GardentFest[1]

Workshop in recognition of Claire Gardent.
Contributions from Claire and colleagues to the field of Computational Linguistics

Christophe Cerisara, Yannick Parmentier (Éds.)

Nancy, France, 21st March 2023

[1] https://synalp.gitlabpages.inria.fr/gardentfest/

Preface

We started to work with Claire some time ago, when writing a PhD thesis on semantic construction with tree-adjoining grammars under her supervision (Yannick, 2003-2007) and as a co-head of the SyNaLP group at Loria (Christophe, since 2012). Over all these years, we were witnesses of the many contributions made by Claire to the field, and of the many collaborations she has been having, either via the supervision of master and PhD theses, or via the numerous projects she investigated or took part in.

Her major impact on the field (and on NLP research in Nancy) has been recognize by various awards, including the 2022 CNRS Silver Medal. At this occasion, we are delighted to organize this workshop in recognition to Claire, and are grateful to Loria for its support.

In this volume, you will find title, abstracts (when available) and URL of papers co-authored by Claire along here career. These are gathered according to Claire's affiliation at the time of writing, and to the contribution's topic. On top of these, you will find titles and abstracts of the five invited talks given by Benoit Crabbé, Marc Dymetman, Shashi Narayan, Mark Steedman and Bonnie Webber. Finally, this volume features seven selected contributions submitted by colleagues of Claire and sorted by topic.

Wishing you a pleasant reading, thank you Claire for all you accomplished (and still do).

Christophe Cerisara, Yannick Parmentier (March 2023)

Invited talks

Invited talk by Bonnie Webber, University of Edinburgh, recipient of the ACL Lifetime Achievement Award (2020)
Title: Supporting Further Advances in Discourse-based Sentence Splitting.

Abstract: In recent work, Claire together with her student Liam Cripwell and colleague Joël Legrand explored sentence-splitting (mapping a complex sentence into a sequence of simpler sentences) from the dual perspectives of sentence-level syntax and discourse) [Cripwell et al, 2021; 2022]. I found the work particularly interesting, and have started speculating on whether the effort could be taken further by taking account of properties of version 3 of the Penn Discourse TreeBank (PDTB 3.0), which annotates several thousand more instances of intra-sentential discourse relations, many modified forms of discourse connectives, and cases where two discourse spans (sentences or clauses) have both an explicitly marked relation between them and one that has been left unmarked.
Liam Cripwell, Joël Legrand, and Claire Gardent (2021). Discourse-based sentence splitting. Findings of the Association for Computational Linguistics (EMNLP 2021), pages 261–273.
Liam Cripwell, Joël Legrand, and Claire Gardent (2022). Controllable Sentence Simplification via Operation Classification. Findings of the Association for Computational Linguistics (NAACL 2022), pages 2091–2103.

Invited talk by Marc Dymetman, Naverlabs, Grenoble
Title: Controlling the Quality of Large Language Models: a Distributional Approach

Abstract: I will cover a line of work and collaborations, started a few years ago at NAVER Labs, where one augments a standard neural language model with constraints over the generative distribution. These help account for aspects of the training data that may be missed by these models (descriptive dimension) but also permit to introduce normative criteria (prescriptive dimension) controlling for biases, offensiveness, or other deficiencies of the standard training process.

Invited talk by Shashi Narayan, Google Inc.
Title: Conditional Generation with Question-Answering Blueprint

Abstract: The ability to convey relevant and faithful information is critical for many tasks in conditional generation and yet remains elusive for neural seq-to-seq models whose outputs often reveal hallucinations and fail to correctly cover important details. In this work, we advocate planning as a useful intermediate representation for rendering conditional generation less opaque and more grounded. We propose a new conceptualization of text plans as a sequence of question-answer (QA) pairs and enhance existing datasets (e.g., for summarization) with a QA *blueprint* operating as a proxy for content selection (i.e., what to say) and planning (i.e., in what order). We obtain blueprints automatically by exploiting state-of-the-art question generation technology and convert input-output pairs into input-blueprint-output tuples. We develop Transformer-based models, each varying in how they incorporate the blueprint in the generated output (e.g., as a global plan or iteratively). Evaluation across metrics and datasets demonstrates that blueprint models are more factual than alternatives which do not resort to planning and allow tighter control of the generation output.

Invited talk by Benoit Crabbé, Université Paris Cité, honorary member of the Institut Universitaire de France (2014)
Title: The promise of language models for language sciences ? let's chat !

Abstract: The field of Computational linguistics is currently is going through a period of paradigm shift Large language models are now ubiquitous with chat GPT creating the last buzz. If you ask chat GPT its promises for the future of language sciences, you get the somewhat confident reply: "Large language models like myself hold great promise for the field of linguistics. They offer improved language understanding, access to vast amounts of data, automatic language analysis, and the ability to test linguistic theories. These tools can help linguists to gain new insights into how language works, identify patterns in language usage, and refine their linguistic theories." In this talk I will put in perspective some key modeling directions in computational linguistics: modeling language structure and modeling language in relation with the world knowledge. And I will explain how we eventually end up with the current language models. We will show that given what they are, current language models achieve sometimes surprising results with respect to the modeling of language structure and highlight some potential research perspectives in language sciences and some of their current limitations.

Invited talk by Mark Steedman, University of Edinburgh, recipient of the ACL Lifetime Achievement Award (2018)
Title: Inference in the Time of GPT ⋆

Abstract: Large pretrained Language Models (LLM) such as GPT3 have upended NLP, calling into question many established methods. In particular, they have been claimed to be capable of doing logical inference when fine-tuned on entailment datasets, or prompted with small numbers of examples of inferential tasks. The talk will review and assess these claims, and propose that we should not give up on alternative methods.
⋆ With apologies to Gabriel Garcia Marquez.

Table of contents

Part I

1989 – 1991 | PhD in Edinburgh, employed by Clermont Ferrand University on EU funded ACORD Project

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

EFFICIENT PARSING FOR FRENCH

Claire Gardent **Gabriel G. Bias** **Pierre-Frangois Jurie** **Karine Baschung**

ABSTRACT

Please see paper using the link below.

27th Annual Meeting of the Association for Computational Linguistics

https://aclanthology.org/P89-1034.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 2–2.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

FRENCH ORDER WITHOUT ORDER

Gabriel G. Bes **Claire Gardent**

ABSTRACT

Please see paper using the link below.

Fourth Conference of the European Chapter of the Association for Computational Linguistics

https://aclanthology.org/E89-1034.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 3–3.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

THE GENERAL ARCHITECTURE OF GENERATION IN ACORD

Dieter Kohl **Agnes Plainfosse** **Claire Gardent**

ABSTRACT

Please see paper using the link below.

COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics

https://aclanthology.org/C90-3082.pdf

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Generating from a Deep Structure

Claire Gardent **Agnes Plainfosse**

Abstract

Please see paper using the link below.

COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics

https://aclanthology.org/C90-2022.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 5–5.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

A UNIFICATION-BASED APPROACH TO MULTIPLE VP ELLIPSIS RESOLUTION

Claire Gardent

ABSTRACT

Please see paper using the link below.

Sixth Conference of the European Chapter of the Association for Computational Linguistics

https://aclanthology.org/E93-1018.pdf

Part II

1991 – 1994 | Post-doc at Utrecht University and University of Amsterdam

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

TALKING ABOUT TREES

Patrick Blackburn **Claire Gardent** **Wilfried Meyer-Viol**

ABSTRACT

Please see paper using the link below.

Sixth Conference of the European Chapter of the Association for Computational Linguistics

`https://aclanthology.org/E93-1004.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

A Specification Language for Lexical Functional Grammars

Patrick Blackburn **Claire Gardent**

Abstract

Please see paper using the link below.

Seventh Conference of the European Chapter of the Association for Computational Linguistics

https://aclanthology.org/E95-1006.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 9–9.

Part III

1994 – 2001 | Post-Doc at Universitaet des Saarlandes

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

HIGHER-ORDER COLOURED UNIFICATION AND NATURAL LANGUAGE SEMANTICS

Claire Gardent **Michael Kohlhase**

ABSTRACT

Please see paper using the link below.

34th Annual Meeting of the Association for Computational Linguistics

`https://aclanthology.org/P96-1001.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

FOCUS AND HIGHER-ORDER UNIFICATION

Claire Gardent **Michael Kohlhase**

ABSTRACT

Please see paper using the link below.

COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics

`https://aclanthology.org/C96-1073.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 12–12.

DESCRIBING DISCOURSE SEMANTICS

Claire Gardent **Bonnie Webber**

ABSTRACT

Please see paper using the link below.

Proceedings of the Fourth International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+4)

`https://aclanthology.org/W98-0113.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 13–13.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

UNIFYING PARALLELS

Claire Gardent

ABSTRACT

Please see paper using the link below.

Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics

`https://aclanthology.org/P99-1007.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

UNDERSTANDING "EACH OTHER"

Claire Gardent **Karsten Konrad**

ABSTRACT

Please see paper using the link below.

1st Meeting of the North American Chapter of the Association for Computational Linguistics

`https://aclanthology.org/A00-2042.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 15–15.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GENERATING WITH A GRAMMAR BASED ON TREE DESCRIPTIONS: A CONSTRAINT-BASED APPROACH

Claire Gardent **Stefan Thater**

ABSTRACT

Please see paper using the link below.

Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics

`https://aclanthology.org/P01-1028.pdf`

Part IV

2001 – now | CNRS researcher at Loria – contributions on Anaphora resolution

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GENERATING MINIMAL DEFINITE DESCRIPTIONS

Claire Gardent

ABSTRACT

Please see paper using the link below.

Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics

https://aclanthology.org/P02-1013.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 18–18.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

WHICH BRIDGES FOR BRIDGING DEFINITE DESCRIPTIONS?

Claire Gardent **Hélène Manuélian** **Eric Kow**

ABSTRACT

Please see paper using the link below.

Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003

https://aclanthology.org/W03-2410.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 19–19.

Part V

2001 – now | CNRS researcher at Loria – contributions on Formal Grammars

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

SEMANTIC CONSTRUCTION IN F-TAG

Claire Gardent **Laura Kallmeyer**

ABSTRACT

Please see paper using the link below.

10th Conference of the European Chapter of the Association for Computational Linguistics

`https://aclanthology.org/E03-1030.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 21–21.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

PARAPHRASTIC GRAMMARS

Claire Gardent **Marilisa Amoia** **Evelyne Jacquey**

ABSTRACT

Please see paper using the link below.

Proceedings of the 2nd Workshop on Text Meaning and Interpretation

`https://aclanthology.org/W04-0910.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 22–22.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

SemTAG, the LORIA toolbox for TAG-based Parsing and Generation

Eric Kow **Yannick Parmentier** **Claire Gardent**

Abstract

Please see paper using the link below.

Proceedings of the Eighth International Workshop on Tree Adjoining Grammar and Related Formalisms

https://aclanthology.org/W06-1516.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 23–23.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

THREE REASONS TO ADOPT TAG-BASED SURFACE REALISATION

Claire Gardent **Eric Kow**

ABSTRACT

Please see paper using the link below.

Proceedings of the Eighth International Workshop on Tree Adjoining Grammar and Related Formalisms

`https://aclanthology.org/W06-1513.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 24–24.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

INTÉGRATION D'UNE DIMENSION SÉMANTIQUE DANS LES GRAMMAIRES D'ARBRES ADJOINTS

Claire Gardent

ABSTRACT

Dans cet article, nous considérons un formalisme linguistique pour lequel l'intégration d'information sémantique dans une grammaire à large couverture n'a pas encore été réalisée à savoir, les grammaires d'arbres adjoints (Tree Adjoining Grammar ou TAG). Nous proposons une méthode permettant cette intégration et décrivons sa mise en oeuvre dans une grammaire noyau pour le français. Nous montrons en particulier que le formalisme de spécification utilisé, XMG, (Duchier et al., 2004) permet une factorisation importante des données sémantiques facilitant ainsi le développement, la maintenance et le déboggage de la grammaire.

Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs

`https://aclanthology.org/2006.jeptalnrecital-long.12.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 25–25.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

COREFERENCE HANDLING IN XMG

Claire Gardent **Yannick Parmentier**

ABSTRACT

Please see paper using the link below.

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions

`https://aclanthology.org/P06-2032.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 26–26.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

SEMTAG: A PLATFORM FOR SPECIFYING TREE ADJOINING GRAMMARS AND PERFORMING TAG-BASED SEMANTIC CONSTRUCTION

Claire Gardent **Yannick Parmentier**

ABSTRACT

Please see paper using the link below.

Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume
Proceedings of the Demo and Poster Sessions

`https://aclanthology.org/P07-2004.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

SEMTAG, UNE ARCHITECTURE POUR LE DÉVELOPPEMENT ET L'UTILISATION DE GRAMMAIRES D'ARBRES ADJOINTS À PORTÉE SÉMANTIQUE

Claire Gardent **Yannick Parmentier**

ABSTRACT

Dans cet article, nous présentons une architecture logicielle libre et ouverte pour le développement de grammaires d'arbres adjoints à portée sémantique. Cette architecture utilise un compilateur de métagrammaires afin de faciliter l'extension et la maintenance de la grammaire, et intègre un module de construction sémantique permettant de vérifier la couverture aussi bien syntaxique que sémantique de la grammaire. Ce module utilise un analyseur syntaxique tabulaire généré automatiquement à partir de la grammaire par le système DyALog. Nous présentons également les résultats de l'évaluation d'une grammaire du français développée au moyen de cette architecture.

Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs

`https://aclanthology.org/2007.jeptalnrecital-long.16.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

INTEGRATING A UNIFICATION-BASED SEMANTICS IN A LARGE SCALE LEXICALISED TREE ADJOINING GRAMMAR FOR FRENCH

Claire Gardent

ABSTRACT

Please see paper using the link below.

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)

`https://aclanthology.org/C08-1032.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 29–29.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

STRUCTURE-DRIVEN LEXICALIST GENERATION

Shashi Narayan **Claire Gardent**

ABSTRACT

Please see paper using the link below.

Proceedings of COLING 2012

`https://aclanthology.org/C12-1124.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 30–30.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GENERATION FOR GRAMMAR ENGINEERING

Claire Gardent **German Kruszewski**

ABSTRACT

Please see paper using the link below.

INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference

`https://aclanthology.org/W12-1507.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 31–31.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

XMG: eXtensible MetaGrammar

Benoît Crabbé **Denys Duchier** **Claire Gardent** **Joseph Le Roux** **Yannick Parmentier**

Abstract

Please see paper using the link below.

Computational Linguistics, Volume 39, Issue 3 - September 2013

`https://aclanthology.org/J13-3005.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Multiple Adjunction in Feature-Based Tree-Adjoining Grammar

Claire Gardent **Shashi Narayan**

Abstract

Please see paper using the link below.

Computational Linguistics, Volume 41, Issue 1 - March 2015

https://aclanthology.org/J15-1003.pdf

Part VI

2001 – now | CNRS researcher at Loria – contributions on Symbolic Natural Language Generation

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GENERATING AND SELECTING GRAMMATICAL PARAPHRASES

Claire Gardent **Eric Kow**

ABSTRACT

Please see paper using the link below.

Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)

`https://aclanthology.org/W05-1605.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

A Symbolic Approach to Near-Deterministic Surface Realisation using Tree Adjoining Grammar

Claire Gardent **Eric Kow**

Abstract

Please see paper using the link below.

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics

`https://aclanthology.org/P07-1042.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 36–36.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

SPOTTING OVERGENERATION SUSPECTS

Claire Gardent **Eric Kow**

ABSTRACT

Please see paper using the link below.

Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)

`https://aclanthology.org/W07-2306.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 37–37.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

UNE RÉALISATEUR DE SURFACE BASÉ SUR UNE GRAMMAIRE RÉVERSIBLE

Claire Gardent **Éric Kow**

ABSTRACT

En génération, un réalisateur de surface a pour fonction de produire, à partir d'une représentation conceptuelle donnée, une phrase grammaticale. Les réalisateur existants soit utilisent une grammaire réversible et des méthodes statistiques pour déterminer parmi l'ensemble des sorties produites la plus plausible ; soit utilisent des grammaires spécialisées pour la génération et des méthodes symboliques pour déterminer la paraphrase la plus appropriée à un contexte de génération donné. Dans cet article, nous présentons GENI, un réalisateur de surface basé sur une grammaire d'arbres adjoints pour le français qui réconcilie les deux approches en combinant une grammaire réversible avec une sélection symbolique des paraphrases.

Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles. Posters

https://aclanthology.org/2007.jeptalnrecital-poster.7.pdf

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

COMPARING THE PERFORMANCE OF TWO TAG-BASED SURFACE REALISERS USING CONTROLLED GRAMMAR TRAVERSAL

Claire Gardent **Benjamin Gottesman** **Laura Perez-Beltrachini**

ABSTRACT

Please see paper using the link below.

Coling 2010: Posters

`https://aclanthology.org/C10-2039.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 39–39.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

RTG BASED SURFACE REALISATION FOR TAG

Claire Gardent **Laura Perez-Beltrachini**

ABSTRACT

Please see paper using the link below.

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)

`https://aclanthology.org/C10-1042.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 40–40.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GÉNÉRATION DE PHRASE : ENTRÉE, ALGORITHMES ET APPLICATIONS (SENTENCE GENERATION: INPUT, ALGORITHMS AND APPLICATIONS)

Claire Gardent

ABSTRACT

Please see paper using the link below.

Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Conférences invitées

`https://aclanthology.org/2011.jeptalnrecital-invite.3.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

KBGEN – TEXT GENERATION FROM KNOWLEDGE BASES AS A NEW SHARED TASK

Eva Banik **Claire Gardent** **Donia Scott** **Nikhil Dinesh** **Fennie Liang**

ABSTRACT

Please see paper using the link below.

INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference

`https://aclanthology.org/W12-1526.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GENERATING ELLIPTIC COORDINATION

Claire Gardent **Shashi Narayan**

ABSTRACT

Please see paper using the link below.

Proceedings of the 14th European Workshop on Natural Language Generation

https://aclanthology.org/W13-2105.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 43–43.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

INCREMENTAL QUERY GENERATION

Laura Perez-Beltrachini **Claire Gardent** **Enrico Franconi**

ABSTRACT

Please see paper using the link below.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics

https://aclanthology.org/E14-1020.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 44–44.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

HYBRID SIMPLIFICATION USING DEEP SEMANTICS AND MACHINE TRANSLATION

Shashi Narayan **Claire Gardent**

ABSTRACT

Please see paper using the link below.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

https://aclanthology.org/P14-1041.pdf

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Surface Realisation from Knowledge-Bases

Bikash Gyawali **Claire Gardent**

Abstract

Please see paper using the link below.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

https://aclanthology.org/P14-1040.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 46–46.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

CONTENT SELECTION AS SEMANTIC-BASED ONTOLOGY EXPLORATION

Laura Perez-Beltrachini **Claire Gardent** **Anselme Revuz** **Saptarashmi Bandyopadhyay**

ABSTRACT

Please see paper using the link below.

Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)

https://aclanthology.org/W16-3508.pdf

Part VII

2001 – now | CNRS researcher at Loria – contributions on Inference

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

ADJECTIVE BASED INFERENCE

Marilisa Amoia **Claire Gardent**

ABSTRACT

Please see paper using the link below.

Proceedings of the Workshop KRAQ'06: Knowledge and Reasoning for Language Processing

`https://aclanthology.org/W06-1805.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 49–49.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

A FIRST ORDER SEMANTIC APPROACH TO ADJECTIVAL INFERENCE

Marilisa Amoia **Claire Gardent**

ABSTRACT

Please see paper using the link below.

Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing

`https://aclanthology.org/W07-1430.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 50–50.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

A Test Suite for Inference Involving Adjectives

Marilisa Amoia **Claire Gardent**

Abstract

Recently, most of the research in NLP has concentrated on the creation of applications coping with textual entailment. However, there still exist very few resources for the evaluation of such applications. We argue that the reason for this resides not only in the novelty of the research field but also and mainly in the difficulty of defining the linguistic phenomena which are responsible for inference. As the TSNLP project has shown test suites provide optimal diagnostic and evaluation tools for NLP applications, as contrary to text corpora they provide a deep insight in the linguistic phenomena allowing control over the data. Thus in this paper, we present a test suite specifically developed for studying inference problems shown by English adjectives. The construction of the test suite is based on the deep linguistic analysis and following classification of entailment patterns of adjectives and follows the TSNLP guidelines on linguistic databases providing a clear coverage, systematic annotation of inference tasks, large reusability and simple maintenance. With the design of this test suite we aim at creating a resource supporting the evaluation of computational systems handling natural language inference and in particular at providing a benchmark against which to evaluate and compare existing semantic analysers.

Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)

`http://www.lrec-conf.org/proceedings/lrec2008/pdf/161_paper.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

RÉÉCRITURE ET DÉTECTION D'IMPLICATION TEXTUELLE

Paul Bédaride **Claire Gardent**

ABSTRACT

Nous présentons un système de normalisation de la variation syntaxique qui permet de mieux reconnaître la relation d'implication textuelle entre deux phrases. Le système est évalué sur une suite de tests comportant 2 520 paires test et les résultats montrent un gain en précision par rapport à un système de base variant entre 29.8 et 78.5 points la complexité des cas considérés.

Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs

`https://aclanthology.org/2008.jeptalnrecital-long.2.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

SEMANTIC NORMALISATION : A FRAMEWORK AND AN EXPERIMENT

Paul Bedaride **Claire Gardent**

ABSTRACT

Please see paper using the link below.

Proceedings of the Eight International Conference on Computational Semantics

https://aclanthology.org/W09-3744.pdf

SYNTACTIC TESTSUITES AND TEXTUAL ENTAILMENT RECOGNITION

Paul Bedaride **Claire Gardent**

ABSTRACT

We focus on textual entailments mediated by syntax and propose a new methodology to evaluate textual entailment recognition systems on such data. The main idea is to generate a syntactically annotated corpus of pairs of (non-)entailments and to use error mining methodology from the parsing field to identify the most likely sources of errors. To generate the evaluation corpus we use a template based generation approach where sentences, semantic representations and syntactic annotations are all created at the same time. Furthermore, we adapt the error mining methodology initially proposed for parsing to the field of textual entailment. To illustrate the approach, we apply the proposed methodology to the Afazio RTE system (an hybrid system focusing on syntactic entailment) and show how it permits identifying the most likely sources of errors made by this system on a testsuite of 10 000 (non-)entailment pairs which is balanced in term of (non-)entailment and in term of syntactic annotations.

Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)

`http://www.lrec-conf.org/proceedings/lrec2010/pdf/379_Paper.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Benchmarking for syntax-based sentential inference

Paul Bedaride **Claire Gardent**

Abstract

Please see paper using the link below.

Coling 2010: Posters

`https://aclanthology.org/C10-2006.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 55–55.

Part VIII

2001 – now | CNRS researcher at Loria – contributions on Natural Language Processing for French

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

EXTRACTION D'INFORMATION DE SOUS-CATÉGORISATION À PARTIR DES TABLES DU LADL

Claire Gardent **Bruno Guillaume** **Guy Perrier** **Ingrid Falk**

ABSTRACT

Les tables du LADL (Laboratoire d'Automatique Documentaire et Linguistique) contiennent des données électroniques extensives sur les propriétés morphosyntaxiques et syntaxiques des foncteurs syntaxiques du français (verbes, noms, adjectifs). Ces données, dont on sait qu'elles sont nécessaires pour le bon fonctionnement des systèmes de traitement automatique des langues, ne sont cependant que peu utilisées par les systèmes actuels. Dans cet article, nous identifions les raisons de cette lacune et nous proposons une méthode de conversion des tables vers un format mieux approprié au traitement automatique des langues.

Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs

https://aclanthology.org/2006.jeptalnrecital-long.11.pdf

SENS, SYNONYMES ET DÉFINITIONS

Ingrid Falk **Claire Gardent** **Évelyne Jacquey** **Fabienne Venant**

ABSTRACT

Cet article décrit une méthodologie visant la réalisation d'une ressource sémantique en français centrée sur la synonymie. De manière complémentaire aux travaux existants, la méthode proposée n'a pas seulement pour objectif d'établir des liens de synonymie entre lexèmes, mais également d'apparier les sens possibles d'un lexème avec les ensembles de synonymes appropriés. En pratique, les sens possibles des lexèmes proviennent des définitions du TLFi et les synonymes de cinq dictionnaires accessibles à l'ATILF. Pour évaluer la méthode d'appariement entre sens d'un lexème et ensemble de synonymes, une ressource de référence a été réalisée pour 27 verbes du français par quatre lexicographes qui ont spécifié manuellement l'association entre verbe, sens (définition TLFi) et ensemble de synonymes. Relativement à ce standard étalon, la méthode d'appariement affiche une F-mesure de 0.706 lorsque l'ensemble des paramètres est pris en compte, notamment la distinction pronominal / non-pronominal pour les verbes du français et de 0.602 sans cette distinction.

Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs

https://aclanthology.org/2009.jeptalnrecital-long.21.pdf

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Classifying French Verbs Using French and English Lexical Resources

Ingrid Falk **Claire Gardent** **Jean-Charles Lamirel**

Abstract

Please see paper using the link below.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

https://aclanthology.org/P12-1090.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 59–59.

Part IX

2001 – now | CNRS researcher at Loria – contributions on Lexicons

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

ÉVALUER SYNLEX

Ingrid Falk **Gil Francopoulo** **Claire Gardent**

ABSTRACT

SYNLEX est un lexique syntaxique extrait semi-automatiquement des tables du LADL. Comme les autres lexiques syntaxiques du français disponibles et utilisables pour le TAL (LEFFF, DICOVALENCE), il est incomplet et n'a pas fait l'objet d'une évaluation permettant de déterminer son rappel et sa précision par rapport à un lexique de référence. Nous présentons une approche qui permet de combler au moins partiellement ces lacunes. L'approche s'appuie sur les méthodes mises au point en acquisition automatique de lexique. Un lexique syntaxique distinct de SYNLEX est acquis à partir d'un corpus de 82 millions de mots puis utilisé pour valider et compléter SYNLEX. Le rappel et la précision de cette version améliorée de SYNLEX sont ensuite calculés par rapport à un lexique de référence extrait de DICOVALENCE.

Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs

`https://aclanthology.org/2007.jeptalnrecital-long.31.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GROUPING SYNONYMS BY DEFINITIONS

Ingrid Falk **Claire Gardent** **Evelyne Jacquey** **Fabienne Venant**

ABSTRACT

Please see paper using the link below.

Proceedings of the International Conference RANLP-2009

`https://aclanthology.org/R09-1015.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 62–62.

IDENTIFYING SOURCES OF WEAKNESS IN SYNTACTIC LEXICON EXTRACTION

Claire Gardent **Alejandra Lorenzo**

ABSTRACT

Previous work has shown that large scale subcategorisation lexicons could be extracted from parsed corpora with reasonably high precision. In this paper, we apply a standard extraction procedure to a 100 millions words parsed corpus of french and obtain rather poor results. We investigate different factors likely to improve performance such as in particular, the specific extraction procedure and the parser used; the size of the input corpus; and the type of frames learned. We try out different ways of interleaving the output of several parsers with the lexicon extraction process and show that none of them improves the results. Conversely, we show that increasing the size of the input corpus and modifying the extraction procedure to better differentiate prepositional arguments from prepositional modifiers improves performance. In conclusion, we suggest that a more sophisticated approach to parser combination and better probabilistic models of the various types of prepositional objects in French are likely ways to get better results.

Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)

`http://www.lrec-conf.org/proceedings/lrec2010/pdf/177_Paper.pdf`

Part X

2001 – now | CNRS researcher at Loria – contribution on Speech Processing

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

VERS LA DÉTECTION DES DISLOCATIONS À GAUCHE DANS LES TRANSCRIPTIONS AUTOMATIQUES DU FRANÇAIS PARLÉ (TOWARDS AUTOMATIC RECOGNITION OF LEFT DISLOCATION IN TRANSCRIPTIONS OF SPOKEN FRENCH)

Corinna Anderson **Christophe Cerisara** **Claire Gardent**

ABSTRACT

Ce travail prend place dans le cadre plus général du développement d'une plate-forme d'analyse syntaxique du français parlé. Nous décrivons la conception d'un modèle automatique pour résoudre le lien anaphorique présent dans les dislocations à gauche dans un corpus de français parlé radiophonique. La détection de ces structures devrait permettre à terme d'améliorer notre analyseur syntaxique en enrichissant les informations prises en compte dans nos modèles automatiques. La résolution du lien anaphorique est réalisée en deux étapes : un premier niveau à base de règles filtre les configurations candidates, et un second niveau s'appuie sur un modèle appris selon le critère du maximum d'entropie. Une évaluation expérimentale réalisée par validation croisée sur un corpus annoté manuellement donne une F-mesure de l'ordre de 40

Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts

`https://aclanthology.org/2011.jeptalnrecital-court.34.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 65–65.

Part XI

2001 – now | CNRS researcher at Loria – contributions on Statistical Natural Language Processing

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Error Mining with Suspicion Trees: Seeing the Forest for the Trees

Shashi Narayan **Claire Gardent**

Abstract

Please see paper using the link below.

Proceedings of COLING 2012

`https://aclanthology.org/C12-1123.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 67–67.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

An End-to-End Evaluation of Two Situated Dialog Systems

Lina M. Rojas-Barahona **Alejandra Lorenzo** **Claire Gardent**

Abstract

Please see paper using the link below.

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue

`https://aclanthology.org/W12-1602.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Error Mining on Dependency Trees

Claire Gardent **Shashi Narayan**

Abstract

Please see paper using the link below.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

https://aclanthology.org/P12-1062.pdf

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

BUILDING AND EXPLOITING A CORPUS OF DIALOG INTERACTIONS BETWEEN FRENCH SPEAKING VIRTUAL AND HUMAN AGENTS

Lina M. Rojas-Barahona **Alejandra Lorenzo** **Claire Gardent**

ABSTRACT

We describe the acquisition of a dialog corpus for French based on multi-task human-machine interactions in a serious game setting. We present a tool for data collection that is configurable for multiple games; describe the data collected using this tool and the annotation schema used to annotate it; and report on the results obtained when training a classifier on the annotated data to associate each player turn with a dialog move usable by a rule based dialog manager. The collected data consists of approximately 1250 dialogs, 10454 utterances and 168509 words and will be made freely available to academic and nonprofit research.

Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)

`http://www.lrec-conf.org/proceedings/lrec2012/pdf/505_Paper.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 70–70.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

LOR-KBGEN, A HYBRID APPROACH TO GENERATING FROM THE KBGEN KNOWLEDGE-BASE

Bikash Gyawali **Claire Gardent**

ABSTRACT

Please see paper using the link below.

Proceedings of the 14th European Workshop on Natural Language Generation

https://aclanthology.org/W13-2131.pdf

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

The KBGen Challenge

Eva Banik **Claire Gardent** **Eric Kow**

Abstract

Please see paper using the link below.

Proceedings of the 14th European Workshop on Natural Language Generation

`https://aclanthology.org/W13-2111.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 72–72.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

A Domain Agnostic Approach to Verbalizing n-ary Events without Parallel Corpora

Bikash Gyawali **Claire Gardent** **Christophe Cerisara**

Abstract

Please see paper using the link below.

Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)

`https://aclanthology.org/W15-4703.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

BUILDING RDF CONTENT FOR DATA-TO-TEXT GENERATION

Laura Perez-Beltrachini **Rania Sayed** **Claire Gardent**

ABSTRACT

In Natural Language Generation (NLG), one important limitation is the lack of common benchmarks on which to train, evaluate and compare data-to-text generators. In this paper, we make one step in that direction and introduce a method for automatically creating an arbitrary large repertoire of data units that could serve as input for generation. Using both automated metrics and a human evaluation, we show that the data units produced by our method are both diverse and coherent.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers

`https://aclanthology.org/C16-1141.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 74–74.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

UNSUPERVISED SENTENCE SIMPLIFICATION USING DEEP SEMANTICS

Shashi Narayan **Claire Gardent**

ABSTRACT

Please see paper using the link below.

Proceedings of the 9th International Natural Language Generation conference

`https://aclanthology.org/W16-6620.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 75–75.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Category-Driven Content Selection

Rania Mohammed **Laura Perez-Beltrachini** **Claire Gardent**

Abstract

Please see paper using the link below.

Proceedings of the 9th International Natural Language Generation conference

`https://aclanthology.org/W16-6616.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 76–76.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

CREATING TRAINING CORPORA FOR NLG MICRO-PLANNERS

Claire Gardent **Anastasia Shimorina** **Shashi Narayan** **Laura Perez-Beltrachini**

ABSTRACT

In this paper, we present a novel framework for semi-automatically creating linguistically challenging micro-planning data-to-text corpora from existing Knowledge Bases. Because our method pairs data of varying size and shape with texts ranging from simple clauses to short texts, a dataset created using this framework provides a challenging benchmark for microplanning. Another feature of this framework is that it can be applied to any large scale knowledge base and can therefore be used to train and learn KB verbalisers. We apply our framework to DBpedia data and compare the resulting dataset with Wen et al. 2016's. We show that while Wen et al.'s dataset is more than twice larger than ours, it is less diverse both in terms of input and in terms of text. We thus propose our corpus generation framework as a novel method for creating challenging data sets from which NLG models can be learned which are capable of handling the complex interactions occurring during in micro-planning between lexicalisation, aggregation, surface realisation, referring expression generation and sentence segmentation. To encourage researchers to take up this challenge, we made available a dataset of 21,855 data/text pairs created using this framework in the context of the WebNLG shared task.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

https://aclanthology.org/P17-1017.pdf

A Statistical, Grammar-Based Approach to Microplanning

Claire Gardent **Laura Perez-Beltrachini**

Abstract

Although there has been much work in recent years on data-driven natural language generation, little attention has been paid to the fine-grained interactions that arise during microplanning between aggregation, surface realization, and sentence segmentation. In this article, we propose a hybrid symbolic/statistical approach to jointly model the constraints regulating these interactions. Our approach integrates a small handwritten grammar, a statistical hypertagger, and a surface realization algorithm. It is applied to the verbalization of knowledge base queries and tested on 13 knowledge bases to demonstrate domain independence. We evaluate our approach in several ways. A quantitative analysis shows that the hybrid approach outperforms a purely symbolic approach in terms of both speed and coverage. Results from a human study indicate that users find the output of this hybrid statistic/symbolic system more fluent than both a template-based and a purely symbolic grammar-based approach. Finally, we illustrate by means of examples that our approach can account for various factors impacting aggregation, sentence segmentation, and surface realization.

Computational Linguistics, Volume 43, Issue 1 - April 2017

`https://aclanthology.org/J17-1001.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Analysing Data-To-Text Generation Benchmarks

Laura Perez-Beltrachini **Claire Gardent**

Abstract

A generation system can only be as good as the data it is trained on. In this short paper, we propose a methodology for analysing data-to-text corpora used for training Natural Language Generation (NLG) systems. We apply this methodology to three existing benchmarks. We conclude by eliciting a set of criteria for the creation of a data-to-text benchmark which could help better support the development, evaluation and comparison of linguistically sophisticated data-to-text generators.

Proceedings of the 10th International Conference on Natural Language Generation

`https://aclanthology.org/W17-3537.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

The WebNLG Challenge: Generating Text from RDF Data

Claire Gardent **Anastasia Shimorina** **Shashi Narayan** **Laura Perez-Beltrachini**

Abstract

The WebNLG challenge consists in mapping sets of RDF triples to text. It provides a common benchmark on which to train, evaluate and compare "microplanners", i.e. generation systems that verbalise a given content by making a range of complex interacting choices including referring expression generation, aggregation, lexicalisation, surface realisation and sentence segmentation. In this paper, we introduce the microplanning task, describe data preparation, introduce our evaluation methodology, analyse participant results and provide a brief description of the participating systems.

Proceedings of the 10th International Conference on Natural Language Generation

`https://aclanthology.org/W17-3518.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

HANDLING RARE ITEMS IN DATA-TO-TEXT GENERATION

Anastasia Shimorina **Claire Gardent**

ABSTRACT

Neural approaches to data-to-text generation generally handle rare input items using either delexicalisation or a copy mechanism. We investigate the relative impact of these two methods on two datasets (E2E and WebNLG) and using two evaluation settings. We show (i) that rare items strongly impact performance; (ii) that combining delexicalisation and copying yields the strongest improvement; (iii) that copying underperforms for rare and unseen items and (iv) that the impact of these two mechanisms greatly varies depending on how the dataset is constructed and on how it is split into train, dev and test.

Proceedings of the 11th International Conference on Natural Language Generation

https://aclanthology.org/W18-6543.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 81–81.

Part XII

2001 – now | CNRS researcher at Loria – contributions on Symbolic Computer Assisted Language Learning

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

USING FB-LTAG DERIVATION TREES TO GENERATE TRANSFORMATION-BASED GRAMMAR EXERCISES

Claire Gardent **Laura Perez-Beltrachini**

ABSTRACT

Please see paper using the link below.

Proceedings of the 11th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+11)

`https://aclanthology.org/W12-4614.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 83–83.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GENERATING GRAMMAR EXERCISES

Laura Perez-Beltrachini **Claire Gardent** **German Kruszewski**

ABSTRACT

Please see paper using the link below.

Proceedings of the Seventh Workshop on Building Educational Applications Using NLP

https://aclanthology.org/W12-2017.pdf

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

REPRESENTATION OF LINGUISTIC AND DOMAIN KNOWLEDGE FOR SECOND LANGUAGE LEARNING IN VIRTUAL WORLDS

Alexandre Denis **Ingrid Falk** **Claire Gardent** **Laura Perez-Beltrachini**

ABSTRACT

There has been much debate, both theoretical and practical, on how to link ontologies and lexicons in natural language processing (NLP) applications. In this paper, we focus on an application in which lexicon and ontology are used to generate teaching material. We briefly describe the application (a serious game for language learning). We then zoom in on the representation and interlinking of the lexicon and of the ontology. We show how the use of existing standards and of good practice principles facilitates the design of our resources while satisfying the expressivity requirements set by natural language generation.

Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)

`http://www.lrec-conf.org/proceedings/lrec2012/pdf/593_Paper.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 85–85.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Using Paraphrases and Lexical Semantics to Improve the Accuracy and the Robustness of Supervised Models in Situated Dialogue Systems

Claire Gardent **Lina M. Rojas Barahona**

Abstract

Please see paper using the link below.

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing

`https://aclanthology.org/D13-1076.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 86–86.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

WEAKLY AND STRONGLY CONSTRAINED DIALOGUES FOR LANGUAGE LEARNING

Claire Gardent **Alejandra Lorenzo** **Laura Perez-Beltrachini** **Lina Rojas-Barahona**

ABSTRACT

Please see paper using the link below.

Proceedings of the SIGDIAL 2013 Conference

`https://aclanthology.org/W13-4056.pdf`

Part XIII

2001 – now | CNRS researcher at Loria – contributions on Neural Natural Language Generation

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

LEARNING EMBEDDINGS TO LEXICALISE RDF PROPERTIES

Laura Perez-Beltrachini **Claire Gardent**

ABSTRACT

Please see paper using the link below.

Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics

https://aclanthology.org/S16-2027.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 89–89.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

THE WEBNLG CHALLENGE: GENERATING TEXT FROM DBPEDIA DATA

Emilie Colin **Claire Gardent** **Yassine M'rabet** **Shashi Narayan** **Laura Perez-Beltrachini**

ABSTRACT

Please see paper using the link below.

Proceedings of the 9th International Natural Language Generation conference

`https://aclanthology.org/W16-6626.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GENERATING PARAPHRASES FROM DBPEDIA USING DEEP LEARNING

Amin Sleimi **Claire Gardent**

ABSTRACT

Please see paper using the link below.

Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)

https://aclanthology.org/W16-3511.pdf

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 91–91.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

ALIGNING TEXTS AND KNOWLEDGE BASES WITH SEMANTIC SENTENCE SIMPLIFICATION

Yassine Mrabet **Pavlos Vougiouklis** **Halil Kilicoglu** **Claire Gardent** **Dina Demner-Fushman**

Jonathon Hare **Elena Simperl**

ABSTRACT

Please see paper using the link below.

Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)

`https://aclanthology.org/W16-3506.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Content Selection through Paraphrase Detection: Capturing different Semantic Realisations of the Same Idea

Elena Lloret **Claire Gardent**

Abstract

Please see paper using the link below.

Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)

`https://aclanthology.org/W16-3505.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 93–93.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

SPLIT AND REPHRASE

Shashi Narayan **Claire Gardent** **Shay B. Cohen** **Anastasia Shimorina**

ABSTRACT

We propose a new sentence simplification task (Split-and-Rephrase) where the aim is to split a complex sentence into a meaning preserving sequence of shorter sentences. Like sentence simplification, splitting-and-rephrasing has the potential of benefiting both natural language processing and societal applications. Because shorter sentences are generally better processed by NLP systems, it could be used as a preprocessing step which facilitates and improves the performance of parsers, semantic role labellers and machine translation systems. It should also be of use for people with reading disabilities because it allows the conversion of longer sentences into shorter ones. This paper makes two contributions towards this new task. First, we create and make available a benchmark consisting of 1,066,115 tuples mapping a single complex sentence to a sequence of sentences expressing the same meaning. Second, we propose five models (vanilla sequence-to-sequence to semantically-motivated models) to understand the difficulty of the proposed task.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing

`https://aclanthology.org/D17-1064.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

DEEP LEARNING APPROACHES TO TEXT PRODUCTION

Claire Gardent **Shashi Narayan**

ABSTRACT

Text production is a key component of many NLP applications. In data-driven approaches, it is used for instance, to generate dialogue turns from dialogue moves, to verbalise the content of Knowledge bases or to generate natural English sentences from rich linguistic representations, such as dependency trees or Abstract Meaning Representations. In text-driven methods on the other hand, text production is at work in sentence compression, sentence fusion, paraphrasing, sentence (or text) simplification, text summarisation and end-to-end dialogue systems. Following the success of encoder-decoder models in modeling sequence-rewriting tasks such as machine translation, deep learning models have successfully been applied to the various text production tasks. In this tutorial, we will cover the fundamentals and the state-of-the-art research on neural models for text production. Each text production task raises a slightly different communication goal (e.g, how to take the dialogue context into account when producing a dialogue turn; how to detect and merge relevant information when summarising a text; or how to produce a well-formed text that correctly capture the information contained in some input data in the case of data-to-text generation). We will outline the constraints specific to each subtasks and examine how the existing neural models account for them.

Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts

`https://aclanthology.org/N18-6002.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GENERATING SYNTACTIC PARAPHRASES

Emilie Colin **Claire Gardent**

ABSTRACT

We study the automatic generation of syntactic paraphrases using four different models for generation: data-to-text generation, text-to-text generation, text reduction and text expansion, We derive training data for each of these tasks from the WebNLG dataset and we show (i) that conditioning generation on syntactic constraints effectively permits the generation of syntactically distinct paraphrases for the same input and (ii) that exploiting different types of input (data, text or data+text) further increases the number of distinct paraphrases that can be generated for a given input.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing

`https://aclanthology.org/D18-1113.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

LORIA / LORRAINE UNIVERSITY AT MULTILINGUAL SURFACE REALISATION 2019

Anastasia Shimorina **Claire Gardent**

ABSTRACT

This paper presents the LORIA / Lorraine University submission at the Multilingual Surface Realisation shared task 2019 for the shallow track. We outline our approach and evaluate it on 11 languages covered by the shared task. We provide a separate evaluation of each component of our pipeline, concluding on some difficulties and suggesting directions for future work.

Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)

`https://aclanthology.org/D19-6312.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 97–97.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

USING LOCAL KNOWLEDGE GRAPH CONSTRUCTION TO SCALE SEQ2SEQ MODELS TO MULTI-DOCUMENT INPUTS

Angela Fan **Claire Gardent** **Chloé Braud** **Antoine Bordes**

ABSTRACT

Query-based open-domain NLP tasks require information synthesis from long and diverse web results. Current approaches extractively select portions of web text as input to Sequence-to-Sequence models using methods such as TF-IDF ranking. We propose constructing a local graph structured knowledge base for each query, which compresses the web search information and reduces redundancy. We show that by linearizing the graph into a structured input sequence, models can encode the graph representations within a standard Sequence-to-Sequence setting. For two generative tasks with very long text input, long-form question answering and multi-document summarization, feeding graph representations as input can achieve better performance than using retrieved text portions.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)

`https://aclanthology.org/D19-1428.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 98–98.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

ENHANCING AMR-TO-TEXT GENERATION WITH DUAL GRAPH REPRESENTATIONS

Leonardo F. R. Ribeiro **Claire Gardent** **Iryna Gurevych**

ABSTRACT

Generating text from graph-based data, such as Abstract Meaning Representation (AMR), is a challenging task due to the inherent difficulty in how to properly encode the structure of a graph with labeled edges. To address this difficulty, we propose a novel graph-to-sequence model that encodes different but complementary perspectives of the structural information contained in the AMR graph. The model learns parallel top-down and bottom-up representations of nodes capturing contrasting views of the graph. We also investigate the use of different node message passing strategies, employing different state-of-the-art graph encoders to compute node representations based on incoming and outgoing perspectives. In our experiments, we demonstrate that the dual graph representation leads to improvements in AMR-to-text generation, achieving state-of-the-art results on two AMR datasets

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)

`https://aclanthology.org/D19-1314.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

SURFACE REALISATION USING FULL DELEXICALISATION

Anastasia Shimorina **Claire Gardent**

ABSTRACT

Surface realisation (SR) maps a meaning representation to a sentence and can be viewed as consisting of three subtasks: word ordering, morphological inflection and contraction generation (e.g., clitic attachment in Portuguese or elision in French). We propose a modular approach to surface realisation which models each of these components separately, and evaluate our approach on the 10 languages covered by the SR'18 Surface Realisation Shared Task shallow track. We provide a detailed evaluation of how word order, morphological realisation and contractions are handled by the model and an analysis of the differences in word ordering performance across languages.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)

https://aclanthology.org/D19-1305.pdf

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

REVISITING THE BINARY LINEARIZATION TECHNIQUE FOR SURFACE REALIZATION

Yevgeniy Puzikov **Claire Gardent** **Ido Dagan** **Iryna Gurevych**

ABSTRACT

End-to-end neural approaches have achieved state-of-the-art performance in many natural language processing (NLP) tasks. Yet, they often lack transparency of the underlying decision-making process, hindering error analysis and certain model improvements. In this work, we revisit the binary linearization approach to surface realization, which exhibits more interpretable behavior, but was falling short in terms of prediction accuracy. We show how enriching the training data to better capture word order constraints almost doubles the performance of the system. We further demonstrate that encoding both local and global prediction contexts yields another considerable performance boost. With the proposed modifications, the system which ranked low in the latest shared task on multilingual surface realization now achieves best results in five out of ten languages, while being on par with the state-of-the-art approaches in others.

Proceedings of the 12th International Conference on Natural Language Generation

`https://aclanthology.org/W19-8635.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GENERATING TEXT FROM ANONYMISED STRUCTURES

Emilie Colin **Claire Gardent**

ABSTRACT

Surface realisation (SR) consists in generating a text from a meaning representations (MR). In this paper, we introduce a new parallel dataset of deep meaning representations (MR) and French sentences and we present a novel method for MR-to-text generation which seeks to generalise by abstracting away from lexical content. Most current work on natural language generation focuses on generating text that matches a reference using BLEU as evaluation criteria. In this paper, we additionally consider the model's ability to reintroduce the function words that are absent from the deep input meaning representations. We show that our approach increases both BLEU score and the scores used to assess function words generation.

Proceedings of the 12th International Conference on Natural Language Generation

`https://aclanthology.org/W19-8614.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Creating a Corpus for Russian Data-to-Text Generation Using Neural Machine Translation and Post-Editing

Anastasia Shimorina **Elena Khasanova** **Claire Gardent**

Abstract

In this paper, we propose an approach for semi-automatically creating a data-to-text (D2T) corpus for Russian that can be used to learn a D2T natural language generation model. An error analysis of the output of an English-to-Russian neural machine translation system shows that 80

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing

https://aclanthology.org/W19-3706.pdf

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

MULTILINGUAL AMR-TO-TEXT GENERATION

Angela Fan **Claire Gardent**

ABSTRACT

Generating text from structured data is challenging because it requires bridging the gap between (i) structure and natural language (NL) and (ii) semantically underspecified input and fully specified NL output. Multilingual generation brings in an additional challenge: that of generating into languages with varied word order and morphological properties. In this work, we focus on Abstract Meaning Representations (AMRs) as structured input, where previous research has overwhelmingly focused on generating only into English. We leverage advances in cross-lingual embeddings, pretraining, and multilingual models to create multilingual AMR-to-text models that generate in twenty one different languages. Our multilingual models surpass baselines that generate into one language in eighteen languages, based on automatic metrics. We analyze the ability of our multilingual models to accurately capture morphology and word order using human evaluation, and find that native speakers judge our generations to be fluent.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)

`https://aclanthology.org/2020.emnlp-main.231.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 104–104.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

The Natural Language Pipeline, Neural Text Generation and Explainability

Juliette Faille **Albert Gatt** **Claire Gardent**

Abstract

End-to-end encoder-decoder approaches to data-to-text generation are often black boxes whose predictions are difficult to explain. Breaking up the end-to-end model into sub-modules is a natural way to address this problem. The traditional pre-neural Natural Language Generation (NLG) pipeline provides a framework for breaking up the end-to-end encoder-decoder. We survey recent papers that integrate traditional NLG submodules in neural approaches and analyse their explainability. Our survey is a first step towards building explainable neural NLG models.

2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence

`https://aclanthology.org/2020.nl4xai-1.5.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

THE 2020 BILINGUAL, BI-DIRECTIONAL WEBNLG+ SHARED TASK: OVERVIEW AND EVALUATION RESULTS (WEBNLG+ 2020)

Thiago Castro Ferreira **Claire Gardent** **Nikolai Ilinykh** **Chris van der Lee** **Simon Mille**

Diego Moussallem **Anastasia Shimorina**

ABSTRACT

WebNLG+ offers two challenges: (i) mapping sets of RDF triples to English or Russian text (generation) and (ii) converting English or Russian text to sets of RDF triples (semantic parsing). Compared to the eponymous WebNLG challenge, WebNLG+ provides an extended dataset that enable the training, evaluation, and comparison of microplanners and semantic parsers. In this paper, we present the results of the generation and semantic parsing task for both English and Russian and provide a brief description of the participating systems.

Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)

`https://aclanthology.org/2020.webnlg-1.7.pdf`

A General Benchmarking Framework for Text Generation

Diego Moussallem **Paramjot Kaur** **Thiago Ferreira** **Chris van der Lee** **Anastasia Shimorina**

Felix Conrads **Michael Röder** **René Speck** **Claire Gardent** **Simon Mille** **Nikolai Ilinykh**

Axel-Cyrille Ngonga Ngomo

Abstract

The RDF-to-text task has recently gained substantial attention due to the continuous growth of RDF knowledge graphs in number and size. Recent studies have focused on systematically comparing RDF-to-text approaches on benchmarking datasets such as WebNLG. Although some evaluation tools have already been proposed for text generation, none of the existing solutions abides by the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles and involves RDF data for the knowledge extraction task. In this paper, we present BENG, a FAIR benchmarking platform for Natural Language Generation (NLG) and Knowledge Extraction systems with focus on RDF data. BENG builds upon the successful benchmarking platform GERBIL, is opensource and is publicly available along with the data it contains.

Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)

`https://aclanthology.org/2020.webnlg-1.3.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

LEARNING HEALTH-BOTS FROM TRAINING DATA THAT WAS AUTOMATICALLY CREATED USING PARAPHRASE DETECTION AND EXPERT KNOWLEDGE

Anna Liednikova **Philippe Jolivet** **Alexandre Durand-Salmon** **Claire Gardent**

ABSTRACT

A key bottleneck for developing dialog models is the lack of adequate training data. Due to privacy issues, dialog data is even scarcer in the health domain. We propose a novel method for creating dialog corpora which we apply to create doctor-patient interaction data. We use this data to learn both a generation and a hybrid classification/retrieval model and find that the generation model consistently outperforms the hybrid model. We show that our data creation method has several advantages. Not only does it allow for the semi-automatic creation of large quantities of training data. It also provides a natural way of guiding learning and a novel method for assessing the quality of human-machine interactions.

Proceedings of the 28th International Conference on Computational Linguistics

`https://aclanthology.org/2020.coling-main.55.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Modeling Global and Local Node Contexts for Text Generation from Knowledge Graphs

Leonardo F. R. Ribeiro **Yue Zhang** **Claire Gardent** **Iryna Gurevych**

Abstract

Recent graph-to-text models generate text from graph-based data using either global or local aggregation to learn node representations. Global node encoding allows explicit communication between two distant nodes, thereby neglecting graph topology as all nodes are directly connected. In contrast, local node encoding considers the relations between neighbor nodes capturing the graph structure, but it can fail to capture long-range relations. In this work, we gather both encoding strategies, proposing novel neural models that encode an input graph combining both global and local node contexts, in order to learn better contextualized node embeddings. In our experiments, we demonstrate that our approaches lead to significant improvements on two graph-to-text datasets achieving BLEU scores of 18.01 on the AGENDA dataset, and 63.69 on the WebNLG dataset for seen categories, outperforming state-of-the-art models by 3.7 and 3.1 points, respectively.1

Transactions of the Association for Computational Linguistics, Volume 8

`https://aclanthology.org/2020.tacl-1.38.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GATHERING INFORMATION AND ENGAGING THE USER COMBOT: A TASK-BASED, SERENDIPITOUS DIALOG MODEL FOR PATIENT-DOCTOR INTERACTIONS

Anna Liednikova **Philippe Jolivet** **Alexandre Durand-Salmon** **Claire Gardent**

ABSTRACT

We focus on dialog models in the context of clinical studies where the goal is to help gather, in addition to the close information collected based on a questionnaire, serendipitous information that is medically relevant. To promote user engagement and address this dual goal (collecting both a predefined set of data points and more informal information about the state of the patients), we introduce an ensemble model made of three bots: a task-based, a follow-up and a social bot. We introduce a generic method for developing follow-up bots. We compare different ensemble configurations and we show that the combination of the three bots (i) provides a better basis for collecting information than just the information seeking bot and (ii) collects information in a more user-friendly, more efficient manner that an ensemble model combining the information seeking and the social bot.

Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations

`https://aclanthology.org/2021.nlpmc-1.3.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 110–110.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Entity-Based Semantic Adequacy for Data-to-Text Generation

Juliette Faille **Albert Gatt** **Claire Gardent**

Abstract

While powerful pre-trained language models have improved the fluency of text generation models, semantic adequacy -the ability to generate text that is semantically faithful to the input- remains an unsolved issue. In this paper, we introduce a novel automatic evaluation metric, Entity-Based Semantic Adequacy, which can be used to assess to what extent generation models that verbalise RDF (Resource Description Framework) graphs produce text that contains mentions of the entities occurring in the RDF input. This is important as RDF subject and object entities make up 2/3 of the input. We use our metric to compare 25 models from the WebNLG Shared Tasks and we examine correlation with results from human evaluations of semantic adequacy. We show that while our metric correlates with human evaluation scores, this correlation varies with the specifics of the human evaluation setup. This suggests that in order to measure the entity-based adequacy of generated texts, an automatic metric such as the one proposed here might be more reliable, as less subjective and more focused on correct verbalisation of the input, than human evaluation measures.

Findings of the Association for Computational Linguistics: EMNLP 2021

`https://aclanthology.org/2021.findings-emnlp.132.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

DISCOURSE-BASED SENTENCE SPLITTING

Liam Cripwell **Joël Legrand** **Claire Gardent**

ABSTRACT

Sentence splitting involves the segmentation of a sentence into two or more shorter sentences. It is a key component of sentence simplification, has been shown to help human comprehension and is a useful preprocessing step for NLP tasks such as summarisation and relation extraction. While several methods and datasets have been proposed for developing sentence splitting models, little attention has been paid to how sentence splitting interacts with discourse structure. In this work, we focus on cases where the input text contains a discourse connective, which we refer to as discourse-based sentence splitting. We create synthetic and organic datasets for discourse-based splitting and explore different ways of combining these datasets using different model architectures. We show that pipeline models which use discourse structure to mediate sentence splitting outperform end-to-end models in learning the various ways of expressing a discourse relation but generate text that is less grammatical; that large scale synthetic data provides a better basis for learning than smaller scale organic data; and that training on discourse-focused, rather than on general sentence splitting data provides a better basis for discourse splitting.

Findings of the Association for Computational Linguistics: EMNLP 2021

https://aclanthology.org/2021.findings-emnlp.25.pdf

AN ERROR ANALYSIS FRAMEWORK FOR SHALLOW SURFACE REALIZATION

Anastasia Shimorina **Yannick Parmentier** **Claire Gardent**

ABSTRACT

Abstract The metrics standardly used to evaluate Natural Language Generation (NLG) models, such as BLEU or METEOR, fail to provide information on which linguistic factors impact performance. Focusing on Surface Realization (SR), the task of converting an unordered dependency tree into a well-formed sentence, we propose a framework for error analysis which permits identifying which features of the input affect the models' results. This framework consists of two main components: (i) correlation analyses between a wide range of syntactic metrics and standard performance metrics and (ii) a set of techniques to automatically identify syntactic constructs that often co-occur with low performance scores. We demonstrate the advantages of our framework by performing error analysis on the results of 174 system runs submitted to the Multilingual SR shared tasks; we show that dependency edge accuracy correlate with automatic metrics thereby providing a more interpretable basis for evaluation; and we suggest ways in which our framework could be used to improve models and data. The framework is available in the form of a toolkit which can be used both by campaign organizers to provide detailed, linguistically interpretable feedback on the state of the art in multilingual SR, and by individual researchers to improve models and datasets.1

Transactions of the Association for Computational Linguistics, Volume 9

https://aclanthology.org/2021.tacl-1.26.pdf

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

AUGMENTING TRANSFORMERS WITH KNN-BASED COMPOSITE MEMORY FOR DIALOG

Angela Fan **Claire Gardent** **Chloé Braud** **Antoine Bordes**

ABSTRACT

Various machine learning tasks can benefit from access to external information of different modalities, such as text and images. Recent work has focused on learning architectures with large memories capable of storing this knowledge. We propose augmenting generative Transformer neural networks with KNN-based Information Fetching (KIF) modules. Each KIF module learns a read operation to access fixed external knowledge. We apply these modules to generative dialog modeling, a challenging task where information must be flexibly retrieved and incorporated to maintain the topic and flow of conversation. We demonstrate the effectiveness of our approach by identifying relevant knowledge required for knowledgeable but engaging dialog from Wikipedia, images, and human-written dialog utterances, and show that leveraging this retrieved information improves model performance, measured by automatic and human evaluation.

Transactions of the Association for Computational Linguistics, Volume 9

`https://aclanthology.org/2021.tacl-1.6.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 114–114.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Exploring the Influence of Dialog Input Format for Unsupervised Clinical Questionnaire Filling

Farnaz Ghassemi Toudeshki **Anna Liednikova** **Philippe Jolivet** **Claire Gardent**

Abstract

In the medical field, we have seen the emergence of health-bots that interact with patients to gather data and track their state. One of the downstream application is automatic questionnaire filling, where the content of the dialog is used to automatically fill a pre-defined medical questionnaire. Previous work has shown that answering questions from the dialog context can successfully be cast as a Natural Language Inference (NLI) task and therefore benefit from current pre-trained NLI models. However, NLI models have mostly been trained on text rather than dialogs, which may have an influence on their performance. In this paper, we study the influence of content transformation and content selection on the questionnaire filling task. Our results demonstrate that dialog pre-processing can significantly improve the performance of zero-shot questionnaire filling models which take health-bots dialogs as input.

Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)

https://aclanthology.org/2022.louhi-1.1.pdf

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

CONTROLLABLE SENTENCE SIMPLIFICATION VIA OPERATION CLASSIFICATION

Liam Cripwell **Joël Legrand** **Claire Gardent**

ABSTRACT

Different types of transformations have been used to model sentence simplification ranging from mainly local operations such as phrasal or lexical rewriting, deletion and re-ordering to the more global affecting the whole input sentence such as sentence rephrasing, copying and splitting. In this paper, we propose a novel approach to sentence simplification which encompasses four global operations: whether to rephrase or copy and whether to split based on syntactic or discourse structure. We create a novel dataset that can be used to train highly accurate classification systems for these four operations. We propose a controllable-simplification model that tailors simplifications to these operations and show that it outperforms both end-to-end, non-controllable approaches and previous controllable approaches.

Findings of the Association for Computational Linguistics: NAACL 2022

https://aclanthology.org/2022.findings-naacl.161.pdf

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GENERATING BIOGRAPHIES ON WIKIPEDIA: THE IMPACT OF GENDER BIAS ON THE RETRIEVAL-BASED GENERATION OF WOMEN BIOGRAPHIES

Angela Fan **Claire Gardent**

ABSTRACT

Generating factual, long-form text such as Wikipedia articles raises three key challenges: how to gather relevant evidence, how to structure information into well-formed text, and how to ensure that the generated text is factually correct. We address these by developing a model for English text that uses a retrieval mechanism to identify relevant supporting information on the web and a cache-based pre-trained encoder-decoder to generate long-form biographies section by section, including citation information. To assess the impact of available web evidence on the output text, we compare the performance of our approach when generating biographies about women (for which less information is available on the web) vs. biographies generally. To this end, we curate a dataset of 1,500 biographies about women. We analyze our generated text to understand how differences in available web evidence data affect generation. We evaluate the factuality, fluency, and quality of the generated texts using automatic metrics and human evaluation. We hope that these techniques can be used as a starting point for human writers, to aid in reducing the complexity inherent in the creation of long-form, factual text.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

`https://aclanthology.org/2022.acl-long.586.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 117–117.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GENERATING QUESTIONS FROM WIKIDATA TRIPLES

Kelvin Han **Thiago Castro Ferreira** **Claire Gardent**

ABSTRACT

Question generation from knowledge bases (or knowledge base question generation, KBQG) is the task of generating questions from structured database information, typically in the form of triples representing facts. To handle rare entities and generalize to unseen properties, previous work on KBQG resorted to extensive, often ad-hoc pre- and post-processing of the input triple. We revisit KBQG – using pre training, a new (triple, question) dataset and taking question type into account – and show that our approach outperforms previous work both in a standard and in a zero-shot setting. We also show that the extended KBQG dataset (also helpful for knowledge base question answering) we provide allows not only for better coverage in terms of knowledge base (KB) properties but also for increased output variability in that it permits the generation of multiple questions from the same KB triple.

Proceedings of the Thirteenth Language Resources and Evaluation Conference

`https://aclanthology.org/2022.lrec-1.29.pdf`

Part XIV

2001 – now | CNRS researcher at Loria – contributions on Neural Natural Language Understanding

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

ORTHOGONALITY REGULARIZER FOR QUESTION ANSWERING

Chunyang Xiao **Guillaume Bouchard** **Marc Dymetman** **Claire Gardent**

ABSTRACT

Please see paper using the link below.

Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics

`https://aclanthology.org/S16-2019.pdf`

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

SEQUENCE-BASED STRUCTURED PREDICTION FOR SEMANTIC PARSING

Chunyang Xiao **Marc Dymetman** **Claire Gardent**

ABSTRACT

Please see paper using the link below.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

`https://aclanthology.org/P16-1127.pdf`

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 121–121.

Part XV

Selected contributions in the field of linguistic resources

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

Creation of an annotated corpus for the processing of definite descriptions

Hélène Manuélian
Laboratoire LT2D
CY Cergy Paris Université
F95000 CERGY
helene.manuelian@cyu.fr

Claire Gardent
CNRS – LORIA, Nancy, France
claire.gardent@loria.fr

1 Introduction

In 2005, the development of annotated corpora was necessary for the evaluation and training of algorithms for the automatic processing of anaphors. It is still the case today, as shown by the DEMOCRAT project [1].

In this paper, we proposed a methodology for the annotation of definite descriptions. We annotated about 5,000 definite descriptions in a consistent and useful way for the solving modules. We then presented the results of this annotation and discussed their implications for automatic resolution. Later, this corpus was known as DeDe corpus and was available on the Atilf website.

In these days, the DeDe corpus was one of the biggest corpora annotated with anaphoric relations as shown in the first part of the paper.

2 Methodology

To ensure the consistency of our annotation we followed four principles.

- First, the corpus was pre-processed to make annotation easier and therefore more reliable.
- Expert linguists annotated the corpus.
- We refined the annotation scheme by several iterations.
- We took time for an important adjudication phase to ensure the consistency of the annotation.

3 Annotation and results

During the first pass of the annotation, we classified the definite descriptions into four non-mutually exclusive categories:

- autonomous descriptions which are interpretable independently of any context,
- coreferential descriptions whose referent had already been mentioned in the text,
- contextual descriptions which include associative and situational descriptions. Their referent has not been mentioned in the previous text, but its presence is inferred from an element of the context.
- non referential descriptions.

When several categories were possible, we proposed strategies to make a choice, in connection with the second pass of annotation.

The second pass of annotation refined the classification.

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 123–124.

Autonomous descriptions (more than half of the corpus, 58%) could be classified in the subcategories: anaphoric, circumstantial, generic, unique, identifying.

Coreferential descriptions, (approximately 11% of the annotated descriptions), could fall into the following categories:

- direct: the description is the same as the previous one.
- lexical: a lexical relationship is identifiable between the two coreferential expressions.
- redescription: there is no lexical link between the two expressions.

Contextual definite descriptions, (24% of the annotated descriptions) were divided between associative definite descriptions and situational definite descriptions. The referent of associative descriptions is inferred from an expression found in the text, whereas the referent of situational descriptions is inferred from the general situation.

Non-referential definite descriptions (appositions, conjunctions, idiomatic expressions, predications, or quantifiers) represented about 9% of the corpus.

4 Conclusion

The annotation covered a corpus of 48,360 words and 4,910 definite descriptions. It allowed us to define of a new annotation strategy for the definite descriptions. We then created a corpus - at this time - close to a reference corpus for the resolution of the definite descriptions.

We concluded at the time that the reference chains remained to be annotated in order to report the ambiguities of categorization that we had encountered. Today, of course, many advances have been made and the production of annotated corpora has progressed (with the ANCOR and DEMOCRAT corpora in particular, [1], [2]). Anyway, the work carried out in 2005 and before laid the foundations for these projects. This is why, on this day of homage to Claire Gardent, I wanted to recall this aspect of her work, perhaps a little marginal in relation to her other work, but still, crucial for research in linguistics.

References

[1] Frederic LANDRAGIN Le corpus DEMOCRAT et son exploitation. Presentation, In *Langages, 2021/4 (Nb 224)*, pages 11–24. DOI : 10.3917/lang.224.0011.

[2] Judith MUZERELLE, Aurore PELLETIER-BOYER, Jean-Yves ANTOINE, Emmanuel SCHANG, Denis MAUREL, Jeanne VILLANEAU, Iris ESHKOL (2012) Annotation en relations anaphoriques d'un corpus de discours oral spontane en francais In *Proc. CMLF'2012, Lyon [HAL-00788164]*,

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

THE WEBNLG DATASET

Laura Perez-Beltrachini
University of Edinburgh
lperez@ed.ac.uk

Shashi Narayan
Google Brain, London
shashinarayan@google.com

Anastasia Shimorina
Orange Labs
anastasia.shimorina@orange.com

Claire Gardent
CNRS-LORIA
gardent@loria.fr

Description

The WebNLG KB-to-text generation benchmark Gardent et al. [2017a] was developed under Claire's French National Research Agency award on Natural Language Generation for the Semantic Web, namely the WebNLG project. The dataset consists of varied, relevant and coherent sets of DBPedia triples Perez-Beltrachini et al. [2016] paired with human authored texts that correctly capture their meaning Gardent et al. [2017a].

The WebNLG dataset has proven to be an extremely valuable resource for the natural language processing and semantic web research communities. It enabled research on neural architectures for data-to-text generation Trisedya et al. [2018], Marcheggiani and Perez-Beltrachini [2018], Jagfeld et al. [2018], Castro Ferreira et al. [2019], Harkous et al. [2020], Li et al. [2020], Xu et al. [2021], Han et al. [2021], Su et al. [2021], Chowdhery et al. [2022], Yin and Wan [2022], Duong et al. [2023]. It is not only still widely used for research in KB-to-text generation but has also been used for other tasks such as text re-writing Narayan et al. [2017], paraphrase generation Colin and Gardent [2018], relation extraction Zeng et al. [2018], Zheng et al. [2021], Su et al. [2023], and entity linking Kilias et al. [2019].

The WebNLG dataset was first used in the WebNLG 2017 KB-to-text generation challenge Colin et al. [2016], Gardent et al. [2017b] in English and give place to subsequent iterations of the challenge featuring new related tasks. The WebNLG 2020 challenge Castro Ferreira et al. [2020] focused on multilingual generation including the Russian version of the dataset Shimorina et al. [2019] and include a bi-directional task using the dataset for semantic parsing (e.g., text-to-KG). This year, the WebNLG 2023 challenge focuses on text generation with under-resourced languages including Maltese, Breton, Welsh and Irish.

References

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada, July 2017a. Association for Computational Linguistics. doi:10.18653/v1/P17-1017. URL https://aclanthology.org/P17-1017.

Laura Perez-Beltrachini, Rania Sayed, and Claire Gardent. Building RDF content for data-to-text generation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1493–1502, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1141.

Bayu Distiawan Trisedya, Jianzhong Qi, Rui Zhang, and Wei Wang. GTR-LSTM: A triple encoder for sentence generation from RDF data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1627–1637, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:10.18653/v1/P18-1151. URL https://aclanthology.org/P18-1151.

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 125–127.

Diego Marcheggiani and Laura Perez-Beltrachini. Deep graph convolutional encoders for structured data to text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 1–9, Tilburg University, The Netherlands, November 2018. Association for Computational Linguistics. doi:10.18653/v1/W18-6501. URL https://aclanthology.org/W18-6501.

Glorianna Jagfeld, Sabrina Jenne, and Ngoc Thang Vu. Sequence-to-sequence models for data-to-text natural language generation: Word- vs. character-based processing and output diversity. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 221–232, Tilburg University, The Netherlands, November 2018. Association for Computational Linguistics. doi:10.18653/v1/W18-6529. URL https://aclanthology.org/W18-6529.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 552–562, Hong Kong, China, November 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-1052. URL https://aclanthology.org/D19-1052.

Hamza Harkous, Isabel Groves, and Amir Saffari. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2410–2424, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi:10.18653/v1/2020.coling-main.218. URL https://aclanthology.org/2020.coling-main.218.

Xintong Li, Aleksandre Maskharashvili, Symon Jory Stevens-Guille, and Michael White. Leveraging large pretrained models for WebNLG 2020. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 117–124, Dublin, Ireland (Virtual), 12 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.webnlg-1.12.

Xinnuo Xu, Ondřej Dušek, Verena Rieser, and Ioannis Konstas. AggGen: Ordering and aggregating while generating. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1419–1434, Online, August 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.113. URL https://aclanthology.org/2021.acl-long.113.

Jiuzhou Han, Daniel Beck, and Trevor Cohn. Generating diverse descriptions from semantic graphs. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 1–11, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.inlg-1.1.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. Plan-then-generate: Controlled data-to-text generation via planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 895–909, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.findings-emnlp.76. URL https://aclanthology.org/2021.findings-emnlp.76.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022.

Xunjian Yin and Xiaojun Wan. How do seq2seq models perform on end-to-end data-to-text generation? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7701–7710, 2022.

Song Duong, Alberto Lumbreras, Mike Gartrell, and Patrick Gallinari. Learning from multiple sources for data-to-text and text-to-data. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023,*, Valencia, Spain, 2023. URL https://arxiv.org/pdf/2302.11269.pdf.

Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. Split and rephrase. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 606–616, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi:10.18653/v1/D17-1064. URL https://aclanthology.org/D17-1064.

Emilie Colin and Claire Gardent. Generating syntactic paraphrases. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 937–943, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi:10.18653/v1/D18-1113. URL `https://aclanthology.org/D18-1113`.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:10.18653/v1/P18-1047. URL `https://aclanthology.org/P18-1047`.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. PRGC: Potential relation and global correspondence based joint relational triple extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6225–6235, Online, August 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.486. URL `https://aclanthology.org/2021.acl-long.486`.

Huizhe Su, Hao Wang, Xiangfeng Luo, and Shaorong Xie. An end-to-end neural framework using coarse-to-fine-grained attention for overlapping relational triple extraction. *Natural Language Engineering*, page 1–24, 2023. doi:10.1017/S1351324923000050.

Torsten Kilias, Alexander Löser, Felix Gers, Ying Zhang, Richard Koopmanschap, and Martin Kersten. Idel: In-database neural entity linking. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–8, 2019. doi:10.1109/BIGCOMP.2019.8679486.

Emilie Colin, Claire Gardent, Yassine M'rabet, Shashi Narayan, and Laura Perez-Beltrachini. The WebNLG challenge: Generating text from DBPedia data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 163–167, Edinburgh, UK, September 5-8 2016. Association for Computational Linguistics. doi:10.18653/v1/W16-6626. URL `https://aclanthology.org/W16-6626`.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain, September 2017b. Association for Computational Linguistics. doi:10.18653/v1/W17-3518. URL `https://aclanthology.org/W17-3518`.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual), 12 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.webnlg-1.7`.

Anastasia Shimorina, Elena Khasanova, and Claire Gardent. Creating a corpus for Russian data-to-text generation using neural machine translation and post-editing. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 44–49, Florence, Italy, August 2019. Association for Computational Linguistics. doi:10.18653/v1/W19-3706. URL `https://aclanthology.org/W19-3706`.

Part XVI

Selected contributions in the field of formal grammars

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

USING A REVERSIBLE GRAMMAR TO MEASURE AND FIX OVER-GENERATION

Eric Kow *
Standard Chartered Bank, Singapore
eric.kow@gmail.com

Yannick Parmentier †
Université de Lorraine – LORIA, Nancy, France
yannick.parmentier@loria.fr

Claire Gardent
CNRS – LORIA, Nancy, France
claire.gardent@loria.fr

ABSTRACT

In this paper, we present a technique combining a semantic parser with a surface realiser to semi-automatically detect the cases of overgeneration in a corpus and provide the list of the grammatical rules that are involved in this overgeneration. The proposed approach relies on two key-characteristics of the grammar: its reversibility and its compilation from a factorized description (metagrammar). We also present an evaluation of the benefits brought by this technique for the development of a tree-adjoining grammar for French.

Keywords Tree-adjoining grammar · reversible grammar · semantic parsing · surface realisation · over-generation

1 Introduction

A generative grammar aims at defining the sentences which belong to a given language. More precisely, a grammar must generate all sentences belonging to a language and only these. While this definition may seem over-restrictive in so far as it prevents such grammars from describing real-life texts, it permits to have a clear characterization of well-formed utterances. In practice, specifying the constraints which apply to a given utterance is a very complex task, and grammars often prove to be too permissive (they can generate sentences which actually do not belong to the target language). In such situation, the grammar is said to over-generate.

In this paper, we present a method to detect and fix over-generation within a Tree-Adjoining Grammar [Joshi et al., 1975] for French. The proposed approach builds upon two main key features of the grammar: its *reversibility* (usability for both parsing and generation) on the one hand, and its *factorized description* (by means of a metagrammar-based compilation) on the other hand.

A grammar is said to be reversible when it can be used for both parsing natural language utterances (e.g. to compute a syntax tree or a semantic representation in our case) and generating texts (verbalising a given meaning, which is usually called surface realization). In this paper, we rely on the SEMFRAG grammar, whose size allows us to generate a sufficient number of sentences to run a large-scale evaluation of the grammar.

Furthermore, the compilation of this grammar from a compact description (so-called metagrammar) allows us to associate every grammar rule (elementary tree) with a description of its linguistic content (e.g. verb form, subject, etc.). In Section 2, we show how this description called hypertag in the litterature, can be used to track the source of over-generation.

*This work is related to a PhD thesis supervised by Claire Gardent and defended on Nov. 2007, 14th.
†This work is related to a PhD thesis supervised by Claire Gardent and defended on Apr. 2007, 6th.

This paper is structured as follows. We first present the SemFrag resource and its compilation. In Sections 3 and 4, we briefly introduce the parsing and generation algorithms. In Section 5, we present the methodology developed to track over-generation. In Section 6, we compare our approach with related work, and finally we conclude in Section 7.

2 SemFrag

SemFrag is a Feature-structure-based Lexicalised Tree-Adjoining Grammar (FBLTAG) [Vijay-Shanker and Joshi, 1988] which includes a support for compositional semantic calculus using unification [Gardent and Kallmeyer, 2003]. An important feature from SemFrag is that it is compiled from a factorized description written in the XMG language [Duchier et al., 2005]. In the following subsections, we detail these concepts.

2.1 Feature-based Lexicalised Tree-Adjoining Grammar (FBLTAG)

An FBLTAG is made of a set of elementary trees, and two rewriting operations: tree substitution and tree adjunction. The result of tree rewritings is called a derived tree.

Elementary trees are lexicalised, which means that they are all associated with at least one lexical item, called anchor (typically a lemma or an inflected form). Tree nodes are labelled with two feature-structures called top and bottom structures. An elementary tree can be either an initial tree (where leaves are either marked for substitution via a $\downarrow$ or terminal symbols), or an auxiliary tree (where a leaf node is called foot node and marked with $\star$ and whose syntactic category is the same as the one of the tree root node).

Substitution consists in replacing a node n marked with $\downarrow$ with a tree whose root r has the same category as n. Adjunction consists in replacing a node n with an auxiliary tree whose root r and foot f nodes have the same category. In these rewritings, top feature-structures of n and r are unified. In adjunction, bottom feature-structures of n and f are unified as well.

2.2 Syntax / semantics interface

Following Gardent and Kallmeyer [2003], each elementary tree is associated with a semantic representation whose missing arguments are unification variables shared with specific features in the tree. When generating a given sentence by means of tree rewritings, these variables get assigned values, as illustrated in Figure 1 below.

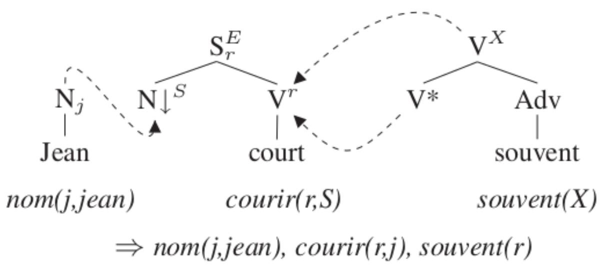


Figure 1: Syntax / semantics interface of "Jean court souvent" (*John often runs*)

2.3 Grammar and metagrammar

SemFrag is built from the compilation of an abstract grammar called a metagrammar. In this metagrammar, each elementary tree is defined as the combination of one or more tree fragments. Each of these fragments encapsulate a specific linguistic feature. For instance, all verb trees whose subject is in a canonical form will instanciate the CanSubject (written SujetCanonique hereafter) tree fragment. Some examples of tree fragments are given in Figure 2.

2.4 Grammar coverage

SemFrag relies on the structure defined in [Crabbé, 2005a]. It has about 4,200 elementary trees describing both the syntax and semantics of a given linguistic structure, which cover roughly the French grammar of Abeillé [2002]. In other words, it covers the syntactic subcategorisation frames listed in this grammar, and for each of these, the various

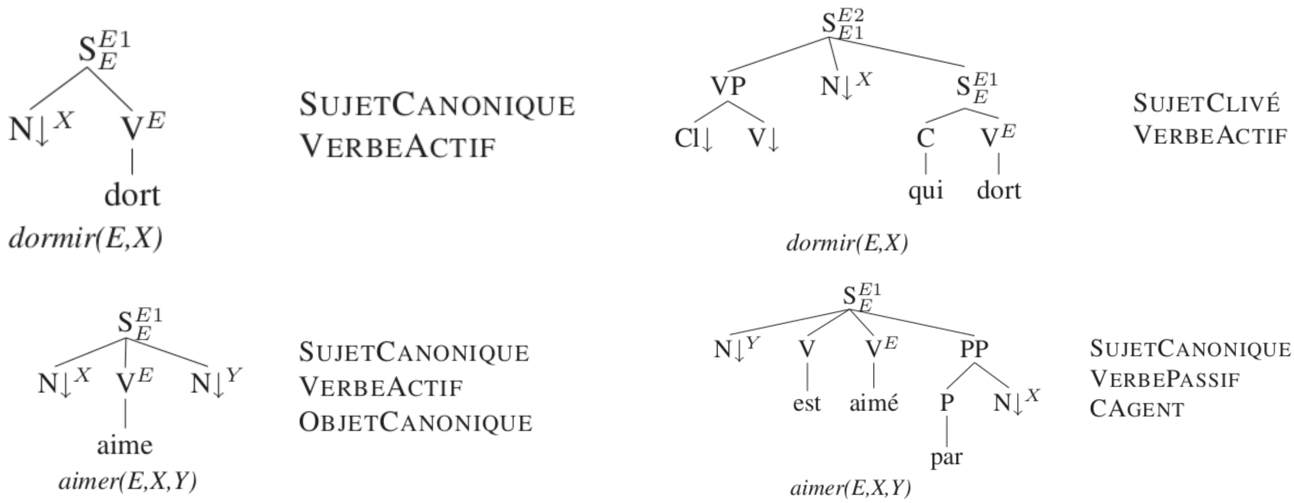


Figure 2: Tree fragments

distributions (e.g. active, passive, reflexive, etc.) and realisations of arguments (e.g. clitics, extraction, etc.). More precisely, the concepts supported by SEMFRAG are the following:

- predicate, verb, nouns and adjectives ;
- predicative arguments ;
- realisations of noun phrases (canonical, clitics, interrogated, relativised, cleft)
- realisations of phrasal arguments (finite completives, infinitives, indirect questions)
- verbal diatheses (active, passive, reflexive, impersonnal, middle)
- verbal subcategorisations (tree families of Abeillé [2002]),
- subject control and raising,
- inverted subject,
- clitics ordering,
- long-distance dependencies,
- modifiers (adjectives, relatives, prepositional phrases),
- coordination,
- adverbial modification,
- phrasal modification.

SEMFRAG has been evaluated on the TSNLP test-set [Lehmann et al., 1996]. SEMFRAG correctly parses 76% of grammatical sentences, and fails at parsing 83% of ungrammatical ones. Syntactic ambiguity amounts to 1.64 parses per sentence [Crabbé, 2005b]. While being uncomplete and far from covering French language, this grammar is a good base for evaluating the impact of the over-generation tracking procedure presented in this paper, thanks to its well defined structure and good documentation.

3 Using SemFrag for parsing

SEMFRAG can be used for semantic construction (that is, for computing semantic representations associated with linguistic expressions by the grammar). As shown by Gardent and Parmentier [2005], this construction can be performed either during or after parsing. SEMFRAG implements the second option, enabling us to stick to the TAG formalism, and to keep our approach modular in so far as syntactic parsing sensu stricto and semantic construction remain independent on each other. Concretely, semantic construction is done in 3 steps:

1. A first step consists in extracting parallel grammars for syntactic and semantic information (rules). From SEMFRAG, two grammars are thus extracted: a syntactic and a semantic one. The syntactic grammar is

made of the elementary trees contained in SEMFRAG without their semantic pieces of information (semantic features labelling nodes and semantic representation). The semantic grammar on the other hand is made of the elementary trees contained in SEMFRAG without their morpho-syntactic pieces of information (node features). Figure 3 below illustrates this for the intransitive verb "dormir" (to sleep in French).

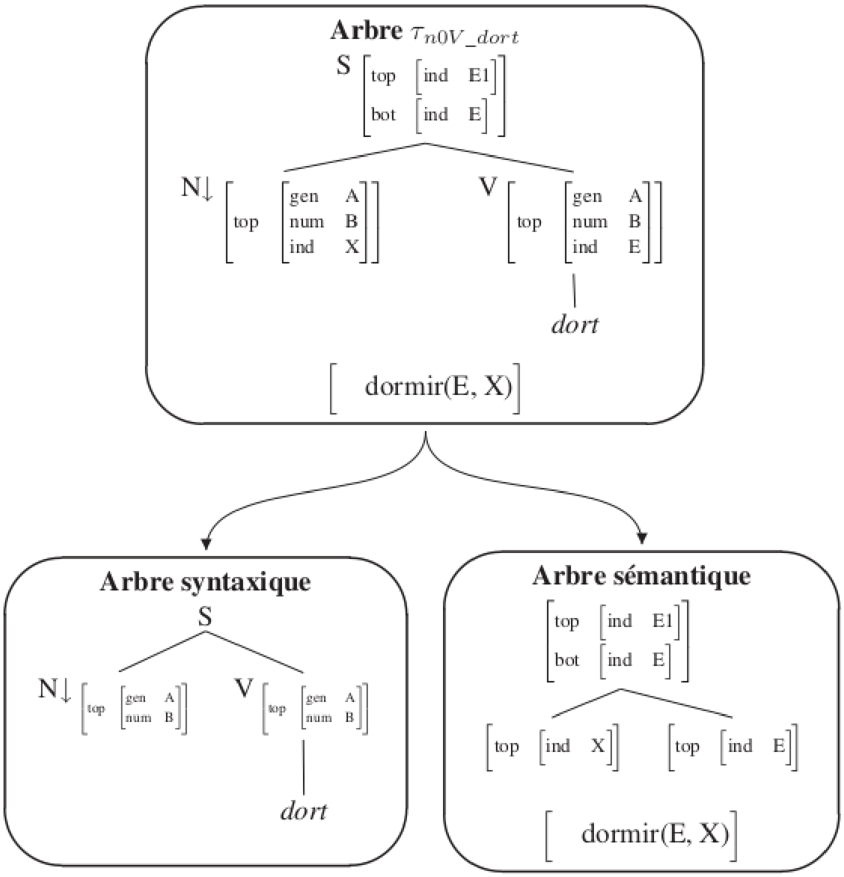


Figure 3: Parallel grammars extraction (example of the intransitive verb "dormir" (*to sleep*) in French

2. A second step consists in performing syntactic parsing by means of the syntactic grammar extracted at step 1. Parsing is done by the DyALog parsing engine of Villemonte de la Clergerie [2005]. This engine computes a derivation forest, which describes in a compact way the set of derivation trees whose yield corresponds to the input sentence. As an example, let us consider the forest depicted in Figure 4 below, which shows the derivations of the ambiguous sentence "Jean regarde Anne avec un télescope" (*John looks at Mary with a telescope*).

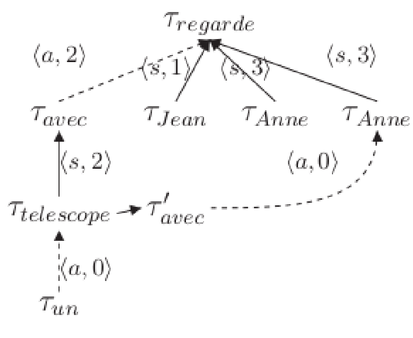


Figure 4: Derivation forest for "Jean regarde Anne avec un télescope" (*John looks at Ann with a telescope*)

3. In a third step, the semantic features get unified. To do so, we use both information included in the lexicon (semantic representations), and in the derivation forest (feature unifications to be applied which are collected via a top-down traversal of the derivation forest). As a result, we obtain an instanciated semantic representation such as $\{sleep(e,j), john(j)\}$.[3]

4 Using SemFrag for generation

SEMFRAG can also be used for surface realisation, which consists in generating the sentences verbalising a given meaning. The realisation algorithm can be summarized as follows:

1. Step 1 (**selection phase**): all elementary trees from SEMFRAG whose semantic description is subsumed by the input semantic representation are selected. These semantic descriptionsare then unified with the input semantic representation.

2. Step 2 (**combination phase**): all selected (and semantically instanciated) elementary trees are combined with each other using first susbtitution, and then adjunction.

3. Step 3 (**extraction phase**): all yields made of terminal symbols only and associated with semantic representations which are identical to the input semantics are extracted.

This surface realisation algorithm uses both an agenda to store the combinations which can be done (valid auxiliary trees) and a chart to store the combinations done so far (derived trees). For more details, please consult [Gardent and Kow, 2005]. Figure 5 gives an overview of the surface realisation process.

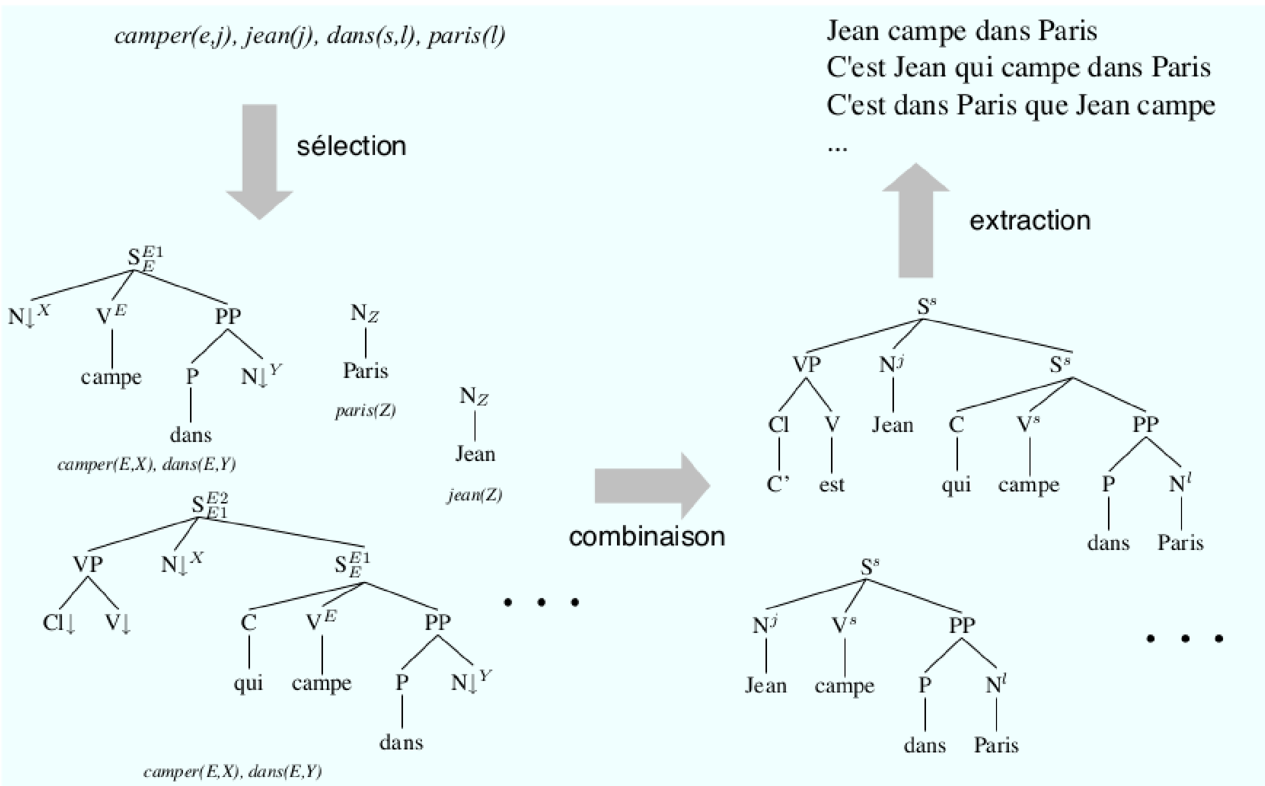


Figure 5: Surface realisation of the meaning $\{camper(e,j), jean(j), dans(s,l), paris(l)\}$ (realisations of the concept of John camping in Paris)

[3] SEMFRAG's semantic representations are flat logical expressions belonging to Predicate Logic Unplugged [Bos, 1995].

5 Tracking over-generation

As shown in previous sections, SEMFRAG is a reversible grammar, which can be used to compute the semantic representation(s) of a given sentence, or to generate the various sentences which convey a given meaning. This feature makes it possible to use SEMFRAG to both visualize the sentences licensed by the grammar (using surface realisation) and to create in a semi-automatic way the semantic representations needed for generation (by manually selecting semantic representations produced at parsing), as we will discuss below.

5.1 Creating a test-suite to measure over-generation

As shown by the development of the Redwood Lingo Treebank [Oepen et al., 2002], an important feature of reversible grammars is that they facilitate the automatic compilation of test-suites for realisation. It merely requires one to parse a set of sentences, and to select those which correctly verbalise the intended meaning. More generally, such grammars facilitate the evaluation of surface realisers, and the detection of over-generation (ungrammatical sentences which are licensed by the grammar).

In order to build a test-suite to evaluate SEMFRAG's over-generation, we use the semantic construction procedure introduced in section 3. This test-suite is composed of pairs of the form: (semantic representation, sentence). Its items are selected in order to cover all the linguistic structures which the grammar should license. Selected sentences contain one, two or three conjugated verbs of the following types:

- standard verbs (e.g. to sleep, eat, accept, read), 26 distinct sub-categorisation frames are supported ;
- control verbs (e.g. want, can, persuade), subject and object control are supported ;
- predicative verbs (e.g. to be, to remain) ;
- verbs with phrasal subjects.

Recursive structures (e.g. John says that Peter thinks that Mary sleeps) are supported as well. Since we are interested in checking over-generation, we only includes canonical sentences in our test-suite (non-canonical sentences are associated with the same meaning representation).

The test-suite is built as follows. For each canonical sentence parsed by the semantic construction module, a linguist annotator selects the correct semantic representation(s). Such representations are then saved in a file, together with the sentence which has been used to compute them. When no semantic representation is computed, the sentence is ignored. This way, we collected 140 pairs associating a meaning representation with a sentence. Even though this test set is relatively small, it is sufficient to perform a systematic evaluation of the syntactic realisations licensed by SEMFRAG.

5.2 Sentences generated by the test set

To evaluate over-generation, the surface realiser is run in a non-deterministic mode on the test set. Figure 6 below shows how many sentences can be generated for each of the input meaning representations before and after over-generation correction. For instance, before correction 48 input representations generate less than 10 sentences. After correction, 64 representations generate less than 10 paraphrases.

More generally, this figure shows the leveraging effect of over-generation tracking. It strongly reduces the number of semantic entries which generate more than 90 paraphrases and it significantly improves the number of semantic entries which generate less than 10 paraphrases.

Without any over-generation tracking, SEMFRAG associates the 140 meaning representations of the test set with 28,167 sentences, with a worst-case of 4,908 sentences for a given representation. This high number of paraphrases is due to two features of the grammar:

- SEMFRAG aims at covering a significant par of French syntax, hence it describes e.g. various realisations of verbal arguments (good feature) ;
- SEMFRAG, like many electronic grammar under development, is too permissive due to erroneous rules. This phenomenon is amplified by the use of a factorised description (so-called metagrammar). Indeed a given error can be propagated in a large number of realisations (and so can be fixes!).

In the following section, we will describe our methodology to identify causes of over-generation in SEMFRAG's meta-description directly. This methodology permitted us to reduce the number of realisations by 70% by applying 13 modifications to the metagrammar, which took a linguist 13 hours.

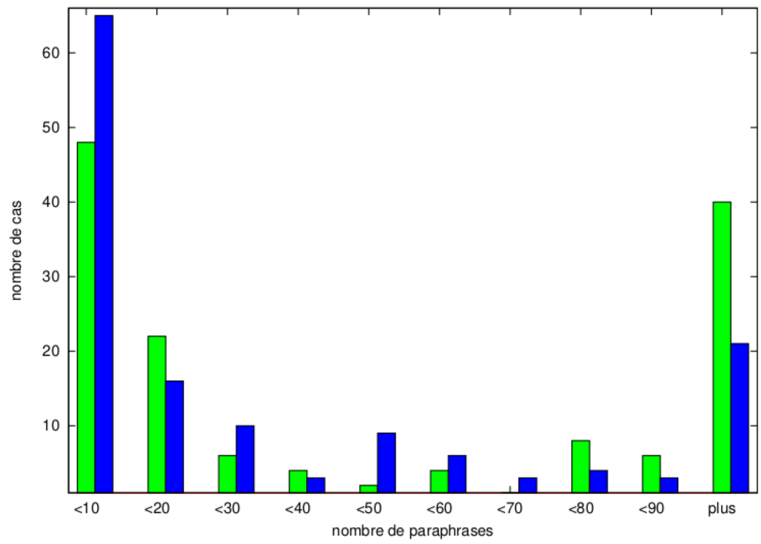


Figure 6: Distribution of the number of sentences generated before (green) and after (blue) over-generation correction (out of 140 input entries)

5.3 Methodology

To identify over-generation, we proceed as follows:

1. surface realiser's output sentences are labelled with either OK or OVER-GENERATION ;
2. linguistic properties (e.g. TAG tree families, morpho-syntactic features) of sentences labelled with OVER-GENERATION are automatically extracted.

This procedure is repeated iteratively (in a continuous integration fashion), while taking care of applying regression tests at the end of each cycle. These tests are used to ensure that the modifications applied to the metagrammar following a given property extraction do not harm its linguistic precision (sentences that were correctly generated before the modifications should still be).

An incremental procedure. As discussed above, SEMFRAG's high over-generation degree is partly due to the fact that this grammar has been designed first for semantic construction, and that it is compiled from a meta-grammar. For some semantic entries, the grammar generates more than 4,000 sentences.

To facilitate the process of tracking over-generation, we start with labelling simple sentences first. For instance, sentences having a single verb are processed prior to those having two or three verbs. Each fix is then applied before labelling the more complex cases (the latter could thus benefit from this fix).

Thanks to this ordering scheme, we managed to reduce the number of realisations by 70%. The remaining 30% are expected paraphrases.

Finding out suspects. For each sentence generated by the surface realiser, we produce the following pieces of information:

- its derivation tree,
- for each elementary tree τ present in this derivation tree:
  - the tree families τ belongs to,
  - its unique identifier (e.g. `Tn0V-615`) in the grammar,
  - its hypertag.

As an illustration, Figure 7 below contains the pieces of information associated with the sentence "Jean demande si c'est Paul qui vient" (*John asks whether Paul comes*). There are three elementary trees involved in this derivation (namely `Tn0ClVs1int-630`, `Tn0V-615`, `TproperName-45` and `TproperName-45`) anchored respectively by the lemmas "demander" (*to ask*), "venir" (*to come*), "paul" and "jean". This log also contains the tree fragments from the metagrammar used to build the various elementary TAG trees (e.g. tree `Tn0ClVs1int-630` is built from tree fragments `CanonicalSubject` and `SententialInterrogative`).

```
Output: jean demande si c'est paul qui vient
demander:n8 <-(s)- venir
demander:n1 <-(s)- jean
venir:n4 <-(s)- paul
demander Tn0ClVs1int-630
  CanonicalSubject
  SententialInterrogative
venir Tn0V-615
  CleftSubje t
paul TproperName-45
jean TproperName-45
```

Figure 7: Example of a log entry.

From an automatic processing of the entries of this log, we automatically compute a list of suspects, that is of linguistic properties (e.g. tree identifier, tree families, derivation trees) appearing only in problematic cases.

Concretely, we compute items of the following forms:

- a lemma associated with trees, tree families and tree fragments from the metagrammar (called tree properties) ;
- a combination operation on two elementary trees: $\tau_i \xleftarrow{Op,n_j} \tau_k$ (operation Op applying tree τ_k on node n_j of tree τ_i).

As an example of the use of this list of suspects, let us consider the sentence "Jean dit accepter/*C'est par Jean qui accepte d'être dit" (*Jean says he accepts/*It is by Jean that it is said to accept*). For this example, we obtain the following log entry:

```
input t90
Lemma: dire
Tn0Vn1 (all) - InfinitiveSubject Passive
  [699] CanonicalCAgent Passive
  [746] CanonicalGenitive dePassive
  [702] CleftCAgentOne Passive
  [752] CleftDont dePassive
  [751] CleftGenitiveOne dePassive
  [750] RelativeGenitive dePassive
```

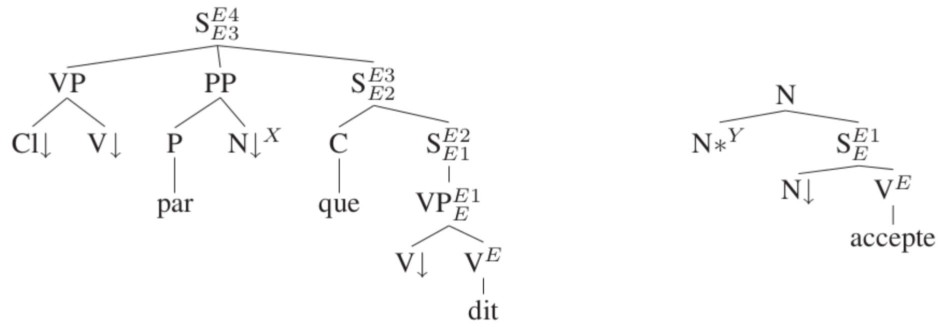
This entry reveals us that all trees of the family `Tn0Vn1` anchored by the verb "dire" (*to say*) are involved in over-generation (there are 6 such trees among the 100 trees in this family). All these trees are using the `InfinitiveSubject` and `Passive` tree fragments (properties). In the current version of the grammar, these trees combine with the verb anchored by "accepter" leading to erroneous derivations. Figure 8 below shows the trees involved in this derivation.

While both trees are valid on their own,[4] they lack some linguistic constraints to prevent this adjunction on the cleft agent. First, the verb anchoring the main phrase should be in infinitive mode. This information is present in a lexical feature `MODE=INF` and should percolate up to the tree root. Furthermore, we should add some constraint on the auxiliary tree anchored by "accepter" to make sure it cannot be adjoined to a tree anchored by a proper noun. Finally, we could define a lexical constraint which would prevent a tree describing a passive voice with a cleft agent from being anchored with "dire" (*to say*).

5.4 Results

We used our over-generation tracking methodology for about a week (corresponding roughly to 13 working hours). Over that time, we went through 10 iterations and applied 13 modifications to the metagrammar. 40 semantic representations (less than one third of the test set) were considered and 1389 sentences manually annotated. Out of

[4]They can be used to describe the syntax of the particular sentence "C'est d'être vu par un lion qui a faim qui est dangereux" (*It is to be seen by a lion which is hungry which is dangerous*).



dire(E,X,E1) *accepter(E,Y)*

Figure 8: Elementary trees involved in an erroneous realisation.

the 140 representations of the test set, the initial grammar generated 28,167 sentences. Once fixed, ungrammatical sentences are cut down by 70% (8,434 sentences).

This high impact from a reduced number of modifications has 2 origins. First, modifications are applied to the metagrammar (and not the grammar itself). Thus, each modification on a given tree fragment (property) is propagated to all elementary trees which use this fragment.[5] Second, the ordered processing of the ill-formed sentences creates a leveraging effect (fixes made on simple sentences are also applied on complex ones).

Let us finally summarize the various sources of over-generation, which were identified during this work.

Missing constraints A frequent cause of over-generation corresponds to missing constraints (e.g. feature constraints) to unlicense some tree rewritings.

Uncomplete constraints or missing percolation equations In some cases, the expected linguistic constraints were encoded in the grammar, but only partly. For instance, some feature equations used to prevent determiners from being adjoined at a given node were missing.

Wrong elementary trees Since trees are compiled by combinations from tree fragments, it may happen that the linguist does not imagine all possible combinations when describing tree families. In such cases, some additional metagrammatical instructions are needed to guide the way fragments get combined (this can be done using either node variables or colors).

Wrong semantic representations In some cases, the semantic representation associated with elemenaty trees may be underspecified.

Lexical exceptions Finally, some lexical entries are underspecified in so far as they are missing some anchoring equations, which would prevent a given lexical item from anchoring all the trees of a given tree family (this kind of specifications are used for instance to deal with verbs which cannot appear in the passive voice).

6 Related work

Using a surface realiser to track over-generation is not new. Karttunen and Kay [1985] for instance used a realiser to detect over-generation in a unification grammar and so did Boguraev et al. [1988] in an Head-drive Phrase Structure Grammar (HPSG) one. The approach taken here differs from above-mentioned ones first from its high degree of automatization. Indeed, Karttunen and Kay [1985] and Boguraev et al. [1988] are mainly resorting to manual (and sometime interactive) inspection of the grammar.

In the context of syntactic parsing, systematic error detection techniques have been proposed over the last few years. In particular, [van Noord, 2004], [Sagot and De La Clergerie, 2006] and [Nicolas et al., 2007] designed techniques to identify suspects (potentiel sources of erroneous parses). These approaches differ from ours in so far as they rely on statistical measures to track under generation (sentences which should be fully parsed using the input grammar but which are only partially). Furthermore, our approach aims at listing candidate trees, rather than candidate lexical entries as sources of errors.

[5] SEMFRAG is made of 4,200 elementary trees compiled from 300 tree fragments.

7 Conclusion

In this paper, we presented an approach to semi-automatically detect causes of over-generation in a Lexicalized Tree-Adjoining Grammar. This approach differs from error mining techniques by its degree of precision, it indeed targets suspect trees or tree fragments. It also differs from existing over-generation tracking approaches by its incremental process which offers some leveraging effect. As a result, most errors can be tracked down in a limited number of iterations, thus saving time.

From its current status, three mid-term objectives are foreseen: first, we would like to increase the size of our semantic dataset (e.g. by generating these semi-automatically). Second, we would like to experiment with probabilistic error mining techniques to see whether we could improve the recall of our approach. Third, we would like to develop techniques to fully automatically validate sentences generated by the surface realiser (by using for instance n-grams).

Finally, let us mention that our approach, while developped for Tree-Adjoining Grammars, could be applied to any type of formal reversible grammar. In particular, it is highly probable that this approach proves useful with resources such as the ERG [Copestake, 2001] or the German Lexical Functional Grammar [Rohrer and Forst, 2006].

References

A. Joshi, L. Levy, and M. Takahashi. Tree adjunct grammars. pages 136–163. 1975. Journal of Comput. Syst. Sci., Vol. 10-1.

K. Vijay-Shanker and Aravind K. Joshi. Feature structures based tree adjoining grammars. In *Proceedings of the international conference on Computational Linguistics (COLING), Budapest, Hungary*, pages 714–719, 1988.

C. Gardent and L. Kallmeyer. Semantic construction in FTAG. In *Proceedings of the European chapter of the Association for Computational Linguistics (EACL'03), Budapest*, pages 123–130, 2003.

D. Duchier, J. Le Roux, and Y. Parmentier. XMG : Un Compilateur de Métagrammaire Extensible. In *Actes de la 12e conférence sur le Traitement Automatique des Langues Naturelles 2005, Dourdan, France*, pages 13–22, 2005.

B. Crabbé. Grammatical development with XMG. Proceedings of the 5th International Conference on the Logical Aspects of Computational Linguistics (LACL05), 2005a.

A. Abeillé. *Une grammaire électronique du français*. CNRS Editions, Paris, 2002.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Hervé Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. TSNLP — Test Suites for Natural Language Processing. In *Proceedings of COLING 1996*, pages 711–716, Kopenhagen, 1996.

B. Crabbé. *Représentation informatique de grammaires fortement lexicalisées : Application à la grammaire d'arbres adjoints*. PhD thesis, Université Nancy 2, 2005b.

C. Gardent and Y. Parmentier. Large scale semantic construction for tree adjoining grammars. In *Proceedings of the 5th International Conference on the Logical Aspects of Computational Linguistics (LACL05), Bordeaux, France*, pages 131–146, 2005.

Éric Villemonte de la Clergerie. DyALog: a tabular logic programming based environment for NLP. In *Proceedings of the 2nd International Workshop on Constraint Solving and Language Processing - CSLP 2005*, pages 18–33, Barcelona, 2005. URL `ftp://ftp.inria.fr/INRIA/Projects/Atoll/Eric.Clergerie/CSLP05.pdf`.

J. Bos. Predicate Logic Unplugged. In *Proceedings of the tenth Amsterdam Colloquium, Amsterdam*, pages 133–142, 1995. URL `citeseer.ist.psu.edu/bos95predicate.html`.

C. Gardent and E. Kow. Generating and selecting grammatical paraphrases. In *Proceedings of the European workshop on Natural Language Generation, Aberdeen, Scotland*, pages 49–57, 2005.

S. Oepen, E. Callahan, C. Manning, and K. Toutanova. LinGO Redwoods: A rich and dynamic treebank for HPSG. In *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT2002), Sozopol, Bulgaria*, 2002.

L. Karttunen and M. Kay. *Natural Language Parsing*, chapter Parsing in a free word order language. Zwicky, 1985.

Bran Boguraev, John Carroll, Ted Briscoe, and Claire Grover. Software support for practical grammar development. In *Proceedings of the 12th conference on Computational linguistics*, pages 54–58, Morristown, NJ, USA, 1988. Association for Computational Linguistics. ISBN 963 8431 56 3. doi:http://dx.doi.org/10.3115/991635.991647.

Gertjan van Noord. Error mining for wide-coverage grammar engineering. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain*, pages 446–453, 2004.

B. Sagot and E. De La Clergerie. Error mining in parsing results. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, Sydney, Australia*, pages 329–336, 2006.

L. Nicolas, J. Farré, and E. De La Clergerie. Confondre le coupable : corrections à un lexique suggérées par une grammaire. In *Farah Benamara, Nabil Hathout, Philippe Muller, and Sylwia Ozdowska, editors, Actes du Colloque TALN 2007, pages 315-324, Toulouse*, Juin 2007.

Ann Copestake. *Implementing Typed Feature Structure Grammars*. Center for the Study of Language and Information - Lecture Notes, 2001.

Christian Rohrer and Martin Forst. Improving coverage and parsing quality of a large-scale lfg for german. In *Proceedings of LREC-2006*, Genoa, Italy, 2006.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

SURFACE REALISATION FROM KNOWLEDGE-BASES

Bikash Gyawali *
Université de Lorraine, LORIA
Villers-lès-Nancy, F-54600, France
bikash.gyawali@loria.fr

Claire Gardent
CNRS, LORIA, UMR 7503
Vandoeuvre-lès-Nancy, F-54500, France
claire.gardent@loria.fr

ABSTRACT

We present a simple, data-driven approach to generation from knowledge bases (KB). A key feature of this approach is that grammar induction is driven by the extended domain of locality principle of TAG (Tree Adjoining Grammar); and that it takes into account both syntactic and semantic information. The resulting extracted TAG includes a unification based semantics and can be used by an existing surface realiser to generate sentences from KB data. Experimental evaluation on the KBGen data shows that our model outperforms a data-driven generate-and-rank approach based on an automatically induced probabilistic grammar; and is comparable with a handcrafted symbolic approach.

1 Introduction

In this paper we present a grammar based approach for generating from knowledge bases (KB) which is linguistically principled and conceptually simple. A key feature of this approach is that grammar induction is driven by the extended domain of locality principle of TAG (Tree Adjoining Grammar) and takes into account both syntactic and semantic information. The resulting extracted TAGs include a unification based semantics and can be used by an existing surface realiser to generate sentences from KB data.

To evaluate our approach, we use the benchmark provided by the KBGen challenge Banik et al. [2012, 2013], a challenge designed to evaluate generation from knowledge bases; where the input is a KB subset; and where the expected output is a complex sentence conveying the meaning represented by the input. When compared with two other systems having taken part in the KBGen challenge, our system outperforms a data-driven, generate-and-rank approach based on an automatically induced probabilistic grammar; and produces results comparable to those obtained by a symbolic, rule based approach. Most importantly, we obtain these results using a general purpose approach that we believe is simpler and more transparent than current state of the art surface realisation systems generating from KB or DB data.

2 Related Work

Our work is related to work on concept to text generation.

Earlier work on concept to text generation mainly focuses on generation from logical forms using rule-based methods. Wang [1980] uses hand-written rules to generate sentences from an extended predicate logic formalism; Shieber et al. [1990] introduces a head-driven algorithm for generating from logical forms; Kay [1996] defines a chart based algorithm which enhances efficiency by minimising the number of semantically incomplete phrases being built; and Shemtov [1996] presents an extension of the chart based generation algorithm presented in Kay [1996] which supports the generation of multiple paraphrases from underspecified semantic input. In all these approaches, grammar and lexicon are developed manually and it is assumed that the lexicon associates semantic sub-formulae with natural language expressions. Our approach is similar to these approaches in that it assumes a grammar encoding a compositional semantics. It differs from them however in that, in our approach, grammar and lexicon are automatically acquired from the data.

* Author's address at the time of original publication of this article. It was published in the ACL 2014 conference.

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 140–150.

With the development of the semantic web and the proliferation of knowledge bases, generation from knowledge bases has attracted increased interest and so called ontology verbalisers have been proposed which support the generation of text from (parts of) knowledge bases. One main strand of work maps each axiom in the knowledge base to a clause. Thus the OWL verbaliser integrated in the Protégé tool Kaljurand and Fuchs [2007] provides a verbalisation of every axiom present in the ontology under consideration and Wilcock [2003] describes an ontology verbaliser using XML-based generation. As discussed in Power and Third [2010], one important limitation of these approaches is that they assume a simple deterministic mapping between knowledge representation languages and some controlled natural language (CNL). Specifically, the assumption is that each atomic term (individual, class, property) maps to a word and each axiom maps to a sentence. As a result, the verbalisation of larger ontology parts can produce very unnatural text such as, *Every cat is an animal. Every dog is an animal. Every horse is an animal. Every rabbit is an animal.* More generally, the CNL based approaches to ontology verbalisation generate clauses (one per axiom) rather than complex sentences and thus cannot adequately handle the verbalisation of more complex input such as the KBGen data where the KB input often requires the generation of a complex sentence rather than a sequence of base clauses.

To generate more complex output from KB data, several alternative approaches have been proposed.

The MIAKT project Bontcheva and Wilks. [2004] and the ONTOGENERATION project Aguado et al. [1998] use symbolic NLG techniques to produce textual descriptions from some semantic information contained in a knowledge base. Both systems require some manual input (lexicons and domain schemas). More sophisticated NLG systems such as TAILOR Paris [1988], MIGRAINE Mittal et al. [1994], and STOP Reiter et al. [2003] offer tailored output based on user/patient models. While offering more flexibility and expressiveness, these systems are difficult to adapt by non-NLG experts because they require the user to understand the architecture of the NLG systems Bontcheva and Wilks. [2004]. Similarly, the NaturalOWL system Galanis et al. [2009] has been proposed to generate fluent descriptions of museum exhibits from an OWL ontology. This approach however relies on extensive manual annotation of the input data.

The SWAT project has focused on producing descriptions of ontologies that are both coherent and efficient Williams and Power [2010]. For instance, instead of the above output, the SWAT system would generate the sentence: *The following are kinds of animals: cats, dogs, horses and rabbits.* . In this approach too however, the verbaliser output is strongly constrained by a simple Definite Clause Grammar covering simple clauses and sentences verbalising aggregation patterns such as the above. More generally, the sentences generated by ontology verbalisers cover a limited set of linguistics constructions; the grammar used is manually defined; and the mapping between semantics and strings is assumed to be deterministic (e.g., a verb maps to a relation and a noun to a concept). In constrast, we propose an approach which can generate complex sentences from KB data; where the grammar is acquired from the data; and where no assumption is made about the mapping between semantics and NL expressions.

Recent work has focused on data-driven generation from frames, lambda terms and data base entries.

DeVault et al. [2008] describes an approach for generating from the frames produced by a dialog system. They induce a probabilistic Tree Adjoining Grammar from a training set aligning frames and sentences using the grammar induction technique of Chiang [2000] and use a beam search that uses weighted features learned from the training data to rank alternative expansions at each step.

The function of a gated channel is to release particles from the endoplasmic reticulum

```
:TRIPLES (
(|Release-Of-Calcium646| |object| |Particle-In-Motion64582|)
(|Release-Of-Calcium646| |base| |Endoplasmic-Reticulum64603|)
(|Gated-Channel64605|  |has-function||Release-Of-Calcium646|)
(|Release-Of-Calcium646|  |agent| |Gated-Channel64605|))
:INSTANCE-TYPES
(|Particle-In-Motion64582| |instance-of| |Particle-In-Motion|)
(|Endoplasmic-Reticulum64603| |instance-of| |Endoplasmic-Reticulum|)
(|Gated-Channel64605| |instance-of| |Gated-Channel|)
 |Release-Of-Calcium646| |instance-of| |Release-Of-Calcium|))
:ROOT-TYPES (
(|Release-Of-Calcium646| |instance-of| |Event|)
(|Particle-In-Motion64582| |instance-of| |Entity|)
(|Endoplasmic-Reticulum64603| |instance-of| |Entity|)
(|Gated-Channel64605| |instance-of| |Entity|)))
```

Figure 1: Example KBGEN Scenario

Lu and Ng [2011] focuses on generating natural language sentences from logical form (i.e., lambda terms) using a synchronous context-free grammar. They introduce a novel synchronous context free grammar formalism for generating from lambda terms; induce such a synchronous grammar using a generative model; and extract the best output sentence from the generated forest using a log linear model.

Wong and Mooney [2007], Lu et al. [2009] focuses on generating from variable-free tree-structured representations such as the CLANG formal language used in the ROBOCUP competition and the database entries collected by Liang et al. [2009] for weather forecast generation and for the air travel domain (ATIS dataset) by Dahl et al. [1994]. Wong and Mooney [2007] uses synchronous grammars to transform a variable free tree structured meaning representation into sentences. Lu et al. [2009] uses a Conditional Random Field to generate from the same meaning representations.

Finally, more recent papers propose approaches which perform both surface realisation and content selection. Angeli et al. [2010] proposes a log linear model which decomposes into a sequence of discriminative local decisions. The first classifier determines which records to mention; the second, which fields of these records to select; and the third, which words to use to verbalise the selected fields. Kim and Mooney [2010] uses a generative model for content selection and verbalises the selected input using WASP^{-1}, an existing generator. Finally, Konstas and Lapata [2012a,b] develop a joint optimisation approach for content selection and surface realisation using a generic, domain independent probabilistic grammar which captures the structure of the database and the mapping from fields to strings. They intersect the grammar with a language model to improve fluency; use a weighted hypergraph to pack the derivations; and find the best derivation tree using Viterbi algorithm.

Our approach differs from the approaches which assume variable free tree structured representations Wong and Mooney [2007], Lu et al. [2009] and data-based entries Kim and Mooney [2010], Konstas and Lapata [2012a,b] in that it handles graph-based, KB input and assumes a compositional semantics. It is closest to DeVault et al. [2008] and Lu and Ng [2011] who extract a grammar encoding syntax and semantics from frames and lambda terms respectively. It differs from the former however in that it enforces a tighter syntax/semantics integration by requiring that the elementary trees of our extracted grammar encode the appropriate linking information. While DeVault et al. [2008] extracts a TAG grammar associating each elementary tree with a semantics, we additionnally require that these trees encode the appropriate linking between syntactic and semantic arguments thereby restricting the space of possible tree combinations and drastically reducing the search space. Although conceptually related to Lu and Ng [2011], our approach extracts a unification based grammar rather than one with lambda terms. The extraction process and the generation algorithms are also fundamentally different. We use a simple mainly symbolic approach whereas they use a generative approach for grammar induction and a discriminative approach for sentence generation.

3 The KBGen Task

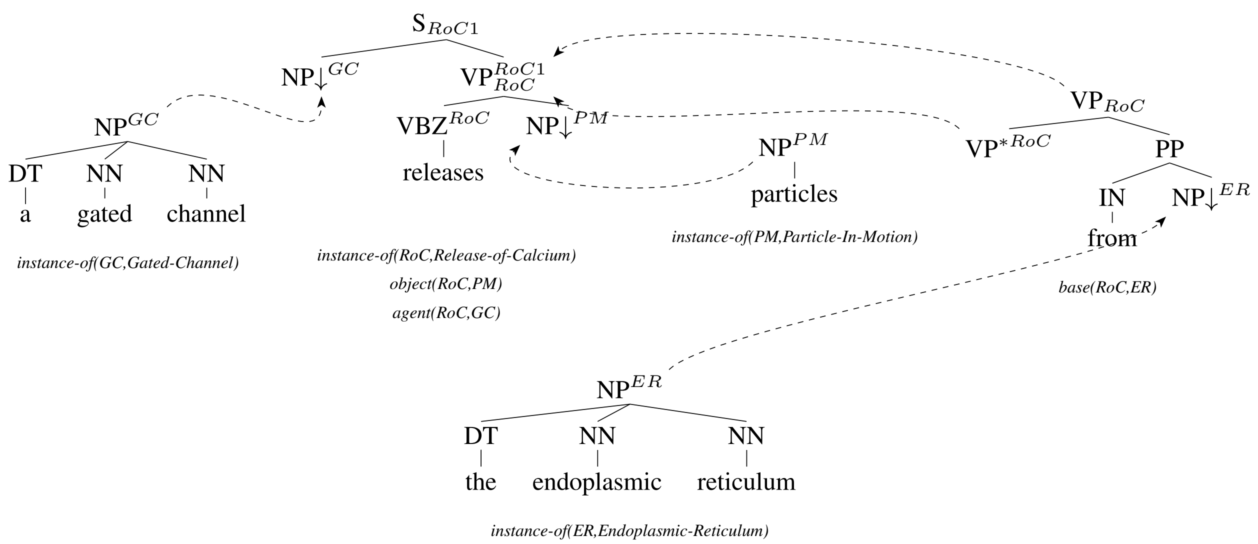


Figure 2: Example FB-LTAG with Unification-Based Semantics. Dotted lines indicate substitution and adjunction operations between trees. The variables decorating the tree nodes (e.g., GC) abbreviate feature structures of the form $[idx : V]$ where V is a unification variable shared with the semantics.

The KBGen task was introduced as a new shared task at Generation Challenges 2013 Banik et al. [2013][2] and aimed to compare different generation systems on KB data. Specifically, the task is to verbalise a subset of a knowledge base. For instance, the KB input shown in Figure 1 can be verbalised as:

(1) The function of a gated channel is to release particles from the endoplasmic reticulum

The KB subsets forming the KBGen input data were pre-selected from the AURA biology knowledge base Gunning et al. [2010], a knowledge base about biology which was manually encoded by biology teachers and encodes knowledge about events, entities, properties and relations where relations include event-to-entity, event-to-event, event-to-property and entity-to-property relations. AURA uses a frame-based knowledge representation and reasoning system called Knowledge Machine Clark and Porter [1997] which was translated into first-order logic with equality and from there, into multiple different formats including SILK Grosof [2012] and OWL2 Motik et al. [2009]. It is available for download in various formats including OWL[3].

4 Generating from the KBGen Knowledge-Base

To generate from the KBGen data, we induce a Feature-Based Lexicalised Tree Adjoining Grammar (FB-LTAG, Vijay-Shanker and Joshi [1988]) augmented with a unification-based semantics Gardent and Kallmeyer [2003] from the training data. We then use this grammar and an existing surface realiser to generate from the test data.

4.1 Feature-Based Lexicalised Tree Adjoining Grammar

Figure 2 shows an example FB-LTAG augmented with a unification-based semantics.

Briefly, an FB-LTAG consists of a set of elementary trees which can be either initial or auxiliary. Initial trees are trees whose leaves are labeled with substitution nodes (marked with a down-arrow) or terminal categories. Auxiliary trees are distinguished by a foot node (marked with a star) whose category must be the same as that of the root node. In addition, in an FB-LTAG, each elementary tree is anchored by a lexical item (lexicalisation) and the nodes in the elementary trees are decorated with two feature structures called *top* and *bottom* which are unified during derivation. Two tree-composition operations are used to combine trees namely, substitution and adjunction. While substitution inserts a tree in a substitution node of another tree, adjunction inserts an auxiliary tree into a tree. In terms of unifications, substitution unifies the top feature structure of the substitution node with the top feature structure of the root of the tree being substituted in. Adjunction unifies the top feature structure of the root of the tree being adjoined with the top feature structure of the node being adjoined to; and the bottom feature structure of the foot node of the auxiliary tree being adjoined with the bottom feature structure of the node being adjoined to.

In an FB-LTAG augmented with a unification-based semantics, each tree is associated with a semantics i.e., a set of literals whose arguments may be constants or unification variables. The semantics of a derived tree is the union of the semantics of the tree contributing to its derivation modulo unification. Importantly, semantic variables are shared with syntactic variables (i.e., variables occurring in the feature structures decorating the tree nodes) so that when trees are combined, the appropriate syntax/semantics linking is enforced. For instance given the semantics:

instance-of(RoC,Release-Of-Calcium),
object(RoC,PM),agent(RoC,GC),base(RoC,ER),
instance-of(ER,Endoplasmic-Reticulum),
instance-of(GC,Gated-Channel),
instance-of(PM,Particle-In-Motion)

the grammar will generate *A gated channel releases particles from the endoplasmic reticulum* but not e.g., *Particles releases a gated channel from the endoplasmic reticulum.*

4.2 Grammar Extraction

We extract our FB-LTAG with unification semantics from the KBGen training data in two main steps. First, we align the KB data with the input string. Second, we induce a Tree Adjoining Grammar augmented with a unification-based semantics from the aligned data.

[2] http://www.kbgen.org
[3] http://www.ai.sri.com/halo/halobook2010/exported-kb/biokb.html

4.2.1 Alignment

Given a Sentence/Input pair (S, I) provided by the KBGen Challenge, the alignment procedure associates each entity and event variable in I to a substring in S. To do this, we use the entity and the event lexicon provided by the KBGen organiser. The event lexicon maps event types to verbs, their inflected forms and nominalizations while the entity lexicon maps entity types to a noun and its plural form. For instance, the lexicon entries for the event and entity types shown in Figure 1 are as shown in Figure 3.

For each entity and each event variable V in I, we retrieve the corresponding type (e.g., `Particle-In-Motion` for `Particle-In-Motion64582`); search the KBGen lexicon for the corresponding phrases (e.g., *molecule in motion,molecules in motion*); and associate V with the phrase in S which matches one of these phrases. Figure 3 shows an example lexicon and the resulting alignment obtained for the scenario shown in Figure 1. Note that there is not always an exact match between the phrase associated in the KBGen lexicon with a type and the phrase occurring in the training sentence. To account for this, we use some additional similarity based heuristics to identify the phrase in the input string that is most likely to be associated with a variable lacking an exact match in the input string. E.g., for entity variables (e.g., `Particle-In-Motion64582`), we search the input string for nouns (e.g., particles) whose overlap with the variable type (e.g., Particle-In-Motion) is not empty.

`Particle-In-Motion`	molecule in motion,molecules in motion
`Endoplasmic-Reticulum`	endoplasmic reticulum,endoplasmic reticulum
`Gated-Channel`	gated Channel,gated Channels
`Release-Of-Calcium`	releases,release,released,release

The function of a (gated channel, `Gated-Channel64605`) is to (release, `Release-Of-Calcium646`) (particles, `Particle-In-Motion64582`) from the (endoplasmic reticulum, `Endoplasmic-Reticulum64603`)

Figure 3: Example Entries from the KBGen Lexicon and example alignment

4.2.2 Inducing a based FB-LTAG from the aligned data

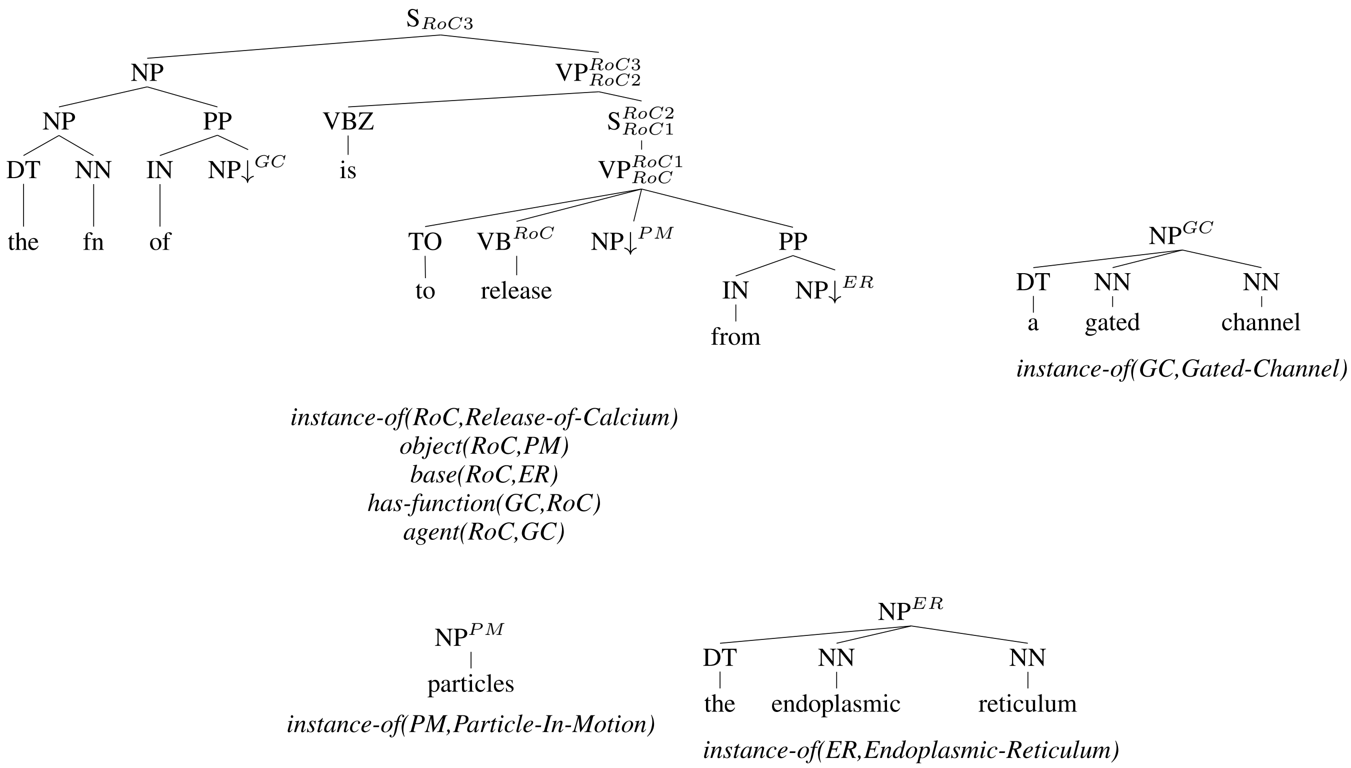


Figure 4: Extracted Grammar for "*The function of a gated channel is to release particles from the endoplasmic reticulum*". Variable names have been abbreviated and the KBGen tuple notation converted to terms so as to fit the input format expected by our surface realiser.

To extract a Feature-Based Lexicalised Tree Adjoining Grammar (FB-LTAG) from the KBGen data, we parse the sentences of the training corpus; project the entity and event variables to the syntactic projection of the strings they are aligned with; and extract the elementary trees of the resulting FB-LTAG from the parse tree using semantic information. Figure 4 shows the trees extracted from the scenario given in Figure 1.

To associate each training example sentence with a syntactic parse, we use the Stanford parser. After alignment, the entity and event variables occurring in the input semantics are associated with substrings of the yield of the syntactic parse tree. We project these variables up the syntactic tree to reflect headedness. A variable aligned with a noun is projected to the NP level or to the immediately dominating PP if it occurs in the subtree dominated by the leftmost daughter of that PP. A variable aligned with a verb is projected to the first S node immediately dominating that verb or, in the case of a predicative sentence, to the root of that sentence[4].

Once entity and event variables have been projected up the parse trees, we extract elementary FB-LTAG trees and their semantics from the input scenario as follows.

First, the subtrees whose root node is indexed with an entity variable are extracted. This results in a set of NP and PP trees anchored with entity names and associated with the predication true of the indexing variable.

Second, the subtrees capturing relations between variables are extracted. To perform this extraction, each input variable X is associated with a set of dependent variables i.e., the set of variables Y such that X is related to Y ($R(X,Y)$). The minimal tree containing all and only the dependent variables $D(X)$ of a variable X is then extracted and associated with the set of literals Φ such that $\Phi = \{R(Y,Z) \mid (Y = X \wedge Z \in D(X)) \vee (Y, Z \in D(X))\}$. This procedure extracts the subtrees relating the argument variables of a semantic functors such as an event or a role e.g., a tree describing a verb and its arguments as shown in the top part of Figure 4. Note that such a tree may capture a verb occurring in a relative or a subordinate clause (together with its arguments) thus allowing for complex sentences including a relative or relating a main and a subordinate clause.

The resulting grammar extracted from the parse trees (cf. e.g., Figure 4) is a Feature-Based Tree Adjoining Grammar with a Unification-based compositional semantics as described in Gardent and Kallmeyer [2003]. In particular, our grammars differs from the traditional probabilistic Tree Adjoining Grammar extracted as described in e.g., Chiang [2000] in that they encode both syntax and semantics rather than just syntax. They also differ from the semantic FB-TAG extracted by DeVault et al. [2008] in that (i) they encode the linking between syntactic and semantic arguments; (ii) they allow for elementary trees spanning discontiguous strings (e.g., *The function of X is to release Y*); and (iii) they enforce the semantic principle underlying TAG namely that an elementary tree containing a syntactic functor also contains its syntactic arguments.

4.3 Generation

To generate with the grammar extracted from the KBGen data, we use the GenI surface realiser Gardent et al. [2007]. Briefly, given an input semantics and a FB-LTAG with a unification based semantics, GenI selects all grammar entries whose semantics subsumes the input semantics; combines these entries using the FB-LTAG combination operations (i.e., adjunction and substitution); and outputs the yield of all derived trees which are syntactically complete and whose semantics is the input semantics. To rank the generator output, we train a language model on the GeniA corpus [5], a corpus of 2000 MEDLINE asbtracts about biology containing more than 400000 words Kim et al. [2003] and use this model to rank the generated sentences by decreasing probability.

Thus for instance, given the input semantics shown in Figure 1 and the grammar depicted in Figure 4, the surface realiser will select all of these trees; combine them using FB-LTAG substitution operation; and output as generated sentence the yield of the resulting derived tree namely the sentence *The function of a gated channel is to release particles from the endoplasmic reticulum.*

However, this procedure only works if the entries necessary to generate from the given input are present in the grammar. To handle new, unseen input, we proceed in two ways. First, we try to guess a grammar entry from the shape of the input and the existing grammar. Second, we expand the grammar by decomposing the extracted trees into simpler ones.

[4]Initially, we used the head information provided by the Stanford parser. In practice however, we found that the heuristics we defined to project semantic variables to the corresponding syntactic projection were more accurate and better supported our grammar extraction process.

[5]`http://www.nactem.ac.uk/genia/`

4.4 Guessing new grammar entries.

Given the limited size of the training data, it is often the case that input from the test data will have no matching grammar unit. To handle such previously unseen input, we start by partitioning the input semantics into sub-semantics corresponding to events, entities and role.

For each entity variable X of type $Type$, we create a default NP tree whose semantics is a literal of the form *instance-of(X,Type)*.

For event variables, we search the lexicon for an entry with a matching or similar semantics i.e., an entry with the same number and same type of literals (literals with same arity and with identical relations). When one is found, a grammar entry is constructed for the unseen event variable by substituting the event type of the matching entry with the type of the event variable. For instance, given the input semantics *instance-of(C,Carry), object(C,X), base(C,Y), has-function(Z,C), agent(C,Z)*, this procedure will create a grammar entry identical to that shown at the top of Figure 4 except that the event type *Release-of-Calcium* is changed to *Carry* and the terminal *release* to the word form associated in the KBGen lexicon with this concept, namely to the verb *carry*.

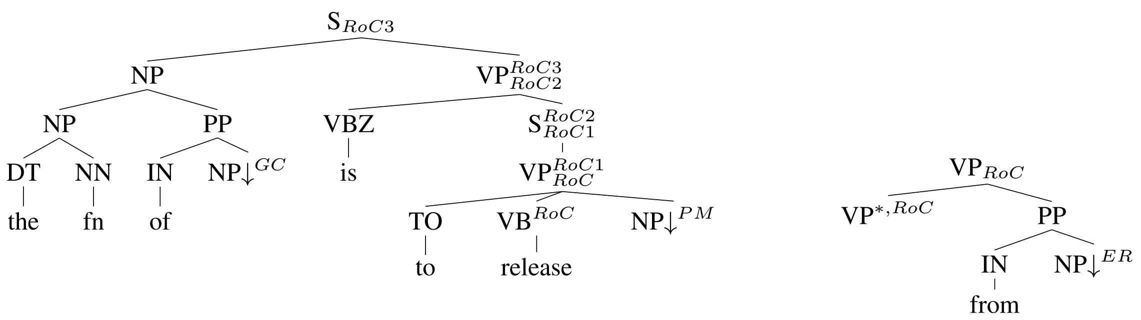
4.5 Expanding the Grammar

While the extracted grammar nicely captures predicate/argument dependencies, it is very specific to the items seen in the training data. To reduce overfitting, we generalise the extracted grammar by extracting from each event tree, subtrees that capture structures with fewer arguments and optional modifiers.

For each event tree τ extracted from the training data which contains a subject-verb-object subtree τ', we add τ' to the grammar and associate it with the semantics of τ minus the relations associated with the arguments that have been removed. For instance, given the extracted tree for the sentence *"Aquaporin facilitates the movement of water molecules through hydrophilic channels."*, this procedure will construct a new grammar tree corresponding to the subphrase *"Aquaporin facilitates the movement of water molecules"*.

We also construct both simpler event trees and optional modifiers trees by extracting from event trees, PP trees which are associated with a relational semantics. For instance, given the tree shown in Figure 4, the PP tree associated with the relation *base(RoC,ET)* is removed thus creating two new trees as illustrated in Figure 5: an S tree corresponding to the sentence *The function of a gated channel is to release particles* and an auxiliary PP tree corresponding to the phrase *from the endoplasmic reticulum*. Similarly in the above example, a PP tree corresponding to the phrase *"through hydrophilic channels."* will be extracted.

As with the base grammar, missing grammar entries are guessed from the expanded grammar. However we do this only in cases where a correct grammar entry cannot be guessed from the base grammar.



instance-of(RoC,Release-of-Calcium)
object(RoC,PM)
has-function(GC,RoC)
agent(RoC,GC)

base(RoC,ER)

Figure 5: Trees Added by the Expansion Process

5 Experimental Setup

We evaluate our approach on the KBGen data and compare it with the KBGen reference and two other systems having taken part to the KBGen challenge.

5.1 Training and test data.

Following a practice introduced by Angeli et al. [2010], we use the term *scenario* to denote a KB subset paired with a sentence. The KBGen benchmark contains 207 scenarii for training and 72 for testing. Each KB subset consists of a set of triples and each scenario contains on average 16 triples and 17 words.

5.2 Systems

We evaluate three configurations of our approach on the KBGen test data: one without grammar expansion (BASE); a second with a manual grammar expansion MANEXP; and a third one with automated grammar expansion AUTEXP. We compare the results obtained with those obtained by two other systems participating in the KBGen challenge, namely the UDEL system, a symbolic rule based system developed by a group of students at the University of Delaware; and the IMS system, a statistical system using a probabilistic grammar induced from the training data.

5.3 Metrics.

We evaluate system output automatically, using the BLEU-4 modified precision score Papineni et al. [2002] with the human written sentences as reference. We also report results from a human based evaluation. In this evaluation, participants were asked to rate sentences along three dimensions: **fluency** (Is the text easy to read?), **grammaticality** and meaning similarity or **adequacy** (Does the meaning conveyed by the generated sentence correspond to the meaning conveyed by the reference sentence?). The evaluation was done on line using the LG-Eval toolkit Kow and Belz [2012], subjects used a sliding scale from -50 to +50 and a Latin Square Experimental Design was used to ensure that each evaluator sees the same number of outputs from each system and for each test set item. 12 subjects participated in the evaluation and 3 judgments were collected for each output.

6 Results and Discussion

System	All	Covered	Coverage	# Trees
IMS	0.12	0.12	100%	
UDEL	0.32	0.32	100%	
Base	0.04	0.39	30.5%	371
ManExp	0.28	0.34	83 %	412
AutExp	0.29	0.29	100%	477

Figure 6: BLEU scores and Grammar Size (Number of Elementary TAG trees

Table 6 summarises the results of the automatic evaluation and shows the size (number of elementary TAG trees) of the grammars extracted from the KBGen data.

	Fluency		Grammaticality		Meaning Similarity	
System	Mean	Homogeneous Subsets	Mean	Homogeneous Subsets	Mean	Homogeneous Subsets
UDEL	4.36	A	4.48	A	3.69	A
AutExp	3.45	B	3.55	B	3.65	A
IMS	1.91	C	2.05	C	1.31	B

Figure 7: Human Evaluation Results on a scale of 0 to 5. Homogeneous subsets are determined using Tukey's Post Hoc Test with $p < 0.05$

The average BLEU score is given with respect to all input (All) and to those inputs for which the systems generate at least one sentence (Covered). While both the IMS and the UDEL system have full coverage, our BASE system strongly undergenerates failing to account for 69.5% of the test data. However, because the extracted grammar is linguistically principled and relatively compact, it is possible to manually edit it. Indeed, the MANEXP results show that, by adding 41 trees to the grammar, coverage can be increased by 52.5 points reaching a coverage of 83%. Finally, the AUTEXP results demonstrate that the automated expansion mechanism permits achieving full coverage while keeping a relative small grammar (477 trees).

In terms of BLEU score, the best version of our system (AUTEXP) outperforms the probabilistic approach of IMS by a large margin (+0.17) and produces results similar to the fully handcrafted UDEL system (-0.03).

In sum, our approach permits obtaining BLEU scores and a coverage which are similar to that obtained by a hand crafted system and outperforms a probabilistic approach. One key feature of our approach is that the grammar extracted from the training data is linguistically principled in that it obeys the extended locality principle of Tree Adjoining Grammars. As a result, the extracted grammar is compact and can be manually modified to fit the need of an application as shown by the good results obtained when using the MANEXP configuration.

We now turn to the results of the human evaluation. Table 7 summarises the results whereby systems are grouped by letters when there is no significant difference between them (significance level: $p < 0.05$). We used ANOVAs and post-hoc Tukey tests to test for significance. The differences between systems are statistically significant throughout except for meaning similarity (adequacy) where UDEL and our system are on the same level. Across the metrics, our system consistently ranks second behind the symbolic, UDEL system and before the statistical IMS one thus confirming the ranking based on BLEU.

7 Conclusion

In Tree Adjoining Grammar, the *extended domain of locality principle* ensures that TAG trees group together in a single structure a syntactic predicate and its arguments. Moreover, the *semantic principle* requires that each elementary tree captures a single semantic unit. Together these two principles ensure that TAG elementary trees capture basic semantic units and their dependencies. In this paper, we presented a grammar extraction approach which ensures that extracted grammars comply with these two basic TAG principles. Using the KBGen benchmark, we then showed that the resulting induced FB-LTAG compares favorably with competing symbolic and statistical approaches when used to generate from knowledge base data.

In the current version of the generator, the output is ranked using a simple language model trained on the GENIA corpus. We observed that this often fails to return the best output in terms of BLEU score, fluency, grammaticality and/or meaning. In the future, we plan to remedy this using a ranking approach such as proposed in Velldal and Oepen [2006], White and Rajkumar [2009].

References

Eva Banik, Claire Gardent, Donia Scott, Nikhil Dinesh, and Fennie Liang. Kbgen: Text generation from knowledge bases as a new shared task. In *Proceedings of the seventh International Natural Language Generation Conference*, pages 141–145. Association for Computational Linguistics, 2012.

Eva Banik, Claire Gardent, Eric Kow, et al. The kbgen challenge. In *Proceedings of the 14th European Workshop on Natural Language Generation (ENLG)*, pages 94–97, 2013.

Juen-tin Wang. On computational sentence generation from logical form. In *Proceedings of the 8th conference on Computational linguistics*, pages 405–411. Association for Computational Linguistics, 1980.

Stuart M Shieber, Gertjan Van Noord, Fernando CN Pereira, and Robert C Moore. Semantic-head-driven generation. *Computational Linguistics*, 16(1):30–42, 1990.

Martin Kay. Chart generation. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 200–204. Association for Computational Linguistics, 1996.

Hadar Shemtov. Generation of paraphrases from ambiguous logical forms. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 919–924. Association for Computational Linguistics, 1996.

K. Kaljurand and N.E. Fuchs. Verbalizing owl in attempto controlled english. *Proceedings of OWLED07*, 2007.

G. Wilcock. Talking owls: Towards an ontology verbalizer. *Human Language Technology for the Semantic Web and Web Services, ISWC*, 3:109–112, 2003.

R. Power and A. Third. Expressing owl axioms by english sentences: dubious in theory, feasible in practice. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1006–1013. Association for Computational Linguistics, 2010.

K. Bontcheva and Y. Wilks. Automatic report generation from ontologies: the miakt approach. In *Ninth International Conference on Applications of Natural Language to Information Systems (NLDB'2004)*. Lecture Notes in Computer Science 3136, Springer, Manchester, UK, 2004.

G. Aguado, A. Bañón, J. Bateman, S. Bernardos, M. Fernández, A. Gómez-Pérez, E. Nieto, A. Olalla, R. Plaza, and A. Sánchez. Ontogeneration: Reusing domain and linguistic ontologies for spanish text generation. In *Workshop on Applications of Ontologies and Problem Solving Methods, ECAI*, volume 98, 1998.

C.L. Paris. Tailoring object descriptions to a user's level of expertise. *Computational Linguistics*, 14(3):64–78, 1988.

VO Mittal, G. Carenini, and JD Moore. Generating patient specific explanations in migraine. In *Proceedings of the eighteenth annual symposium on computer applications in medical care. McGraw-Hill Inc*, 1994.

E. Reiter, R. Robertson, and L.M. Osman. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1):41–58, 2003.

D. Galanis, G. Karakatsiotis, G. Lampouras, and I. Androutsopoulos. An open-source natural language generator for owl ontologies and its use in protégé and second life. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, pages 17–20. Association for Computational Linguistics, 2009.

Sandra Williams and Richard Power. Grouping axioms for more coherent ontology descriptions. In *Proceedings of the 6th International Natural Language Generation Conference (INLG 2010)*, pages 197–202, Dublin, 2010.

David DeVault, David Traum, and Ron Artstein. Making grammar-based generation easier to deploy in dialogue systems. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 198–207. Association for Computational Linguistics, 2008.

David Chiang. Statistical parsing with an automatically-extracted tree adjoining grammar. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 456–463. Association for Computational Linguistics, 2000.

Wei Lu and Hwee Tou Ng. A probabilistic forest-to-string model for language generation from typed lambda calculus expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1611–1622. Association for Computational Linguistics, 2011.

Yuk Wah Wong and Raymond J Mooney. Generation by inverting a semantic parser that uses statistical machine translation. In *HLT-NAACL*, pages 172–179, 2007.

Wei Lu, Hwee Tou Ng, and Wee Sun Lee. Natural language generation with tree conditional random fields. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 400–409. Association for Computational Linguistics, 2009.

Percy Liang, Michael I Jordan, and Dan Klein. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 91–99. Association for Computational Linguistics, 2009.

Deborah A Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. Expanding the scope of the atis task: The atis-3 corpus. In *Proceedings of the workshop on Human Language Technology*, pages 43–48. Association for Computational Linguistics, 1994.

Gabor Angeli, Percy Liang, and Dan Klein. A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 502–512. Association for Computational Linguistics, 2010.

Joohyun Kim and Raymond J Mooney. Generative alignment and semantic parsing for learning from ambiguous supervision. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 543–551. Association for Computational Linguistics, 2010.

Ioannis Konstas and Mirella Lapata. Unsupervised concept-to-text generation with hypergraphs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 752–761. Association for Computational Linguistics, 2012a.

Ioannis Konstas and Mirella Lapata. Concept-to-text generation via discriminative reranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 369–378. Association for Computational Linguistics, 2012b.

D. Gunning, V. K. Chaudhri, P. Clark, K. Barker, Shaw-Yi Chaw, M. Greaves, B. Grosof, A. Leung, D. McDonald, S. Mishra, J. Pacheco, B. Porter, A. Spaulding, D. Tecuci, and J. Tien. Project halo update - progress toward digital aristotle. *AI Magazine*, Fall:33–58, 2010.

Peter Clark and Bruce Porter. Building concept representations from reusable components. In *AAAI/IAAI*, pages 369–376. Citeseer, 1997.

B. Grosof. The silk project: Semantic inferencing on large knowledge. Technical report, SRI, 2012. `http://silk.semwebcentral.org/`.

Boris Motik, Peter F Patel-Schneider, Bijan Parsia, Conrad Bock, Achille Fokoue, Peter Haase, Rinke Hoekstra, Ian Horrocks, Alan Ruttenberg, Uli Sattler, et al. Owl 2 web ontology language: Structural specification and functional-style syntax. *W3C recommendation*, 27:17, 2009.

K. Vijay-Shanker and AK Joshi. Feature structures based tree adjoining grammars. In *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, Hungary, 1988.

Claire Gardent and Laura Kallmeyer. Semantic construction in feature-based tag. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 123–130. Association for Computational Linguistics, 2003.

Claire Gardent, Eric Kow, et al. A symbolic approach to near-deterministic surface realisation using tree adjoining grammar. In *ACL*, volume 7, pages 328–335, 2007.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182, 2003.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

Eric Kow and Anja Belz. Lg-eval: A toolkit for creating online language evaluation experiments. In *LREC*, pages 4033–4037, 2012.

Erik Velldal and Stephan Oepen. Statistical ranking in tactical generation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 517–525. Association for Computational Linguistics, 2006.

Michael White and Rajakrishnan Rajkumar. Perceptron reranking for ccg realization. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 410–419. Association for Computational Linguistics, 2009.

Part XVII

Selected contribution in the field of Machine-learning-based Natural Language Processing

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

From Task-Oriented Dialogue to Conversational Question Answering

Lina M. Rojas-Barahona
Orange Innovation
2 Avenue Pierre Marzin, 22300 Lannion
linamaria.rojasbarahona@orange.com

Claire Gardent
CNRS-LORIA, UMR 7503
Vandoeuvre-les-Nancy F-54500, France
claire.gardent@loria.fr

1 Introduction

I got the privilege to work with Claire Gardent, first as post-doctorate, later as part of a research collaboration, formalised in the Innovative Training Networks (ITN) European project, Marie Skłodowska-Curie: NL4XAI[1] involving partners such as Orange (where I am working) and academics such as the CNRS (Claire's affiliation). This work summarises some selected publications of joint work with Claire. It covers task-oriented dialogue in serious games and conversational question answering grounded in knowledge graphs (KGs). While task-oriented dialogues are usually domain specific and concentrated to solve a quite simple task: provide information to the user, Conversational Question Answering (ConvQA) covers general knowledge. It consist of pairs of questions-answers related to either a paragraph in a text or to a knowledge graph. However, task-oriented dialogues usually offer richer interactions such as confirmation, corrections, inquiries,etc; while ConvQA support a simpler interaction in a teacher-student scenario: one user always asks questions, the other always answers.

On the one hand, for task-oriented dialogue we not only collected the French Emospeech dialogue corpus, a dialogue corpus in a serious game. We also define the annotation scheme and implemented state of the art models at that time: Logistic Regresion (LR) and Support Vector Machines (SVM) for the task of Natural Language Understanding (NLU). We will describe in Section 2 briefly this work that gave origin to four publications [Rojas-Barahona et al., 2012a,b, Rojas-Barahona and Gardent, 2012, Gardent and Rojas-Barahona, 2013][2]. On the other hand, for ConvQA our contribution[3] was to enrich existing datasets with information about the ellipsis and coreferences [Brabant et al., 2021] presented in Section 3.1 and to create of a new dataset KB-Conv, that we will soon made public (Section 3.3).

2 Dialogues in a Serious Game

The French Emospeech corpus gathers Human-Human dialog data, that was collected through Wizard-of-OZ (WOZ) experiments in 2011 [Rojas-Barahona et al., 2012a]. The serious game is a multiplayer quest where the players (3 teenagers) seek to build a video game joystick in order to free their uncle trapped in the game. To build this joystick, the players must explore a factory and achieve 17 mandatory goals (find the plans, get the appropriate mould, retrieve some raw material from the storing shed, etc). In addition, they can increase their score by achieving optional goals which, when reached, provide them with extra information about the industry (therefore increasing their knowledge). In total, the players can achieve up to 28 goals by conducting 12 separate subdialogs in various parts of the virtual world. That is, dialogs in the game are long dialogs involving multiple players in various settings.

Table 1 summarises the characteristics of the subdialogs conducted within the game highlighting three distinguishing features of game dialogs. First, the dialog participants vary whereby both the game agent and the player can change. Thus in the game, the player alternatively plays any of the three children involved in the quest while the game agent is

[1] https://nl4xai.eu/
[2] This is a joint work with the researcher Alejandra Lorenzo.
[3] This part is mainly the work of Quentin Brabant a young researcher at Orange under my supervision and the supervision of Claire Gardent.

successively, Lucas, M. Jasper, Samir, Melissa, an operator, Serge and Sophia. Second, game dialogues are task-driven whereby each sub-dialogue is related to a step in the game and each dialog turn aims to achieve a game goal and improve the player score. Third, the context in which each sub-dialogue takes place varies as the player moves around the world.

NLU annotation schema for dialogue is not necessarily dictated by speech act theory alone but might also consider more practical issues namely, how well it will support interpretation and/or dialogue. To enhance learning, the annotation schema designed for the game combines core communicative acts [Bunt et al., 2010] with domain specific information. The domain specific information specifies the goals being pursued/discussed/achieved etc. while the communicative act can be viewed as specifying how the current information state is updated by the speaker's utterance.

Id	VC	Player	Mandatory Goals	Location
1	Lucas	Ben	Find the address of the enterprise.	Unlce's place.
2	M.Jasper	Lucas	The manufacturing first step	Enterprise reception
3	Samir	Julie	Find the plans of the joystick	Designing Office
4	Samir	Julie	Find out what to do next	Designing Office
5	Melissa	Lucas	Manufacturing process ...	Plant
6	Melissa	Lucas	Find the right machine	Plant
7	Melissa	Lucas	Find out what to do next	Plant
8	Operator	Julie	Knowing about the material space ...	Material Space
9	Serge	Ben	Perform quality tests	Laboratory Tests
10	Serge	Ben	Find out what to do next	Laboratory Tests
11	Sophia	Julie	Find the electronic components.	Finishing
12	Sophia	Julie	Finishing process	Finishing

Table 1: Description of the subdialogs in the MP Game.

dialog: 01_dialogDirecteur-Tue Jun 14 11:04:23 2011

M.Jasper: Bonjour, je suis M.Jasper le directeur.
Hello, I am the director, Mr. Jasper. →***greet***

M.Jasper: Qu'est-ce que je peux faire pour vous?
What can I do for you? → ***ask(task(X))***

Lucas:ie dois sauver mon oncle
I must rescue my uncle → ***first_step***

M.Jasper: Pour faire votre manette, il vous faut des plans. Allez voir dans le bureau d'études, ils devraient y être.
To build the joystick you will need the plans. You will find them in the Designing Office. → ***inform(do(first_step))***

M.Jasper: J'aurais aimé continuer à répondre à vos questions mais je dois reprendre mon travail! Bonne Chance!
I have to go back to work! Good Luck! → ***quit***

Figure 1: Excerpt from a dialogue in the EmoSpeech corpus. The corresponding user semantics is shown highlighted on the right.

The full list of Dialogue Acts used for annotation together with the corresponding dialog acts and a gloss of their meaning is described in [Rojas-Barahona et al., 2012a]. Labels used are very specific to the game to facilitate the integration within the game (same goals as defined in the serious game) and to bypass much of the pragmatic reasoning necessary to associate a dialog turn with a communicative function. For instance, in the dialog above, the turn *je dois sauver mon oncle (I must rescue my uncle)* does not explicitly state that the player (i) is seeking to achieve the game goal "rescueing one's uncle" and (ii) is asking the game agent for the first step towards achieving that goal.

2.1 Experimental setup

We experimented with both an SVM and an LR[4] classifier using different sets of features on different data sets with and without TF*IDF (term frequency*Inverse Document Frequency) filtering.

We compared a single classifier on the whole dataset (the whole game) against 12 distinct classifiers, one for each subdialog. In both cases the categories to be learned are restricted to the speaker's intent. Taking into account the game goals, the total number of categories to be learned is 27. When learning on subdialogs, the number of categories to be learned is smaller but so is the size of the training set. The features for the machine learning models were bag of words, in which stop words were filtered out, utterances were deaccented and converted to lower-case. In addition, we experimented with various context length using as features the 0 to 4 previous dialogue acts. Subdialog identifiers were also used when training the classifier on the whole dialogue. More details are given in [Rojas-Barahona et al., 2012a].

[4] We used MALLET [McCallum, 2002] for the LR classifier with L1 Regularisation.

	Whole Dialog		Subdialogs	
	w/o Tf*Idf	w/ Tf*Idf	w/o Tf*Idf	w/ Tf*Idf
LR	79.74	**90.26**	86.41	88.22
SVM	78.79	88.55	76.45	83.99

Table 2: Global Results for the Logistic Regression (LR), the SVM (SVM) and the SVM Classifier with Penalisation (SVM(P))

We also experimented using tf*idf filtering to limit the impact of frequent uninformative words. Moreover, we experimented penalising those categories with more training instances, since the data was highly skewed. Dialogue acts that relate to optional goals were often not followed up by the players resulting in data sparseness.

2.2 Results

Table 2 shows the results for the 6 main configurations: training on the whole dialog or on subdialogs, with and without tf*idf filtering and using LR and SVM. The best results are obtained using the LR classifier on the whole dataset with tf*idf filtering.

*Impact of the tf*idf filtering*. Globally, the tf*idf filtering has a positive impact leading to an increase in accuracy ranging from 2.81 to 11.52 points. For the SVM classifier, the tf*idf filtering consistently lead to better results. However, for the LR classifier the filtering adversely impacts performance on short subdialogs (6 and 7), where one unique goal is being discussed. We conjecture that for these cases, the tf*idf filtering removed words which helped the classifier distinguish between turns about the unique goal from other turns. SVM with penalisation yields worse results with the tf*idf filtering than without, thus suggesting overfitting. In the next section we present how can we exploiting synonyms to improve generalisation.

Impact of contextual features. Having a notion of context is crucial for correctly interpreting dialog acts. As mentioned above, we use the dialog acts of the previous turns to model context. However the further back we look into the previous turns, the more features there will be to train on. In other words, depending on the number of previous turns considered, the data to learn from will be more or less sparse. We experimented with 3 setups: a null context, the dialog moves of the two previous turns and the dialog moves of the four previous moves. Table 3 shows the results.

	Whole Dialog			Subdialogs		
	0	2	4	0	2	4
LR	88.43	90.26	90.26	84.43	87.59	88.22
SVM	84.36	86.76	88.55	78.04	82.06	83.99

Table 3: The impact of context on accuracy. 0,2 and 4 indicates that the context is captured by having as features the dialog moves of 0, 2 and 4 previous turns respectively

Impact of dialog acts. The accuracy varies per dialog acts from 48% to 99%. with most of the acts having an accuracy above 80%. Unsurprisingly, the acts with lowest accuracy are also those with fewest training data. The data is split randomly for the 30-fold evaluation with the risk of having insufficient data for optional goals.

2.3 Data Augmentation

This pioneer work on data augmentation was published in [Gardent and Rojas-Barahona, 2013]. We explored four ways of modifying the content features used for classification: lemmatising the training and the test data; augmenting the training data with automatically acquired paraphrases; and substituting unknown words with synonyms or its distributional neighbours at run-time.

For Lemmatisation, we used the French version of Treetagger [5] to lemmatise both the training and the test data. Lemmas without any filtering were used to train classifiers. We then compare performance with and without lemmatisation. As we shall see, the lemma and the POS tag provided by TreeTagger are also used to lookup synonym dictionaries and EuroWordNet when using synonym handling at run-time.

We were among the first to exploit automatically acquired paraphrases and to use these not only to increase the size of the training corpus but also to better balance it[6]. We proceed as follows.

[5] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[6] The Emospeech data is highly skewed with some classes being populated with many utterances and others with few.

First, we generated paraphrases using a pivot machine translation approach where each user utterance in the training corpus (around 3610 utterances) was translated into some target language and back into French. Using six different languages (English, Spanish, Italian, German, Chinese and Arabian), we generated around 38000 paraphrases. We used Google Translate API for translating.

Second, we eliminate from these paraphrases, words that are likely to be incorrect lexical translations by removing words with low normalised term frequency (< 0.001) across translations i.e., lexical translations given by few translations and/or translation systems. We then preprocessed the paraphrases in the same way the utterances of the initial training corpus were preprocessed i.e., utterances were unaccented, converted to lower-case and stop words were removed, the remaining words were filtered with TF*IDF. After preprocessing, duplicates were removed.

Third, we added the paraphrases to the training data seeking to improve the balance between dialog moves per dialog. The process to balance data was guided by the deviation of the category with lowest examples compared to the standard deviation. If the deviation is lower than the standard deviation then we add paraphrases by keeping as much as possible the data balanced after replacement. We invite the interested reader to find more details about the algorithm proposed for balancing data in the paper [Gardent and Rojas-Barahona, 2013].

Substituting Synonyms for Unknown Words A word is unknown, if it is a well-formed French word[7] and if it does not appear in the training corpus. When an unknown word w is detected in a player utterance at run-time, we search for a word w' which occurs in the training data and is either a synonym of w or a distributional neighbour. After disambiguation, we substitute the unknown word for the synonym.

H		Lemmatisation			
H-H	Orig.	Lemmas	+EWN	+DIC	+RI
Orig.	$65.70\% \pm 5.62$	$\mathbf{66.04\% \pm 6.49}$	$68.17\% \pm 6.98$	$67.92\% \pm 4.51$	$66.83\% \pm 5.92$
Parap.	$70.89\% \pm 6.45$	$\mathbf{74.31\% \pm 4.78}$*	$\mathbf{74.60\% \pm 5.99}$*	$\mathbf{73.07\% \pm 7.71}$*	$\mathbf{72.63\% \pm 5.82}$*
H-C	Orig.	Lemmas	+EWN	+DIC	+RI
Orig.	$59.71\% \pm 16.42$	$59.88\% \pm 7.19$	$61.14\% \pm 16.65$	$61.41\% \pm 16.59$	$60.75\% \pm 17.39$
Parap.	$59.82\% \pm 15.53$	$59.48\% \pm 14.02$	$\mathbf{61.70\% \pm 14.09}$*	$\mathbf{62.01\% \pm 14.37}$*	$61.16\% \pm 14.41$*

Table 4: Accuracy on the H-H and on the H-C corpus. The star denotes statistical significance with the Wilcoxon test ($p < 0.005$) used for the HH corpus and the McNemar test ($p < 0.005$) for the HC corpus.

To identify synonyms, we make use of two lexical resources namely, the French version of EuroWordNet (EWN) [Vossen, 1998], which includes 92833 synonyms, hyperonyms and hyponyms pairs, and a synonym lexicon for French (DIC) [8] which contains 38505 lemmas and 254149 synonym pairs. While words are categorised into Noun, Verbs and Adjectives in EWN, DIC contains no POS tag information.

To identify distributional neighbours, we constructed semantic word spaces for each subdialog in the EmoSpeech corpus[9] using random indexing (RI) on the training corpus expanded with paraphrases. Using the cosine measure as similarity metrics, we then retrieve for any unknown word w, the word w' which is most similar to w and which appear in the training corpus.

For lexical disambiguation, two methods are compared. We use the POS tag provided by TreeTagger. In this case, disambiguation is syntactic only. Or we pick the synonym with highest probability based on a trigram language model trained on the H-H corpus.

2.3.1 Results and Discussion

Table 4 summarises the results obtained in four main configurations: (i) with and without paraphrases; (ii) with and without synonym handling; (iii) with and without lemmatisation; and (iv) when combining lemmatisation with synonym handling. We also compare the results obtained when evaluating using 10-fold cross validation on the training data (H-H dialogs) vs. evaluating the performance of the system on H-C interactions.

[7] A word is determined to be a well-formed French word if it occurs in the LEFFF dictionary, a large-scale morphological and syntactic lexicon for French [Sagot, 2010]

[8] DICOSYN (http://elsap1.unicaen.fr/dicosyn.html).

[9] We also used *distributional semantics* from the Gigaword corpus but the results were poor probably because of the very different text genre and domains between the the Gigaword and the game.

Overall Impact The largest performance gain is obtained by a combination of the three techniques namely, data expansion, synonym handling and lemmatisation (+8.9 points for the cross-validation experiment and +2.3 for the H-C evaluation).

Impact of Lexical Substitution at Run Time We found that lexical resources are only useful when combined with lemmatisation. This is unsurprising since synonym dictionaries and EuroWordNet only contain lemmas. Indeed when distributional neighbours are used, lemmatisation has little impact (e.g., 65.11% usingdistributional neighbours without lemmatisation on the H-H corpus without paraphrases vs. 66.41% when using lemmatisation).

Another important issue when searching for a word synonym concerns lexical disambiguation: the synonym used to replace an unknown word should capture the meaning of that word in its given context. We tried using a language model trained on the training corpus to choose between synonym candidates (i.e., selecting the synonym yielding the highest sentence probability when substituting that synonym for the unknown word) but did not obtain a significant improvement. In contrast, it is noticeable that synonym handling has a higher impact when using EuroWordNet as a lexical resource. Since EuroWordNet contain categorial information while the synonym dictionaries we used do not, this suggests that the categorial disambiguation provided by TreeTagger helps identifying an appropriate synonym in EuroWordNet.

Finally, it is clear that the lexical resources used for this experiment are limited in coverage and quality. We observed in particular that some words which are very frequent in the training data (and thus which could be used to replace unknown words) do not occur in the synonym dictionaries. For instance when using paraphrases and dictionaries (fourth row and fourth column in Table 4) 50% of the unknown words were solved, 17% were illformed and 33% remained unsolved. To compensate this deficiency, we tried combining the three lexical resources in various ways (taking the union or combining them in a pipeline using the first resource that would yield a synonym). However the results did not improve and even in some cases worsened due probably to the insufficient lexical disambiguation. Interestingly, the results show that paraphrases always improves synonym handling presumably because it increases the size of the known vocabulary thereby increasing the possibility of finding a known synonym.

In sum, synonym handling helps most when (i) words are lemmatised and (ii) unknown words can be at least partially (i.e., using POS tag information) disambiguated. Moreover since data expansion increases the set of known words available as potential synonyms for unknown words, combining synonym handling with data expansion further improves accuracy.

Impact of Lemmatisation When evaluating using cross validation on the training corpus, lemmatisation increases accuracy by up to 3.42 points indicating that unseen word forms negatively impact accuracy. Noticeably however, lemmatisation has no significant impact when evaluating on the H-C corpus. This in turn suggests that the lower accuracy obtained on the H-C corpus results not from unseen word forms but from unseen lemmas.

Impact of Paraphrases On the H-H corpus, data expansion has no significant impact when used alone. However it yields an increase of up to 8.27 points and in fact, has a statistically significant impact, for all configurations involving lemmatisation. Thus, data expansion is best used in combination with lemmatisation and their combination permits creating better, more balanced and more general training data. On the H-C corpus however, the impact is negative or insignificant suggesting that the decrease in performance on the H-C corpus is due to content words that are new with respect to the training data i.e., content words for which neither a synonym nor a lemma can be found in the expanded training data.

While classifiers are routinely trained on dialog data to model the dialog management process, the impact of such basic factors as lemmatisation, automatic data expansion and synonym handling has remained largely unexplored. The empirical evaluation described here suggests that each of these factors can help improve performance but that the impact will vary depending on their combination and on the evaluation mode. Combining all three techniques yields the best results. We conjecture that there are two main reasons for this. First, synonym handling is best used in combination with POS tagging and lemmatisation because these supports partial lexical semantic disambiguation. Second, data expansion permits expanding the set of known words thereby increasing the possibility of finding a known synonym to replace an unknown word with.

3 Conversational Question Answering

Conversational Question Answering (QA) is a relatively recent area of research that groups reading comprehension, QA and dialogue. Typically, it consist in a sequence of questions and answers related to a paragraph or to a knowledge graph.

Our contribution in this field was (i) to enrich existing datasets with information about the ellipsis and coreferences (Section 3.1) and (ii) to create a new dataset KB-Conv, that we will soon made public (Section 3.3).

3.1 Detection of Ellipsis and Co-reference in conversational corpora

This work was published in [Brabant et al., 2021]. We make several contributions to the task of ellipsis and coreference detection in dialogue corpora. We create labelled data by enriching three existing dialogue datasets with annotations indicating whether a turn contains an ellipsis and/or a coreference. As these annotations are incomplete (not every turn can be automatically labelled), we draw on inferential relations between incompleteness, pronominalisation, ellipsis and coreference to both extend and complement these annotations. We then use these annotated data to train a classifier based on DistilBERT [Sanh et al., 2020], which assigns to each question in a dialogue two labels indicating whether it contains an ellipsis and/or a coreference. We also explore how active learning, multilabel approaches and fine-tuning can be used to train this model.

A *coreference* occurs when an entity is referred via two or more expressions in the same conversation. However, we are only interested in detecting a particular kind of coreference. In this paper, we say that a coreference happens in a turn if and only if (1) it contains an expression referring to an entity already mentioned in a previous turn and (2) this entity cannot be identified outside of the conversational context. The *resolution* of a coreference consists in replacing the referring expression by an unambiguous reference to the entity.

In linguistics, an *ellipsis* is the omission of one or several words from a clause that preserves the meaning in context. When a turn is not understandable without its context (i.e. without the conversation history), we call it *incomplete*. In this paper, we assume that any turn contains an ellipsis if and only if it is still incomplete after coreferences have been resolved. It follows from this definition that an incomplete sentence contains either a coreference, an ellipsis, or both.

We are mostly interested in detecting ellipsis and coreference for conversational question-answering. Therefore, a conversations is a sequence of alternating questions and answers that starts with a question and ends with an answer: $(q_1, a_1, q_2, a_2 \ldots, q_n, a_n)$. In many available conversational question answering datasets questions are sentences produced by humans (e.g. [Choi et al., 2018, Christmann et al., 2019, Elgohary et al., 2019, Quan et al., 2019, Reddy et al., 2019]), while answers are often given by an automated system, and often not in the form of a sentence. For this reason, we focus on ellipsis and coreference detection in questions. Moreover, we will sometimes use the term question to refer to turns that are not question per-say, but are produced by the user, and not by an automated system.

We propose a model whose purpose is to predict whether any given question q_i of a dialogue $(q_1, a_1, \ldots, q_n, a_n)$ contains an ellipsis and/or a coreference; since any dialogue turn can normally be understood based on the context of previous turns, our task can be seen as the classification of q_i with the given context $c = (q_1, a_1, \ldots, q_{i-1}, a_{i-1})$. We thus formulate our task as a 2-labels classification: for a given input question q_i and an input context c, output two values $(coref, ellipsis) \in \{0, 1\}^2$ where 1 denotes the presence of the phenomenon and 0 denotes its absence. We call *instance* of our task the couple formed by a question, and its context. An instance is *annotated* when it is associated with an annotation of the form $(coref, ellipsis)$.

3.1.1 Active learning (AL)

We use the following values for annotating the datasets: 1 for the presence of a phenomenon (positive class), 0 for its absence (negative class). Cases where no label is assigned are denoted by the value -1. Note that -1 does not denote a class, but only the absence of information about the actual class. We describe how we process each dataset in order to obtain train instances for our task.

ConvQuestions Christmann et al. [2019]. Many conversations of ConvQuestions are centered on the same entity; those conversations tend to be similar to each other, as they often have questions in common. In order to maximise the benefits of manual annotations, we created subsets of the original data containing exactly one conversation per topic entity. This resulted in train/dev/test sets containing respectively 905/330/335 questions in total. Based on these new sets, we created an instance of our task for each question (except the first one) of each conversation. Some of these conversations where manually annotated with $(coref, ellipsis)$ values. We obtained train/dev/test of 247/329/331 annotated instances.

GECOR Quan et al. [2019]. We create instances as follows. For each dialogue $(q_1, a_1, \ldots, q_n, a_n)$ in the GECOR dataset, each $i \in \{2, \ldots, n\}$, and each variant $q'_i \in \{q_i(e), q_i(r), q_i(c)\}$ of the question q_i: if q'_i is not empty, then we create the instance $((q_1, a_1, \ldots, a_{i-1}), q'_i)$ and annotate it with $(coref, ellipsis)$ values. Those values can sometimes be deduced by using the following rules:

- $q_i(e)$ contains an ellipsis;

Piece of dialogue from CANARD:

q_1	What is On the Sunday of Life?
$q_1(c)$	What is On the Sunday of Life?
a_1	In 1992, Delerium released On the Sunday of Life as an edition of 1,000 copies, complete with a deluxe gatefold sleeve.
q_2	Did it do well?
$q_2(c)$	Did Porcupine Tree, On the Sunday of Life do well?
a_2	On the Sunday of Life... had accumulated sales of more than 20,000 copies.
q_3	Was it rereleaesd?
$q_3(c)$	Was Porcupine Tree, On the Sunday of Life rereleaesd?

Corresponding instances of the task:

Context	Question	Coref	Ellipsis	Coref	Ellipsis	Incomp.	Pronoun
(q_1, a_1)	q_2	-1	-1	1	-1	1	1
(q_1, a_1)	$q_2(c)$	0	0	0	0	0	0
(q_1, a_1, q_2, a_2)	q_3	-1	-1	1	-1	1	1
(q_1, a_1, q_2, a_2)	$q_3(c)$	0	0	0	0	0	0

Table 5: Example of dialogue from CANARD and the corresponding instances of the task. Columns with gray headers show the result of label filling.

- $q_i(r)$ contains a coreference;
- $q_i(c)$ contains no ellipsis nor coreference;
- if $q_i(e) = q_i(r)$ we infer that both $q_i(e)$ and $q_i(r)$ contain an ellipsis and a coreference;
- if $q_i(e)$ is empty, we infer that q_i contains no ellipsis and thus $q_i(r)$ neither;
- if $q_i(r)$ is empty, we infer that q_i contains no coreference and thus $q_i(e)$ neither.

These rules are not sufficient to deduce ellipsis and coreference label values in all cases. By default, the value -1 is assigned.

CANARD Elgohary et al. [2019]. Instances were extracted similarly as from the GECOR dataset. The two main differences are: for each created dialogue, two variants (original and complete) of the last question are used. When the complete variant is used, we assign 0 to both $coref$ and $ellipsis$; otherwise, we assign -1. An example is given in Table 5.

At this point many labels are missing in the instances of the task. In particular, instances from CANARD do not contain any positive label. We addressed this issue via two approaches: multilabel learning and label filling.

Multilabel classification can be seen as a particular case of multitask learning, since a single model is trained on several binary classification tasks. One justification for using this approach is that the parameters can be shared among tasks, which has been shown in the literature to be empirically beneficial.

The 4-labels classification task considers the following labels: coreference, ellipsis, incompleteness, and pronoun detection. Formally, it means that annotations of the form $(coref, ellipsis)$ are replaced by annotations of the form $(coref, ellipsis, inc, pronoun)$. We used automatic pronoun detection to provide a 0 or 1 value to $pronoun$ in all questions. By default, the value of inc is set to -1, except for instances from CANARD where the value is known.

We then replace some of the -1 values by taking advantage of the logical dependencies between labels: a pronoun always indicates a coreference; incompleteness is either due to a coreference or an ellipsis; coreferences and ellipses always cause incompleteness. We therefore applied the following rules to each instance, in order:

1. if $pronoun = 1$ then $coref \leftarrow 1$,
2. if $coref = 1$ or $ellipsis = 1$ then $inc \leftarrow 1$,
3. if $coref = 0$ and $ellipsis = 0$ then $inc \leftarrow 0$,
4. if $inc = 0$ then $coref \leftarrow 0$ and $ellipsis \leftarrow 0$.

Remark that in some cases these rules are not sufficient to get rid of all unknown values. Such cases can be found in the examples of Table 5.

Active learning (AL) is a human-in-the-loop method that aims at maximizing the performance gains relatively to the number of manual annotations. It is especially interesting when few labeled data are available and only a small fraction of unlabeled data can be manually annotated in reasonable time. We apply several rounds of AL for labeling (separately) ellipses and coreferences. Each round consists in the following steps:

1. *Train and evaluate a model.* We use CANARD/GECOR as a training set. All CANARD instances that have already been manually annotated during previous rounds are included. The evaluation is done on ConvQuestions test set.
2. *Run the model on unlabeled data.* The model trained in step 1 associates a prediction $(coref^*, ellipsis^*)$ to each instance.
3. *Select a subset of unlabeled data.* We select the 50 CANARD dialogues on which the model display the least certainty. Since one dialogue is the source of several instances, we define the certainty of a dialogue as the average certainty of the corresponding instances. The certainty of the model (for a given label, on a given instance) is defined as the distance from 0.5 of the output corresponding to the predicted label value, i.e.: $|coref^* - 0.5|$ for coreference and $|ellipsis^* - 0.5|$ for ellipsis.
4. *Manually label the selected subset.* We label the selected dialogues (either for ellipsis or coreference). Labeled dialogues are used during training in the next loop.

We stop repeating these steps when the evaluation score stops increasing.

3.1.2 Experiments

We evaluate the following model variants.

- *Baseline*. The baseline is a DistilBert[Sanh et al., 2019] model trained on the 4-label classification task on CANARD/ GECOR.
- *Fine tuning only*. The model is fine-tuned on the 4-label classification task on the training set of ConvQuestions.
- *Baseline + AL.* The model is fine-tuned on the 4-label classification task on CANARD/GECOR, but labelled instances of CANARD are added via AL. Each round of AL adds 50 instances that are labelled for either coreference or ellipsis. We evaluate several versions of this variant: three versions use instances that were annotated for coreference via, respectively, 1, 2, and 3 rounds of AL. Three others versions use instances that were annotated for ellipsis via 1, 2, and 3 rounds.
- *Baseline + all AL.* Identical to baseline + AL, but using all annotations produced for coreference and ellipsis (3 rounds for each).
- *Baseline + all AL + fine tuning*. Identical to *Baseline + all AL.*, but training on CANARD/GECOR is followed by a fine-tuning step on the training set of ConvQuestions.
- *2-label variants*. We evaluate three of them. They are respectively identical to *baseline*, to *baseline + all AL*, and to *baseline + all AL + fine tuning*, with the difference that the model is trained on the 2-labels classification task.

We use GECOR and CANARD for training our models, while ConvQuestions is used for evaluation and fine tuning. In this way we can better assess how well the classifier behaves on unseen data, data that is different from the data the model was trained on. During training, labels with -1 value are simply ignored (no error is retro-propagated). During evaluation, we measure the recall, precision, and F-measure on ellipsis and coreference detection.

3.2 Results

The results are displayed in Table 6. Each line corresponds to a variant of the model.

Generally, the results show that coreference detection performs better than ellipsis detection. Moreover, by looking at lines 2 to 9 in the table, we see that AL is clearly beneficial; the *all AL labels* variant improves F1 scores for coreference and ellipsis detection by 10 and 13 points compared to the baseline. The same conclusion is drawn when comparing lines 11 and 12. The effects of training on 4 labels versus 2 are less clear: by comparing lines 2, 9, 10 to lines 11, 12, 13, we see that 4-labels variants perform roughly as well as their 2-labels counterparts on coreference detection. For ellipsis detection, they score significantly higher on F1 score when no fine tuning is applied, but the scores are too low

		Coreference			Ellipsis		
		P	R	F1	P	R	F1
1	fine tuning only	81	65	72	51	67	57
2	baseline	**97**	64	77	64	36	46
3	+ AL for ellipsis (1 round)	92	63	75	71	48	56
4	+ AL for ellipsis (2 rounds)	89	72	80	83	41	55
5	+ AL for ellipsis (3 rounds)	85	79	82	74	48	57
6	+ AL for coref. (1 round)	87	84	85	72	46	56
7	+ AL for coref. (2 rounds)	92	81	86	71	39	50
8	+ AL for coref. (3 rounds)	95	79	86	67	31	43
9	+ all AL labels	94	81	87	84	46	59
10	+ fine tuning	94	**93**	**94**	83	**71**	**77**
11	baseline, 2-labels variant	89	68	77	**100**	10	19
12	+ all AL labels	91	86	89	88	35	50
13	+ all AL labels + fine-tuning	94	**93**	93	84	70	76

Table 6: Results of the experiments. Scores are given as percentages.

to propose a meaningful interpretation. Fine tuning increases scores for both ellipsis and coreference detection; however the increase is way larger in the case of ellipsis. In fact, coreference detection arguably performs reasonably well without fine-tuning, contrary to ellipsis detection. A possible explanation is that the kinds of ellipses occurring in one dataset can be different from those occurring in another. In contrast, coreferences cover a narrower set of phenomena.

In addition to measuring performances, we looked at the output of the model on the test set: we noticed that coreferences due to pronouns use are well recognized, while, many false negatives correspond to cases where an entity is referred to via its type or function, as in: “To which continent does Germany belong? What size is the country?”.

3.3 A conversational QA corpus grounded in Wikidata

After the great success of ChatGPT that spread out non factual generative neural models to the big audience, guiding semantically these models to enable explainability and to reduce their typical errors: hallucinations, distortions, omissions and repetitions[Faille et al., 2021, Narayan et al., 2022, Nie et al., 2019] is an urgent need. We propose KBConv, a corpus of Conversational Question Answering (CQA) grounded on Wikidata[10] to constraint the generation with a Knowledge Graph (KG).

KBConv is composed of conversations between two participants, one that always asks questions and the other one that always answers based on facts. Thus, it contains sequences of question-answer pairs. The grounded sequences are composed of Wikidata triples of the form: (s, p, o), in which s is the subject, p is the property and o corresponds to the object of a fact belonging to a KG.

In total KBConv gathers 71K conversations (604K question-answer pairs in total), where each pair relates to an underlying fact from the public KG Wikidata[11]. Each conversation is focused on a given *root* entity. As illustrated by Table 7, the first question is directly about this root entity, while the next ones explore new facts about any entity discovered during the conversation (including the root entity itself). This corpus can be used for distinct tasks such as factual question generation, question rewriting as well as generation of sequence of questions and answers from a given Knowledge-graph or vice-versa.

4 Conclusions

This work summarised some selected publications of joint work with Claire Gardent related to dialogue and conversational QA. Regarding task-oriented dialogue we were pioneers in data augmentation thought back-translation and in distributional semantics, before the deep learning era. Regarding Conversational QA, we were interested in annotating important phenomena in conversations: ellipsis and co-reference and in creating a dataset grounded in a knowledge graph to control semantically the generation process, reducing common errors of end-to-end generative models such as: hallucinations, distortions and omissions.

[10] `https://www.wikidata.org/`

[11] `https://www.wikidata.org/`

#1		**Triple**	(*NGC 4833*, *part of*, *Milky Way*)
	Question variants	original	NGC 4833 is part of what astronomical object?
		subject	NGC 4833
		rewritten	NGC 4833 is part of what astronomical object?
		original	Where is NGC 4833 located?
		subject	NGC 4833
		rewritten	Where is NGC 4833 located?
		Answer	Milky Way
#2		**Triple**	(*NGC 4833*, *discoverer or inventor*, *Nicolas Louis de Lacaille*)
	Question variants	original	Who was behind the discovery of NGC 4833?
		subject	NGC 4833
		rewritten	Who was behind the discovery?
		original	What was the name of the discoverer of NGC 4833?
		subject	NGC 4833
		rewritten	Who discovered this object?
		original	Who found NGC 4833?
		subject	NGC 4833
		rewritten	Who found this object?
		Answer	Nicolas Louis de Lacaille
#3		**Triple**	(*Nicolas Louis de Lacaille*, *religion or worldview*, *Catholic Church*)
	Question variants	original	What was his religion?
		subject	his
		rewritten	What was his religion?
		original	What faith did he follow?
		subject	he
		rewritten	What faith did he follow?
		Answer	Catholic Church

Table 7: Excerpt of a question-answer conversation along with the related triples. The root entity is NGC 4833, from the theme "space object". The rewritten corresponds to the in-context question that has been automatically generated by a T5 model.

	entities	properties	triples	conv.	questions			
					train	dev	test	total
person	31671	327	71915	25918	184939	29352	11386	225677
country	2171	171	3475	703	5085	817	214	6116
ideology	1220	169	1677	450	3112	581	228	3921
space object	2586	116	6360	5961	0	0	50158	50158
molecular entity	17798	151	38314	23033	154511	24587	9531	188629
historical event	4695	189	7770	4972	35270	5684	2247	43201
food	2532	166	4012	2099	15050	2230	1011	18291
taxon	3190	215	5408	1902	0	0	16099	16099
with_unseen_properties	13651	404	24123	5558	0	0	51813	51813
whole dataset	63345	458	142691	70596	397967	63251	142687	603905

Table 8: For each theme, the table gives: the number of different entities and properties appearing in conversations, the number of conversations, and the number of questions for each split. Note that in the entities and properties columns, the "total" values are not the sum of the cells above; this is because some entities and properties appear in several themes.

References

Lina Maria Rojas-Barahona, Alejandra Lorenzo, and Claire Gardent. Building and exploiting a corpus of dialog interactions between french speaking virtual and human agents. In *The eighth international conference on Language Resources and Evaluation (LREC)*, pages 1428–1435, 2012a.

Lina Maria Rojas-Barahona, Alejandra Lorenzo, and Claire Gardent. An end-to-end evaluation of two situated dialog systems. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–19, 2012b.

Lina Maria Rojas-Barahona and Claire Gardent. What should i do now? supporting conversations in a serious game. In *SeineDial 2012-16th Workshop on the Semantics and Pragmatics of Dialogue*, 2012.

Claire Gardent and Lina Maria Rojas-Barahona. Using Paraphrases and Lexical Semantics to Improve the Accuracy and the Robustness of Supervised Models in Situated Dialogue Systems. In *Conference on Empirical Methods in Natural Language Processing*, pages 808–813, Seattle, United States, October 2013. SIGDAT, the Association for Computational Linguistics special interest group on linguistic data and corpus-based approaches to NLP. URL `https://hal.inria.fr/hal-00905405`.

Quentin Brabant, Lina Maria Rojas-Barahona, and Claire Gardent. Active Learning and Multi-label Classification for Ellipsis and Coreference Detection in Conversational Question-Answering. In *12th International Workshop on Spoken Dialog System Technology (IWSDS 2021)*, Singapour/Virtual, Singapore, November 2021. URL `https://hal.archives-ouvertes.fr/hal-03533906`.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. Towards an iso standard for dialogue act annotation. In *Proceedings of LREC 2010, the Seventh International Conference on Language Resources and Evaluation*, Malta, 2010.

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

Benoît Sagot. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, 2010. URL `http://hal.inria.fr/inria-00521242`.

Piek Vossen, editor. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA, 1998. ISBN 0-7923-5295-5.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*, February 2020. URL `http://arxiv.org/abs/1910.01108`. arXiv: 1910.01108.

E. Choi, H. He, M. Iyyer, M. Yatskar, W-T. Yih, Y. Choi, P. Liang, and L. Zettlemoyer. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi:10.18653/v1/D18-1241. URL `https://www.aclweb.org/anthology/D18-1241`.

P. Christmann, R. Saha Roy, A. Abujabal, J. Singh, and G. Weikum. Look before you Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 729–738, New York, NY, USA, November 2019. Association for Computing Machinery. ISBN 978-1-4503-6976-3. doi:10.1145/3357384.3358016. URL `https://doi.org/10.1145/3357384.3358016`.

A. Elgohary, D. Peskov, and J. Boyd-Graber. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5918–5924, Hong Kong, China, November 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-1605. URL `https://www.aclweb.org/anthology/D19-1605`.

J. Quan, D. Xiong, B. Webber, and C. Hu. GECOR: An End-to-End Generative Ellipsis and Co-reference Resolution Model for Task-Oriented Dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4547–4557, Hong Kong, China, November 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-1462. URL `https://www.aclweb.org/anthology/D19-1462`.

S. Reddy, D. Chen, and C. D. Manning. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, May 2019. ISSN 2307-387X. doi:10.1162/tacl_a_00266. URL `https://doi.org/10.1162/tacl_a_00266`.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Juliette Faille, Albert Gatt, and Claire Gardent. Entity-based semantic adequacy for data-to-text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1530–1540, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.findings-emnlp.132. URL `https://aclanthology.org/2021.findings-emnlp.132`.

Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins, and Mirella Lapata. A well-composed text is half done! composition sampling for diverse conditional generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1319–1339, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.94. URL `https://aclanthology.org/2022.acl-long.94`.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, 2019.

Part XVIII

Selected contributions in the field of Deep-learning-based Natural Language Processing

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

LEARNING HEALTH-BOTS FROM TRAINING DATA THAT WAS AUTOMATICALLY CREATED USING PARAPHRASE DETECTION AND EXPERT KNOWLEDGE

Anna Liednikova
ALIAE
Université de Lorraine
5 rue Jacques Villemaux
54000 Nancy
anna.liednikova@aliae.io

Philippe Jolivet
ALIAE
5 rue Jacques Villemaux
54000 Nancy
philippe.jolivet@aliae.io

Alexandre Durand-Salmon
ALIAE
5 rue Jacques Villemaux
54000 Nancy
alexandre.durand-salmon@aliae.io

Claire Gardent
615, rue du Jardin Botanique
54600 Villers-lès-Nancy
CNRS
claire.gardent@loria.fr

ABSTRACT

A key bottleneck for developing dialog models is the lack of adequate training data. Due to privacy issues, dialog data is even scarcer in the health domain. We propose a novel method for creating dialog corpora which we apply to create doctor-patient interaction data. We use this data to learn both a generation and a hybrid classification/retrieval model and find that the generation model consistently outperforms the hybrid model. We show that our data creation method has several advantages. Not only does it allow for the semi-automatic creation of large quantities of training data. It also provides a natural way of guiding learning and a novel method for assessing the quality of human-machine interactions.

1 Introduction

Current data-driven dialog models require large quantities of training data. Because of privacy issues, the situation is even worse in the health domain, where data is particularly scarce. In this work, we propose a novel method for automatically creating the training data necessary to learn a chatbot which can mimic a doctor in doctor-patient interactions. Specifically, we combine expert knowledge provided by physicians with automatic paraphrase extraction techniques. We first ask experts (physicians) to specify typical doctor-patient interactions occurring in the context of clinical studies when talking about the four main topics generally discussed in these studies namely, sleep, mood, anxiety, leisure. Formally, the specification takes the form of a dialog tree whose nodes are labelled with either an example doctor question or an example patient input. Each node in the tree is associated with a unique identifier, which can be viewed as a simple form of dialog state.

We then enrich this initial dialog data by extracting paraphrases for patient turns from an online forum.

This data generation method has several advantages. First, it allows for a straightforward integration of expert knowledge in data generation, model learning and model evaluation as we can use the dialog turn identifiers both to guide learning and to assess the model (by comparing the sequences it follows with the expert defined sequences). More generally, the association of each dialog turn with a dialog turn identifier which reflects its position in the dialog tree and the consistent use of this identifier during data creation, model learning and model evaluation allows for increased interpretability. Second, this method helps achieve good coverage, as we can ensure that the data does contain all possible dialog paths. This is not the case with Wizard-of-Oz (WoZ) and crowdsourcing data collections approaches, where the coverage of

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 165–174.

the possible dialog paths depends on the crowd-worker decisions and input. Third, by instantiating each dialog with different paraphrases, we can increase linguistic diversity, i.e., we can create dialogs that have the same structure but different wording.

In sum, our work makes the following contributions. We propose a novel method for creating training data for dialog models. We apply this method to create training data for a bot mimicking doctor-patient interaction in the context of clinical studies. We use the created data to learn a generation and a hybrid classification/retrieval dialog model, we show that the generation model generally outperforms the classification model, and we provide a detailed analysis of the models results using automatic metrics, human evaluation and qualitative analysis.

2 Related Work

Various methods have been proposed to facilitate the creation of training data for dialog. Previous work has explored WoZ experiments in which two humans interact based on some pre-defined scenario and the dialogs resulting from these interactions are collected [Green and Wei-Haas, 1985, El Asri et al., 2017] or crowdsourcing settings where workers provide continuations to incomplete dialogs [Wen et al., 2017]. Both approaches are time intensive. Crowdsourcing is also expensive while the human-human dialogs that are collected by both approaches may be very different from the human-machine interactions that should be learned to support efficient human-machine communication where, typically, chat messages are restricted in length. Other work has relied on already available dialog data or on question/answer pairs extracted from online forums [Wei et al., 2018, Lin et al., 2019, Xu et al., 2019]. In the health domain, however, such data is extremely scarce and difficult to obtain. When obtainable, it also requires extensive pre-processing due to anonymization constraints. Another line of research has been to acquire data through machine-machine simulations [Xu et al., 2019, Majumdar et al., 2019, Shah et al., 2018]. In particular, Majumdar et al. [2019] combines pre-defined dialog outlines with template-based verbalization of dialog turns to automatically create a synthetic dialog corpus. Our work is similar to Majumdar et al. [2019] but differs from it in two main ways. First, instead of using templates, we use automatically extracted paraphrases to enrich the initial dialogs. Second, we experiment with two dialog models to investigate how domain knowledge (in the form of dialog tree positional information) can best be exploited to guide learning and to support error analysis.

3 Creating Dialog Corpora

To create training data for the dialog bot, we start by collecting typical dialog outlines from an expert. We then extract paraphrases for the patient turns from a Health forum and filter out dialog interactions with low coherence.

Collecting an Initial Corpus from an Expert. Studies have shown that the closed questionnaires traditionally used in the context of clinical studies are ineffective in gathering correct and precise information about the patient status because the patients get used to the questions and routinely input the same answers from one interaction to the next. Our long-term goal is to develop a Human-Machine dialog system that would complement standard clinical questionnaires by regularly engaging the patient in a dialog about the questionnaire topics. Since our target users are chronic pain patients, it is more important to keep them engaged for a long period rather than getting all information at the first interaction. To create our dialog corpus, we asked a physician to formalize typical patient-doctor interactions occurring in the context of a clinical study in the form of a dialog tree describing which questions need to be asked and, for each question, which answers are possible. The interactions cover four domains namely, sleep, mood, anxiety and leisure activities and the dialog tree has 58 nodes. A fragment of the dialog tree created for the sleep domain is shown in Figure 1 on the left, and an example dialog for the SLEEP domain in the same figure on the right. We call the data collected from the expert D_{init}.

Extracting paraphrases. We extract paraphrases for the patient turns provided by the expert from the HealthBoard [1] forum in several steps as follows.

As patient turns are mostly assertive responses to the doctor questions, we start by filtering out questions from the forum data to keep only those utterances which are assertions[2]. To this end, we use a binary stacked Bi-LSTM classifier trained on the Switchboard dataset.

We then compare each patient turn in D_{init} together with its context (P, the preceding doctor turn) with the assertive utterances extracted from the forum. For each sequence $D + P$ of contextualised patient turns in D_{init} and each

[1] healthboards.com

[2] As noted by a reviewer, this is a simplification as in fact, users tend to formulate clarification and disambiguation questions. We leave this for future work.

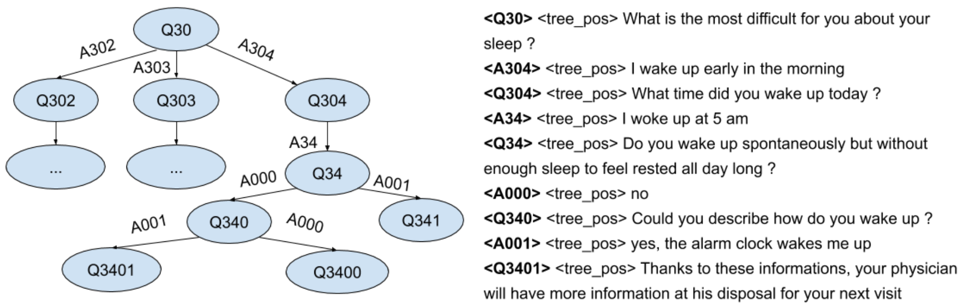


Figure 1: Fragment of dialog tree for the sleep domain and a corresponding dialog

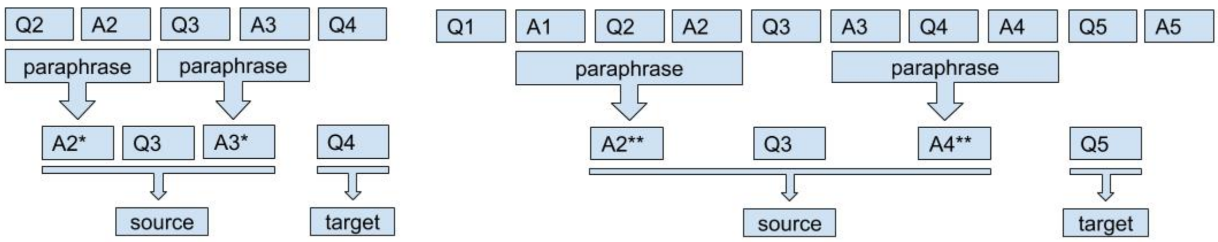


Figure 2: Paraphrasing source-target pairs from $INIT_{long}$ (left) and $INIT_{short}$ (right)

(assertive) utterance U in the forum, we create an S-BERT embedding (cf. Figure 2 left). We then retrieve from the forum all utterances U whose cosine similarity with a contextualized patient turn $D + P$ is higher than 0.70. Finally, we use Maximal Marginal Relevance (MMR) to select from this pool of candidates a subset of paraphrases which maximises both similarity (the paraphrases should be semantically similar to the input turn) and diversity (the resulting set of paraphrases should be maximally diverse[3]). We stop selecting sentences as soon as MMR score becomes negative, as a negative MMR score indicates that adding more paraphrases will not increase diversity.

As illustrated in Figure 2, we apply this paraphrase extraction process not only to create paraphrases for a single turn, but also to create paraphrases which summarise 3 consecutive turns. In this way, we can derive compressed versions of the initial dialogs. For instance, we can derive the short dialog in (3) from the longer dialog interaction shown in (3).

D1: Do you sleep well ?
P1: No
D2: What keeps you awake ?
P2: I have pain in the legs

D1: Do you sleep well ?
P1D2P2: No, I have pain in the legs and that keeps me awake.

We refer to the set of paraphrases that summarise three consecutive turns as SHORT and those that summarise a single turn as LONG.

Filtering Paraphrases. We compute cosine and BertScore on the S-BERT embeddings of each pair $\langle C, D\rangle$ of context-doctor interactions (where the context is the string concatenation of the three preceding turns) created in the previous step and keep only those pairs for which both scores are higher than the corresponding scores for the corresponding turn in the initial corpus (INIT).

Training Data. Table 1 summarises the training data we created. INIT is the dialog data collected from the expert; FORUM , the dataset obtained by replacing patient turns in INIT with their paraphrases, and ALL, the dataset left

[3]MMR is a measure for quantifying the extent to which a new item is both is similar to those already selected and similar to the target (here the patient turn). It is defined as: $Arg \max_{P_i \in C_U \setminus S} (\lambda\ Sim1(P_i, U) - (1-\lambda) \max_{P_j \in S} Sim2(P_i, P_j))$ where U is a contextualized user turn, C_U is a pool of candidate paraphrases for U, P_i; P_j are paraphrases in C_U, and S is the set of already selected paraphrases. A high λ value favors similarity. Conversely, a low λ value results in higher diversity. We set this parameter to be 0.5. We use BertScore recall as function $Sim1$ as this permits checking similarity on a word basis and cosine as function $Sim2$ since we do not need precise comparison between forum sentences, we just want them to be diverse.

	INIT	FORUM	ALL	LONG	SHORT
Nb of src-tgt pairs	388	4 010 696	733 104	373 220	359 884
Nb of distinct turns	483	60 346	28 734	9 761	19 027
Nb of tokens	18 180	204 309 290	37 705 130	19 546 597	18 158 533
Avg Nb of tokens per turn	13.81	12.726	12.858	13.09	12.61
Vocabulary size	426	13169	11314	10 631	9593
Cosine		0.51	0.52	0.5931	0.4456
BertScore		0.83	0.83	0.8446	0.8272

Table 1: Corpus statistics (INIT: dialog data collected from the expert; FORUM: extension of INIT with paraphrases; LONG: filtered FORUM dataset with only the single turn paraphrases; SHORT: filtered FORUM dataset with only the three-turn paraphrases; ALL = SHORT+LONG

after filtering. ALL is the combination of LONG and SHORT. As the table shows, the filtered dataset ALL is 5.5 times smaller but has similar coherence (identical or near identical cosine and BertScore scores) while retaining 86% of the vocabulary and 48% of the unique turns present in FORUM. To facilitate learning and reduce training time, we therefore use the filtered datasets in our experiments.

4 Health Bot Models

We aim to learn a model which mimics a physician in the kind of doctor-patient interaction that is typical of clinical studies conversations.

As we derive the training data from the dialog tree, each patient turn and each doctor query is associated with a dialog state (a node in that dialog tree). We use this dual information (dialog turn and dialog state) to train and compare two models for response generation: a classification model which, given the last three turns of a doctor-patient interaction, predicts a dialog state and outputs the corresponding doctor query; and a generative sequence-to-sequence model which auto-regressively generates an answer while conditioning on the last three dialog turns. For both models, we use a pre-training and fine-tuning approach similarly to that presented in Radford and Salimans [2018].

Classification model. Given a dialog context (3 dialog turns), the classification model predicts a dialog state and outputs the corresponding doctor query. Thus, the classification model is a multi-class classifier with 58 target classes, the 58 dialog states defined by the expert dialog tree. We use the PyTorch implementation of Radford and Salimans [2018]'s pre-training and fine-tuning approach provided by Huggingface[4] and the default hyper-parameter settings.

The input to the model consists of three turns $\langle p_1 d_1 p_2 \rangle$. We concatenate these three turns, prefixing each turn with its dialog state identifier and separating them with a delimiter token. Each token is represented by the sum of three embeddings: a word and a position embeddings which are learnt in the pre-training phase; and a turn embedding (learned during fine-tuning) indicating whether the token belongs to a patient or to a doctor turn. The input to the model is the sum of all three types – word, position and turn embeddings for each token in the input sequence.

The pre-trained model is the Generative Pre-trained Transformer-based (GPT-2) Language Model trained on the BooksCorpus dataset (7,000 books from different genres including Adventure, Fantasy, and Romance). The parameters are initialized to the smallest version of the GPT-2 model weights open-sourced by Radford and Salimans [2018].

We fine-tune the pretrained language model on our data by passing the input turns through the pre-trained model and feeding the final transformer block's activation to an added linear output layer followed by a softmax to predict a probability distribution over the target classes.

Generative model. To generate (rather than retrieve) doctor queries, we use the TransferTransfo[5] model Wolf et al. [2019] which combines a pretrained language model with a Transformer-based generation model fine-tuned on dialog data using multi-task learning. Multi-tasking combines a language modeling loss with a next turn classification loss. For the latter, the model is trained to distinguish a correct continuation from one randomly chosen distractor. As for the classification model, we use the GPT-2 language model pretrained on the BooksCorpus. For fine-tuning, we use the same augmented representations as for classification, i.e. each input consists of the three previous turns with a separator and a dialog state identifier between each turn. From this sequence of input tokens, a sequence of input embeddings for the Transformer is constructed by summing the word and positional embeddings learned during the pre-training phase

[4]https://github.com/huggingface/pytorch-openai-transformer-lm
[5]https://github.com/huggingface/transfer-learning-conv-ai

and the turn embeddings learned during fine-tuning. Multi-task learning is done, as in the TransferTransfo model, by jointly optimizing the language modeling and the next-turn classification loss.

5 Experiments

5.1 Data and Experimental Setting

We train our models on LONG, SHORT and ALL (cf. Table 1) using a 80/20 train/validation ratio.

We created test data for both long and short interactions by manually specifying six distinct paraphrases for each user turn ($TEST_{LONG}$) or 3 turn sequences ($TEST_{SHORT}$) in $INIT$. Paraphrasing the tree user turns permits capturing alternative formulations of the same content, thereby allowing for an evaluation that better takes into account the paraphrasing capacity of natural language. Models trained on the ALL dataset are evaluated on $TEST_{ALL}$ which is a concatenation of $TEST_{LONG}$ and $TEST_{SHORT}$. $TEST_{LONG}$ has 4248 source-target pairs and $TEST_{SHORT}$ 2172.

Both models are 12-layer decoder-only transformer with masked self-attention heads (768 dimensional states and 12 attention heads) a dropout probability of 0.1 on all layers (residual, embedding, and attention). They use learned positional embeddings with supported sequence lengths up to 512 tokens. The input sentences are pre-processed and tokenized using bytepair encoding (BPE) vocabulary with 40,000 merges Sennrich et al. [2016]. Relu activation function is used. CLASSIF is a transformer with a language modelling and a classification head on top, the two heads are two linear layers. The classification head has dropout of 0.1. The model was fine-tuned with a batch size of 8, using OpenAI Adam with a learning rate of 6.25e-5 and a linear learning rate decay schedule with warmup over 0.2% of training. λ was set to 0.5. The GEN model is a transformer with a language modelling and a multiple-choice classification head on top, the two heads are two linear layers. The model was fine-tuned with a batch size of 4, using AdamW with a learning rate of 6.25e-5, $\beta 1$ = 0.9, $\beta 2$ = 0.999, L2 weight decay of 0.01. The learning rate was linearly decayed to zero over the course of the training. For both models, we trained for 3 epochs using cross entropy loss.

5.2 Evaluation

We assess the output of our models using both automatic metrics and human evaluation.

Automatic Metrics. In our data, each dialog turn is associated with a node (or dialog state) in the initial dialog tree drawn by the expert. We use this dual information (dialog turn and dialog state) for the evaluation. We compute F1 on dialog state labels to analyse the coherence of the system response with the current dialog context (For the generative model, if no label was predicted, the score is 0). We also compute BLEU-4 and BertScore between the model output and the reference turn to assess the similarity of the generated output with the reference.

Human evaluation. We ask annotators, coming from the ALIAE company working on health bots and from academia, to interact with a bot which at each new user turn outputs the doctor query suggested by one of our two models. The annotators are instructed to input free-text answers to the chatbot queries, and the interaction stops when the bot repeats a previously output question or when the annotator outputs a closing turn ('Bye!').

To assess the quality of the bots response given the dialog context, annotators are required to score each system response on a 5 point Likert scale with respect to coherence ('Is the bot question coherent with the dialog so far ?') where 1 is totally incoherent and 5 is perfectly coherent. For the generation bot, we additionally ask the annotators to rate fluency ('Is the bot response well-formed ?') where 1 is unreadable and 5 is perfectly readable. The annotators are non native but their English is fluent. For each model (CLASSIF and GEN trained on LONG), we collect 50 dialogs from 20 annotators. Each annotator interacts at most 5 times with the bot.

We also evaluate the quality of the full dialogs resulting from these human-bot interaction. At the end of each human-bot conversation, the annotator is asked to rate satisfaction on a scale from 1 to 5. In addition, we applied the evaluation protocole proposed by Li et al. [2019]. Using the 50 dialog pairs collected for bot response evaluation, we show the annotators pairs of collected dialogs, one dialog from the generation model and the other from the classification model and ask them the questions recommended by the protocole: 'Who would you prefer to talk to for a long conversation?' 'If you had to say one of the speakers is interesting and one is boring, who would you say is more interesting?' 'Which speaker sounds more human?' 'Which speaker has more coherent responses in the conversation?'. For this task, we had 16 annotators annotating 50 dialog pairs. Each pair was rated 3 times except 2 pairs which were only rated twice. Each annotator annotated at most 10 dialog pairs.

We report the percentage of time one model was chosen over the other. We also compute the average user turn length (number of tokens), the average dialog length (number of turns) and the proportion of turn sequences of length at least two which occur in the dialog tree (Sequence Rate). By assessing how often the bot reproduces a sequence of dialog states that is present in the expert dialog tree, this latter metrics provides an estimate both of a task success (i.e. how much of the required information has been collected, what proportion of the dialog tree has been covered) and how much the collected dialog deviates from the dialog tree (how many turns are not about the medical topics covered by the dialog tree).

6 Results

We compare the classification and the generation models using both automatic metrics and human evaluation. We present various ablation settings to analyse the impact of dialog state information on performance. And we display an example dialog between a human and the generative model in Table 7.

6.1 Automatic Metrics

Model	F1			BLEU-4			BERTScore		
	L	S	A	L	S	A	L	S	A
CLASSIF Oracle	0.7943	0.4323	0.7794	0.8343	0.4228	0.7538	0.9668	0.9144	0.9674
CLASSIF	0.6259	0.3780	0.4780	0.6528	0.3897	0.4789	0.9460	0.9124	0.9217
CLASSIF (predict only)	0.6250	0.3697	0.4044	0.6553	0.3673	0.4166	0.9457	0.9088	0.9108
GEN Oracle	0.8269	0.6794	0.8545	0.6176	0.5240	0.6217	0.9595	0.9497	0.9652
GEN	0.6586	0.3942	0.4956	0.4889	0.3360	0.3689	0.9497	0.9269	0.9320
GEN (predict only)	0.6109	0.3765	0.4687	0.4619	0.1336	0.3530	0.9455	0.9228	0.9305
GEN (no d-state)	-	-	-	0.5202	0.3614	0.3987	0.8725	0.8530	0.8020

Table 2: **Results** on **L**ong, **S**mall and **A**ll Datasets

Table 2 shows results for different versions of the generative and classification models, depending on which dialog state information is provided in the source and the target, at test and at training time.

In the Oracle setting (Oracle), dialog state information is provided for all dialog turns in the input, at training and at test time. This gives an upper bound of how the system would perform given perfect dialog state information. We compare this Oracle setting with a standard setting (CLASSIF and GEN) in which only the dialog state associated with the doctor queries are given. At training time, this is the reference dialog state associated with the doctor query. At test time, it is the dialog state of the doctor query predicted by the model.

To analyse the impact of dialog state information on performance, we also execute an ablation study considering models where (i) no dialog state information is given in the input, but the model is trained to predict the output dialog state (predict only) and (ii) a model where dialog state information is not used at all (no dialog states).

Generation outperforms classification. The F1-score is consistently better for the generation models across all datasets, which suggests that learning to generate the system response also helps to predict the correct system dialog state. As regards similarity with the reference, the generation models also consistently show better BERT score but lower BLEU-4 scores. This is coherent with the specificities of each model. Because the generative model generates the system response rather than select it from the training data (as is the case for the classification model), the similarity in terms of word overlap (as measured by BLEU) with the reference is lower. Nonetheless, the high BERT score indicates a strong semantic similarity between the reference and the generated output.

Predicting the output dialog states helps. For both classification and generation model, dialog state information helps improve performance. As expected the improvement is strongest for the Oracle setting. The ablation study further demonstrates that predicting and using predicted dialog state information (CLASSIF, GEN) yields better results compared to settings where dialog state information is only predicted (CLASSIF/GEN predict only) or not used at all (GEN no d-state).

Shorter interactions are hard to learn. Contrasting the results from Short and Long in Table 2, we see that scores for the SHORT dataset are lower across the board – it is harder to handle short interactions. This is because, in that setting, the model needs to handle patient turns which convey multiple information – often from different domains – and, based on this, must decide on the correct response i.e. move to the correct dialog state. For instance, in Example (3), the model must (i) detect that the patient turn conveys information about both sleep and pain domain and (ii) decide to skip the dialog state corresponding to D2 in example (3).

Domain analysis. Table 3 shows the results per domain for the generation and the classification models trained on LONG[6]. Unsurprisingly, results are better for domains (Leisure and Anxiety) with a small number of classes (fewer transitions to learn) and when the training data is larger (Anxiety vs. Leisure and Sleep vs. Mood). This suggests two directions for further research: other paraphrasing techniques could be used to create more training data for those domains where the training data is small and the dialog tree drawn by the expert could be refined to yield more balanced domain subtrees.

Domain	**# D-States**	**% Tg Data**	**CLASSIF**			**GEN**		
			F1	**BLEU-4**	**BERTScore**	**F1**	**BLEU-4**	**BERTScore**
Mood	18	10.44	0.359	0.5505	0.903	0.638	0.5364	0.950
Sleep	33	63.40	0.467	0.6375	0.921	0.601	0.5061	0.947
Leisure	5	4.67	0.554	0.6752	0.932	0.531	0.4692	0.930
Anxiety	7	21.49	0.889	0.9999	0.980	1.000	0.6924	0.973

Table 3: **Results per Domain** (GEN and CLASSIF models trained on LONG)

6.2 Error Analysis

We use the expert dialog tree to analyze how far off the model's predictions are from the correct predictions and compute the proportion of cases where the predicted dialog state is the expected one (Correct), the child of this state in the dialog tree (Child Node) or its parent (Parent Node). We also compute the proportion of cases where the predicted and the expected dialog state have the same grandparent (Same Gd Parent) and for all remaining cases whether they occur as different leaves of the tree (Diff. Leaves) and are or are not in the same domain (In Domain, Out of Domain). Table 4 shows the results.

Error Type	**CLASSIF**	**GEN**
Correct	62.59	65.86
Child Node	4.4	3.28
Parent Node	9.53	10.87
Same Gd Parent	1.69	1.31
Diff. Leaves	15.96	11.51
In Domain	4.28	6.65
Out of Domain	1.53	0.26

Table 4: **Error Analysis on Predicted Dialog States** (GEN and CLASSIF models trained on LONG)

Most predictions are correct or almost correct. We find that together the cases where the prediction is almost correct (Child or Parent Node) covers 13.93% and 13.57% of the cases for the generative and the classification model respectively. This means that the prediction of the dialog state is correct or almost correct 76.52% and 80.01% of the time for the classification and the generative model, respectively.

Most errors are an artifact of the dialog tree. Most predictions which are very far off the expected dialog state are transitions associated with the end of the dialog (Diff. Leaves). This is because although turns concluding a dialog are similar for all domains and all dialog paths, they are associated in the dialog tree with different dialog state identifiers. This could be fixed by assigning each leaf node the same identifier and restarting the chatbot using a turn from another domain when reaching such a node. More generally, this shows that alternative design choices for representing the expert knowledge might impact performance.

Interestingly, the use of dialog states derived from the expert dialog tree increases interpretability and allows for a detailed analysis of the errors made by the models suggesting possible directions for improvement such as, for instance, using the same dialog state identifier for the end of dialog transitions in all domains and all dialog paths (to reduce the proportion of Diff. Leaves error) and focusing on identifying these factors which would help better differentiate between turns associated with closely related dialog states Child or Parent Node.

[6]For the other datasets (SHORT and ALL), we observe the same trends.

6.3 Human Evaluation and Qualitative Analysis

Tables 6 and 5 show the results of the human evaluation.

Response quality. We find that the generative model (GEN, fluency: 4.08) succeeds in generating well-formed responses. Responses that are rated low are often incomplete (e.g., 'in the long run remaining with such unpleasant thoughts doesn't really seem to me to be 'ten' instead of 'tenable'). This is likely due to the model learning an average sentence length which is below that of longer turns and could be remedied by improved tuning. Both models provide reasonably coherent answers (CLASSIF:3.14, GEN:3.32) and while the generative slighty edges out the classification model, the difference (we used a t-test) is not statistically significant at $p < 0.05$.

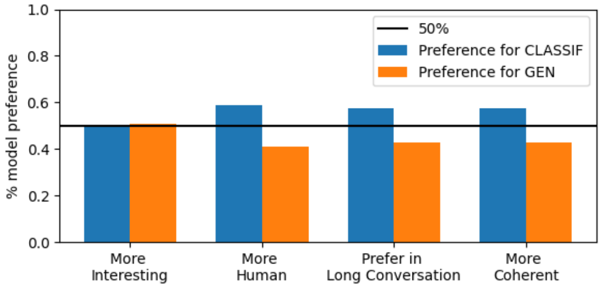


Table 5: **Human Comparison of the two Models**

Criteria	CLASSIF	GEN
Responses		
Coherence	3.14 ± 1.43	3.32 ± 1.40
Fluency	-	4.08 ± 1.28
Whole Dialogs		
Satisfaction	3.12 ± 1.19	3.38 ± 1.03
User turn length	4.92 ± 3.93	5.04 ± 4.24
dialog length	24.42 ± 3.79	28.98 ± 5.63
Sequence rate	0.35 ± 0.13	0.26 ± 0.12

Table 6: **Human Evaluation Scores** (differences are not significant)

Dialog quality. Dialogs are quite long, which indicates that the bot succeeds in driving a non-trivial conversation with the user.

We also observe that the user turns are much shorter than in our training dataset because annotators often respond to questions by a simple yes or no statement rather than a full sentence. This raises the question of how to encourage the user to be more collaborative and provide more informative responses. We leave this here as an open question for further research.

The sequence rate is around a third for each model. Recall that this metrics is the ratio of bot turns that correspond to sub-sequences in the dialog tree (with length more than two). A high score indicates that the model is consistent and capable of engaging the user in a conversation according to the tree. A low score indicates that the model diverges from the dialog tree without creating the expected series of questions, but it also indicates that contrary to a finite state dialog approach where the model is constrained to follow the transitions defined by the finite state automaton, our models can learn new dialog transitions. The observed sequence rates (0.35 and 0.26) suggest both that the models have correctly learned transitions sequences that were defined as natural sounding by the expert and that they can deviate from those, learning new ways to conduct the dialog. We leave a detailed exploration of how these deviations could be used to create alternative dialog paths and thereby enrich the model for further research.

The Acute-Eval results are more nuanced. While the satisfaction (Table 6) and the interest scores (Table 5) are higher for the generative model, the classification model is found more human sounding, more coherent and is preferred for a long conversation. This is in line with previous results Zhang et al. [2018], Dinan et al. [2018] where retrieval models (approximated here by our hybrid classification/retrieval model) were found to score very well in dialog level evaluations because they return human-written utterances from the training set and thus do not suffer from decoding mistakes present in generative models.

7 Conclusion

Using paraphrase identification techniques and a dialog tree to model expert knowledge about doctor-patient interactions, we proposed a novel method to create training data for dialog models, and we used data created using this method to learn health chatbots that cover the main topics standardly used in the questionnaires of clinical studies. We compared two models, a generative and a hybrid classification/retrieval model, and we showed that the expert knowledge captured by the dialog tree both helps guide learning and facilitate error analysis.

Results analysis highlights three main directions for future research. First, additional paraphrase techniques could be explored to create a more balanced dataset. As shown in Table 3, the quantity of training data available for each domain

I am a bot and I will ask you some questions about your health. What is the most difficult for you about your sleep ?
I wake up multiple times during the night
is there a specific event or a change in your activity which could explain this pain?
I believe it's because of my incoming exams
are they frequent or infrequent?
Only twice a year
even if it is not always easy to treat chronic sleep issues why don't you speak with
I don't think it's that urgent for now
okay! do you have any other things to tell me?
I really feel tired all day long
would you agree to discuss anxiety and stress?
Sure!
if you already feel those troubles for the last days why don't you speak with your physician
I don't think I really need it for now
Do you feel like events have been able to affect your mood lately?
My exams
do you ever find yourself thinking that it would be best not to wake up the next mornin
No, it's just my exams!!
do you sometimes think about ending your life?
I just told you I'm ok
in the long run remaining with such unpleasant thoughts doesn't really seem to me to be ten
I think I should be good after my exams are finished
Were you rather anxious?
Yes, I still am, I don't want to fail
What activities you enjoy participating in?
I like to play baseball with my university's team
Okay! Thank you for sharing. So far I have nothing new to tell you.

Table 7: **Conversation between Human and Generative Model**.

varies greatly. We are currently exploring whether paraphrase generation (rather than paraphrase extraction) could help address this issue. Second, longer, richer dialogs could be obtained by extending the expert dialog tree. Here, the American Medical Association Family Medical Guide Kunz [1982] may be used to obtain a new dataset with longer and more precise interaction between doctor and patient, giving more advice and information about patients' state. Third, even in a clinical study context, human dialogs will often mix open-ended chit-chat with targeted health domain interactions. It would be interesting to extend our approach with strategies that engage the user to talk more about his/her problems, e.g., by using ensemble of bots Papaioannou et al. [2017].

Acknowledgements

We thank the anonymous reviewers for their feedback. We gratefully acknowledge the support of the ALIAE company, the French National Center for Scientific Research, and the ANALGESIA Institute Foundation. We also would like to thank testers for their help with human evaluation and for their valuable comments.

References

Paul Green and Lisa Wei-Haas. The rapid development of user interfaces: Experience with the wizard of oz method. *Proceedings of the Human Factors Society Annual Meeting*, 29(5):470–474, 1985. doi:10.1177/154193128502900515. URL `https://doi.org/10.1177/154193128502900515`.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi:10.18653/v1/W17-5526. URL `https://www.aclweb.org/anthology/W17-5526`.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, April 2017. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/E17-1042`.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:10.18653/v1/P18-2033. URL `https://www.aclweb.org/anthology/P18-2033`.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. Enhancing dialogue symptom diagnosis with global attention and symptom graph. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5033–5042, Hong Kong, China, November 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-1508. URL `https://www.aclweb.org/anthology/D19-1508`.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of AAAI*, volume 33, 2019.

Sourabh Majumdar, Serra Sinem Tekiroglu, and Marco Guerini. Generating challenge datasets for task-oriented conversational agents through self-play. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 693–702, Varna, Bulgaria, September 2019. INCOMA Ltd. doi:10.26615/978-954-452-056-4_081. URL `https://www.aclweb.org/anthology/R19-1081`.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana, June 2018. Association for Computational Linguistics. doi:10.18653/v1/N18-3006. URL `https://www.aclweb.org/anthology/N18-3006`.

Alec Radford and Tim Salimans. Improving Language Understanding by Generative Pre-Training. *OpenAI*, pages 1–12, 2018. URL `https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language{_}understanding{_}paper.pdf`.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*, 2019.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, volume 3, pages 1715–1725. Association for Computational Linguistics (ACL), 2016. ISBN 9781510827585. doi:10.18653/v1/p16-1162.

Margaret Li, Jason Weston, and Stephen Roller. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*, 2019.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:2204–2213, 2018. doi:10.18653/v1/p18-1205.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *ICLR*, 2018.

Jeffrey R. M. Kunz. *The American Medical Association Family Medical Guide*. New York: Random House, 1982.

Ioannis Papaioannou, Amanda Cercas Curry, Jose Part, Igor Shalyminov, Xu Xinnuo, Yanchao Yu, Ondrej Dusek, Verena Rieser, and Oliver Lemon. Alana: Social dialogue using an ensemble model and a ranker trained on user feedback. In *2017 Alexa Prize Proceedings*, 2017.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

GATHERING INFORMATION AND ENGAGING THE USER COMBOT : A TASK-BASED, SERENDIPITOUS DIALOG MODEL FOR PATIENT-DOCTOR INTERACTIONS

Anna Liednikova
ALIAE
Université de Lorraine
5 rue Jacques Villemaux
54000 Nancy
`anna.liednikova@aliae.io`

Philippe Jolivet
ALIAE
5 rue Jacques Villemaux
54000 Nancy
`philippe.jolivet@aliae.io`

Alexandre Durand-Salmon
ALIAE
5 rue Jacques Villemaux
54000 Nancy
`alexandre.durand-salmon@aliae.io`

Claire Gardent
615, rue du Jardin Botanique
54600 Villers-lès-Nancy
CNRS
`claire.gardent@loria.fr`

ABSTRACT

A key bottleneck for developing dialog models is the lack of adequate training data. We focus on dialog models in the context of clinical studies where the goal is to help gather, in addition to the closed set of information collected based on a questionnaire, serendipitous information that is medically relevant. To promote user engagement and address this dual goal (collecting both a predefined set of data points and more informal information about the state of the patients), we introduce an ensemble model made of three bots: a task-based, a follow-up and a social bot. We introduce a generic method for developing follow-up bots. We compare different ensemble configurations, and we show that the combination of the three bots (i) provides a better basis for collecting information than just the information seeking bot and (ii) collects information in a more efficient manner than an ensemble model combining the information seeking and the social bot.

1 Introduction

Current work on Human-Machine interaction focuses on three main types of dialogs: task-based, open domain and question answering conversational dialogs. The goal of task-based models is to gather the information needed for a given task, e.g., gathering the price, location and type of restaurant needed to recommend this restaurant. Usually trained on social media data Roller et al. [2020] Adiwardana et al., open domain conversational models aim to mimic open domain conversation between two humans. Finally, question answering conversational models seek to model dialogs where a series of interconnected questions is asked about a text passage.

in the context of clinical studies i.e., dialog models which are used to collect the information needed by the medical body to assess the impact of the clinical trial on a cohort of patients (e.g., information about their mood, their activity, their sleeping patterns). In the context of these clinical studies, the goal of the dialog model is two-fold. A first goal is to collect a set of pre-defined data points, i.e., answers to a set of pre-defined questions specified in a questionnaire. A second goal is to gather relevant serendipitous information, i.e., health related information that is not addressed by the questionnaire, but that is provided by the user during the interaction and which may be relevant to understand the impact of the therapy investigated by the clinical study. This requires keeping the user engaged and prompting him/her with relevant follow-up questions.

Contributions from Claire and colleagues to the field of Computational Linguistics, pages 175–184.

To model these three goals (collecting a predefined set of data points, keeping the user engaged and gathering more informal information about the state of the patient), we introduce an ensemble model which combines three bots: a task-based bot (MEDBOT) whose goal is to collect information about the mood, the daily life, the sleeping pattern, the anxiety level and the leisure activities of the patients; a follow-up bot (FOLLOWUPBOT) designed to extend the task-based exchanges with health-related, follow-up questions based on the user input; and an empathy bot (EMPATHYBOT) whose task is to reinforce the patient engagement by providing empathetic and socially driven feedback.

Our work makes the following contributions.

- We introduce a model where interactions are driven by three main goals: maintaining user engagement, gathering a predefined set of information units and encouraging domain related user input.
- We provide a generic method to create training data for a bot that can follow-up on the user response while remaining in a given domain (in this case, the health domain).
- We show that such a follow-up bot is crucial to support both information gathering and user engagement and we provide a detailed analysis of how the three bots interact.

2 Related Work

Several approaches have explored the use of ensemble models for dialog. While Song et al. [2016] proposed an ensemble model for human-machine dialog which combines a generative and a retrieval model, further ensemble models for dialog have focused on combining agents/bots designed to model different conversation strategies. Yu et al. [2016] focus on open domain conversation and combines three agents, two to improve dialog coherence (ensuring that pronouns can be resolved and maximising semantic similarity with the current context) and one to handle topic switch (moving to a new topic when the retrieval confidence score is low). The ALANA ensemble model Papaioannou et al. [2017a,b], developed for the Amazon Alexa Challenge i.e., for open domain chitchat, combines domain specific bots used to provide information from different sources with social bots to smooth the interactions (by asking for clarification, expressing personal views or handling profanities). Similarly, Yu et al. [2017] introduces a dialog model which interleaves a social and a task-based bot. Conversely, Gunson et al. [2020] showed that success of interleaving depends on the context and that in a public setting, users either prefer purely task-based systems or fail to see a difference between task-based and a richer ensemble model combining task-based and social bots.

Our work differs from these previous approaches in that we combine a standard, task-based model with both a social bot and a domain specific, follow-up bot. This allows both for more natural dialogs (by following up on the user input rather than systematically asking about an item in the predefined set of topics) and for additional relevant, health related information to be gathered.

3 ComBot, an ensemble Model for Repeated Task-Based Interactions

We introduce the three bots making up our ensemble model and the ensemble model combining them.

3.1 Medical Bot

MEDBOT is a retrieval model which uses the pre-trained ConveRT dialog response selection model Henderson et al. [2019] to retrieve a query from the MedTree Corpus Liednikova et al. [2020]. It is designed to collect information from the user based on a predefined set of questions contained in a questionaire.

The MedTree Dataset. The MedTree corpus Liednikova et al. [2020] was developed to train a task-based, information seeking, health bot on five domains: sleep, mood, anxiety, daily tasks and leisure activities. It was derived from a dialog tree provided by a domain expert (i.e., a physician) and designed to formalise typical patient-doctor interactions occurring in the context of a clinical study. In that tree, each branch captures a sequence of (Doctor Question, Patient Answer) pairs and each domain is modeled by a separate tree with the root introducing the conversation (initial question) and the leaves providing a closing statement. The MedTree corpus is then derived from this tree by extracting from each branch of the tree, all context-question pairs, where the context consists of a sequence of patient-doctor-patient turns present on that branch and the question is the following doctor question. A fragment of the decision tree created for the sleep domain and an example dialog are shown in Figure 1.

There are two versions of the MedTree corpus: one consisting of only the context/question pairs derived from the dialog tree (INIT) and the other including variants of these pairs based on paraphrases extracted from forum data (ALL). In Liednikova et al. [2020], the ALL corpus is used to train a generative and a classification model. In our work, we use (a slightly modified version[1] of) the INIT corpus instead, as its small size facilitates retrieval (the number of candidates is small) and preliminary experimentations showed better results when using the INIT corpus.

Model. ConverT is a Transformer-based Encoder-Decoder which is trained on Reddit (727M input-response pairs) to identify the dialog context most similar to the current context and to retrieve the dialog turn following this context. In order to retrieve from the MedTree corpus, the question that best fits the current dialog context, the MEDBOT model compares the last three turns of the current dialog with contexts from the MedTree Corpus. The model identifies the MedTree corpus context with the highest similarity score[2] and outputs the question following that context. If the selected question has already been asked in the dialog generated so far and provided it is not a question such as "What other things would you like to share with me ?", we retrieve the next best question that is not a repetition. No fine-tuning is done due to the small amount of data.

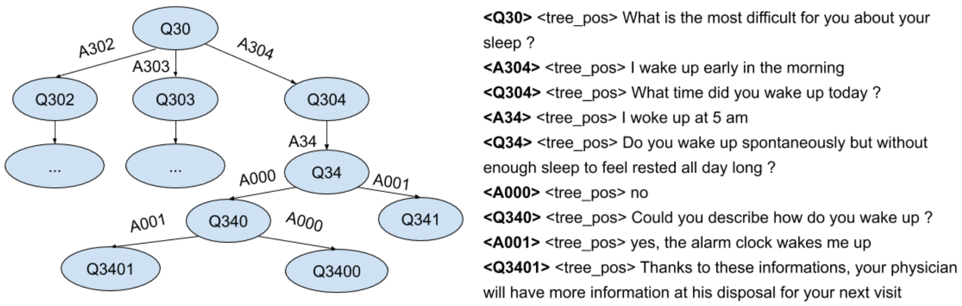


Figure 1: **Fragment of decision tree for the sleep domain and a corresponding dialog**

3.2 Follow-Up Bot

One main motivation behind the use of a health-bot in clinical studies is to complement the information traditionally gathered through a fixed questionnaire filled in each week by the patients with serendipitous information, i.e., information that is not actively queried by the questionnaire, but that is useful to analyse the cohort results.

The MEDBOT model introduced in the previous section is constrained to address only those topics which are present in the dialog tree, in effect, modeling a closed questionnaire. To allow for the collection of serendipitous health information, we develop the FOLLOWUPBOT whose function is to generate health-related questions which are not predicted by the dialog tree but which naturally follow from the user input. The main difference of FOLLOWUPBOT from MEDBOT is the way it retrieves questions that are not in the sequence, but the ones that occur in the same context, even if the question itself doesn't share the lexicons with the previous turns. Rather than artificially restricting the dialog to the limited set of topics pre-defined by the dialog tree, the combined model (MEDBOT + FOLLOWUPBOT) allows for transitions based either on the dialog tree or on health-related, follow-up questions. In that sense, FOLLOWUPBOT allows not only for the collection of health-related serendipitous information, but also for smoother dialog transitions.

Like MEDBOT, FOLLOWUPBOT used the pre-trained ConveRT model to retrieve context appropriate queries from a dialog dataset. In this case however, the queries are retrieved from the HealthBoard dataset, a new dataset we created to support follow-up questions in the health domain.

The Healthboard Dataset. This dataset consists of (s, q) pairs where s is a (health related) statement and q is a follow-up question for that statement. We extract this dataset from the Healthboard forum [3] as follows. We first select 16 forum categories (listed in Table 1) that are relevant to our five domains. In the forum, each category includes multiple conversational threads, each thread consists of multiple posts and each post is a text of several paragraphs that can be split into sentences. In total, we collect 175,789 posts from 31,042 threads with 5.68 posts in average per thread. We then segment each post into sentences using the default NLTK sentence segmenter. We label each sentence with a dialogue act classifier in order to distinguish statements ("sd" label) from questions ("qo" label). For this labelling, we fine-tune the Distilbert Transformer-based classification model [4] on the Switchboard Corpus Stolcke et al.

[1]The modifications consists in shortening the questions, changing all leaves to statements and adding meta-statements about the dialog to account for cases where the user indicates misunderstanding or agreement

[2]Both contexts are encoded using ConveRT as average of embeddings of the last turn and concatenation of preceding ones. The inner product is used to compute similarity.

[3]https://www.healthboards.com/

[4]https://huggingface.co/distilbert-base-uncased

Category	Threads	Posts	Avg
anxiety	6852	38523	5.63
anxiety tips	42	71	1.69
chronic fatigue	670	3856	5.77
chronic pain	646	4893	7.59
depression	5327	32998	6.21
depression tips	27	51	1.89
exercise fitness	1583	8142	5.16
general health	7279	29858	4.11
healthy lifestyle	104	621	5.97
pain management	4985	38738	7.79
panic disorders	1314	8376	6.39
share your anxiety story	42	42	1
share your depression story	55	71	1.29
share your pain story	28	42	1.50
sleep disorders	1671	7656	4.59
stress	415	1973	4.76

Table 1: Forum Categories used for the Creation of the HealthBoard Dataset

[2000] using 6 classes "qo" (Open-Question), "sd" (Statement-non-opinion), "ft" (Thanking), "aa" (Agree/Accept), "%" (Uninterpretable) and "ba" (Appreciation). For each question q (i.e., sentence labelled "qo") in each thread T, we gather all statements (i.e., all sentences labeled as "sd") which precede q in T into a pool of candidate statements[5]. As dialogue turns in bots should remain short, we filter sentences that have more than 100 tokens. For each candidate statement, we calculate its similarity with the question using the dot product of their ConveRT embeddings. We filter out all candidate statements whose score with the question is less than 0.6. If after filtering, the resulting pool contains at least one candidate, we select the top-ranked statement and add the statement-question pair to the dataset. The resulting dataset contains 3,181 (statement, question) pairs.

Model. Similar to the MEDBOT model, the FOLLOWUPBOT model used the pre-trained ConveRT model to compare the current dialog context (the preceding three turns) with the statements contained in the HealthBoard dataset using the inner product.

The top-20 candidates are then retrieved and filtered using Maximal Marginal Relevance (MMR) Carbonell and Goldstein [1998] with $\lambda = 0.5$ to control for repetitions[6]. Next, we compute the similarity between the remaining selected questions and the questions included in the current dialog context (all preceding dialog turns) and we exclude candidates with similarity score 0.8 or higher. After filtering, the top ranking candidate is selected and the associated follow-up question is output.

3.3 Empathy Bot

As the name suggests, the role of the EMPATHYBOT is to engage the user by showing empathy. For this bot, we use Roller et al. [2020] generative model which was pre-trained on a variant of Reddit discussion Baumgartner et al. [2020] and fine-tuned on the ConvAI2 Zhang et al. [2018], Wizard of Wikipedia Dinan et al. [2019], Empathetic Dialogues Rashkin et al. [2019], and Blended Skill Talk datasets (BST) Smith et al. [2020] to optimize engagigness and humanness in open-domain conversation.

3.4 Ensemble Model (ComBot)

Each bot provides a single candidate. To rank them, we encode the whole current dialog context and each candidate response using the ConveRT encoder, we calculate similarity (dot product) for each candidate/context pair, and we select the candidate with the highest similarity score. In case all candidates scores are less than 0.1, we consider that there is no good response and we end the conversation.

[5]We do not restrict the set of candidates at that stage i.e., we consider all posts that precede the question within the question thread and all statements in these posts no matter how far away the statement is from the question. In practice, the set of such statements has limited size and distance does not seem to matter too much, although an investigation of that factor would be interesting. We leave this question open for further research, as it is not central to our paper.

[6]MMR is a measure for quantifying the extent to which a new item is both dissimilar to those already selected and similar to the target (here a selected question). A λ value of 0.5 favors similarity and diversity equally, both matter equally.

4 Experiments

4.1 Data

Table 2 shows some statistics for the corpora used for pretraining (ConveRT, Blender) and for retrieval (INIT, HealthBoard). For MEDBOT and FOLLOWUPBOT, we use the ConveRT model from PolyAI [7]. For EMPATHYBOT, we use the Blender model with 90M parameters from the ParlAI library [8] .

One benefit of the ensemble approach is that several models can be combined, each modelling different types of dialog requirements. We compare different configurations of our three bots: COMBOT (which combines the three bots), MEDBOT (using only the task-based bot), MED+EMPATHYBOT an ensemble model which combines the task-based (MEDBOT) and the social bot (EMPATHYBOT) and MEDBOT+ FOLLOWUPBOT, a bot combining the task-based and the follow-up question bot.

We first use automatic metrics and global satisfaction scores to compare the four models. We restrict the Acute-Eval, human-based model comparison to the two best performing systems namely, COMBOT and MEDBOT.

	Reddit	ConvAI2	WoW	EmpaDial	BSD	INIT	HealthBoard
Nb of context-question pairs		211803	83011	76673	27018	168	3181
Nb of distinct turns	1.50B	267945	165213	88757	53335	154	73140
Nb of tokens	568B	3791971	2720426	2625338	912857	3688	202389
Nb of tokens per turn (Avg, Max, Min)		8.95	16.39	17.12	16.89	6.92	11.5
Vocabulary size		20707	95590	59438	52561	306	7321

Table 2: Corpus statistics (Reddit: pre-training corpus for ConveRT and the Empathy bot. ConvAI2, WoW, EmpaDial and BSD: Datasets used to fine-tune the Empathy Bot. INIT: used for the MedBot retrieval step. HealthBoard: for FollowUp Bot Fine-Tuning and Retrieval .)

4.2 Evaluation

As there does not exist a dataset of well-formed health-related dialogs whose aim is both to answer a clinical study questionaire and to allow for serendipitous interactions, we have no test set on which to compare the output of our dialog models. Moreover, as has been repeatedly argued, reference-based, automatic metrics such as BLEU or METEOR, fail to do justice to the fact that a dialog context usually has many possible continuations. We therefore use reference-free automatic metrics and human assessment for evaluation.

Human evaluation. We use the MTurk platform to collect human-bot dialogs for our four models (COMBOT, MEDBOT and MED+EMPATHYBOT) and ask the crowdworkers to provide a satisfaction rate at the end of their interaction with the bot. We then run a second MTurk crowdsourcing task to grade and compare dialogs pairs produced by different models.

To collect dialogs, we ask participants to interact with the bot for as long as they want. The conversation starts randomly with one of the initial questions of MEDBOT. The interaction stops either when all candidates scores are less than 0.1 (cf. Section 3.4) or when the user ends the conversation. For each model, we collect 50 dialogs. Each annotator interacts at most once with a bot.

At the end of each human-bot conversation, the annotator is asked to rate satisfaction on a 1-5 Likert scale (a higher score indicates more satisfaction).

Assigning a satisfaction score to a single dialog is a highly subjective task, however, with scores suffering from different bias and variance per annotators Kulikov et al. [2019]. As argued by Li et al. [2019], comparing two dialogs, each produced by different models, and deciding on which dialog is best with respect to a predefined set of questions, helps support a more objective evaluation. We therefore use the Acute-Eval human evaluation framework to compare the dialogs collected using different bots. Since the automatic evaluation (cf. Section 5.1) shows that COMBOT and MEDBOT are the best systems, we compare only these two systems asking annotators to read pairs of dialogs created by these two bots and to then answer the pre-defined set of questions recommended by Li et al. [2019]'s evaluation protocol namely:

- Who would you prefer to talk to for a long conversation?

[7]https://github.com/connorbrinton/polyai-models/releases/tag/v1.0

[8]https://parl.ai/projects/recipes/

- If you had to say one of the speakers is interesting and one is boring, who would you say is more interesting?
- Which speaker sounds more human?
- Which speaker has more coherent responses in the conversation?

We report the percentage of time one model was chosen over the other.

For this comparison, we consider 50 dialog pairs (one dialog produced by COMBOT, the other by MEDBOT) and for each Acute-Eval question, collected 50 judgments, one per dialog pair. We had ten annotators, each annotating at most 5 dialog pairs. To maximise similarity between the dialogs being compared, we create the dialog pairs by computing Euclidean distance between context embeddings of MEDBOT and COMBOT dialogue sets. Then we composed a pair of two closest items and excluded them from the choice in the next iteration.

Automatic Metrics. After collecting dialogues, we perform their automatic evaluation. All scores are computed on the 50 bot-human dialogs collected for a given model. Table 3 shows the result scores averaged over 50 dialogs.

To measure *coherence*, we exploit the unsupervised model CoSim introduced by Mesgar et al. [2019], Xu et al. [2018], Zhang et al. [2017]. This model measures the coherence of a dialog as the average of the cosine similarities between ConveRT embedding vectors of its adjacent turns.

To assess *task success*, we count the number of unique medical entities (Slots) mentioned. We do this using the clinical NER-model from the Stanza library Zhang et al. [2020] [9], a model trained on the 2010 i2b2/VA dataset Uzuner et al. [2011] to extract named entities denoting a medical problem, test or treatment. We report the average number of medical entities, both per dialog and in the user turns (to assess how much medical information comes from the user).

Model	Satisf.	CoSim	Slots	ConvLen	InfoGain	UserQ
MEDBOT	3.94	0.26	6.24 (1.68)	28.46	108.82 (3.82)	0.08 (4)
MEDBOT+ FOLLOWUPBOT	3.18	0.34	11.65 (3.22)	36.06	153.23 (4.25)	0.47 (23)
MEDBOT+ EMPATHYBOT	3.77	0.34	3.87 (1.46)	30.29	140.19 (4.63)	0.68 (33)
COMBOT	3.72	0.36	7.12 (2.82)	21.96	124.82 (5.68)	0.48 (24)

Table 3: Satisfaction Scores (Satisf.) and Results of the Automatic Evaluation. CoSim: Average Cosine Similarity between adjacent turns. Slots: Average Number of Medical Entities per dialogue (in brackets: average number in the user turns). ConvLen: Average Number of turns per dialog. InfoGain: Average number of unique tokens per dialog (in brackets: normalised by dialog length). UserQ: number of questions asked by Human (in bracket: total number for 50 dialogs). All metrics are averaged over the 50 Human-Bot dialogs collected for each model.

Following Yu et al. [2017], we also calculate *Information gain (InfoGain)*, the average number of unique tokens per dialog and *Conversation Length (ConvLen)*, the average number of turns in the overall dialog.

Finally, we compute the number of questions asked by the user (*UserQ)* as an indication of the user trust and engagement. We compute both the total number of questions present in the 50 dialog collected for a given model and the average number of question per dialog.

5 Results and Discussion

We compare four models using automatic metric and absolute satisfaction scores. Based on this first evaluation, we compare two of these models using the Acute-Eval human evaluation framework. We display an example dialog and discuss the respective use of each bot in the COMBOT model.

5.1 Automatic Evaluation and Absolute Satisfaction Scores

Table 3 shows the absolute satisfaction scores (i.e., scores provided on the basis of a single dialog rather than by comparing dialogs produced by different models) and the results of the automatic evaluation for the four models mentioned above.

ComBot provides a better basis for collecting information than MedBot. The automatic scores show that COMBOT consistently outperforms MEDBOT on informativity (Slots, InfoGrain) while allowing for shorter dialogs (ConvLen).

[9] http://stanza.run/bio

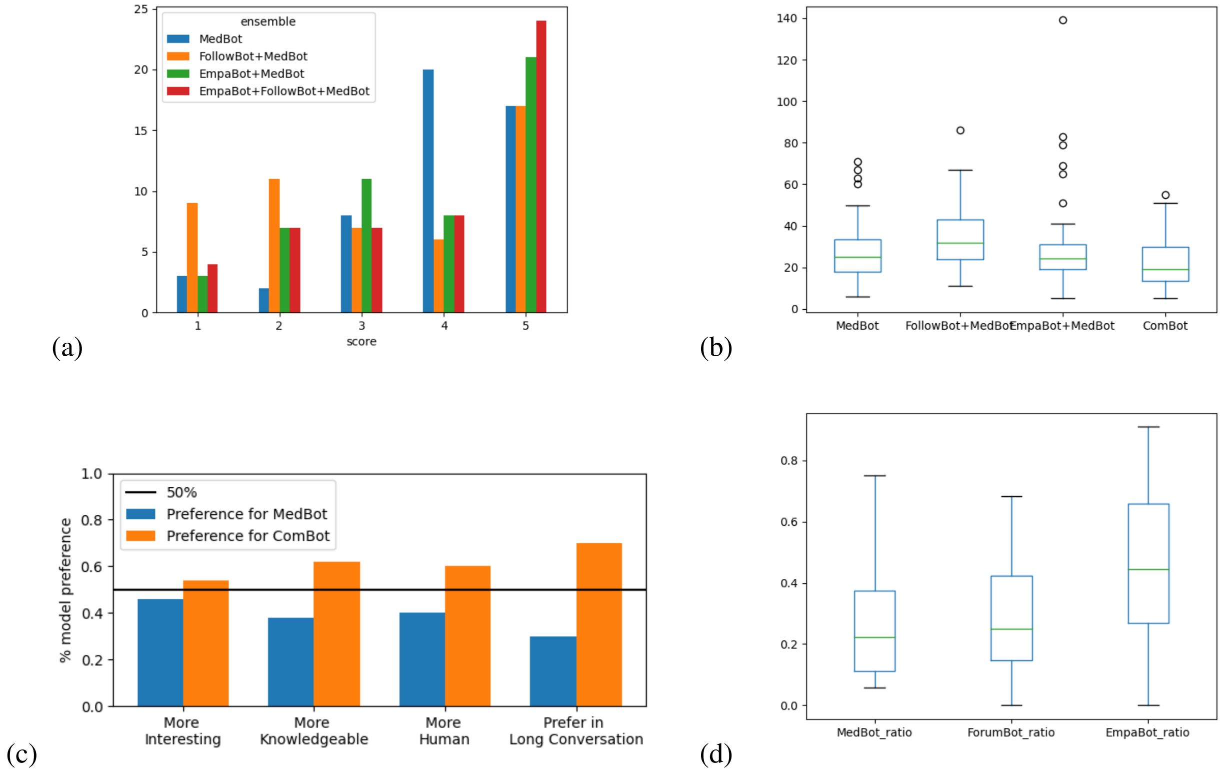


Figure 2: (a) Distribution of the Satisfaction Scores for each configuration, (b) Conversation length distribution for MedBot and ComBot, (c) Acute-Eval results for both systems, (d) Majority bot ratio in COMBOT

In other words, COMBOT allows for a larger range of informational units (words and medical named entities) to be discussed in fewer turns.

ComBot collects information in a more user-friendly, more efficient manner than Med+EmpathyBot. While the InfoGain scores are higher for MED+EMPATHYBOT and MEDBOT+FOLLOWUPBOT than for COMBOT (InfoGain: 140.19 and 153.23 vs. 124.82), this is achieved at the cost of much longer dialogs (ConvLen: 30.29 and 36.06 vs. 21.96; cf. also Figure 2b) In fact, when normalising InfoGain by the number of dialog turns (ConvLen), we see that in average, a turn in COMBOT dialogs contains a much higher number of unique tokens (i.e., is more informative) than for MEDBOT (3.82), MEDBOT+EMPATHYBOT (4.63) or MEDBOT+FOLLOWUPBOT (4.25).

ComBot allows for more coherent dialogs. In terms of quality, the differences in satisfaction scores between the three models is not statistically significant ($p < 0.05$, T-test). For dialog coherence (Measured by CoSim) however, COMBOT scores highest (0.36) and the difference with MEDBOT is statistically significant ($p < 0.05$, T-test). This suggests that follow up questions help support smoother transitions between dialog turns.

5.2 Comparative Human Evaluation

The results of the comparative human evaluation are presented in Figure 2.

ComBot is judged more knowledgeable, more interesting, more human and better for long conversations. COMBOT outperforms MEDBOT on all Acute-Eval questions (Figure 2c).

In particular, users find COMBOT more knowledgeable by a large margin. This is in line with the automatic metrics results (higher COMBOT values for Slots and InfoGain) and is likely due to the fact that the COMBOT model supports the use of health-related, follow-up questions which in turn allows for a wider range of medical issues to be discussed than just those present in the MEDBOT corpus.

Users also show a clear preference for COMBOT in long conversations (Figure 2a). While this seems to contradict the fact that both models have similar satisfaction score, we conjecture that the high MEDBOT satisfaction score is an artefact of the MEDBOT model. Since the MEDBOT coverage is restricted, the users have low expectations and correspondingly give high satisfaction scores (they are easily satisfied because their expectations are low). An indication of these low user expectations is given by the number of questions asked : when users feel that the system they interact with is unrestricted, they will feel comfortable asking questions and will start to do so. Conversely, if they feel the model is restricted, they will refrain from asking questions. The results show a much higher number of questions for users interacting with COMBOT (Table 3).

5.3 Component analysis

BOT: : *[opening question]* What would you like to share today?
USER: : i feel soo tired
BOT: : *[FOLLOWUPBOT]* Is something making you hyper at night?
USER: : maybe i am so stressed that i cannot sleep
BOT: : *[MEDBOT]* What wakes you up at night ?
USER: : nothing, i just cannot fall asleep
BOT: : *[FOLLOWUPBOT]* Are your sleep issues anxiety related?
USER: : i think so
BOT: : *[FOLLOWUPBOT]* Also do you snore when you do sleep possibly sleep apnea?
USER: : i do not know
BOT: : *[MEDBOT]* Do you feel anxiety or stress during nights awakenings ?
USER: : yes, a bit
BOT: : *[MEDBOT]* Anxiety and sleep may be connected, it is time to speak with your physician
USER: : okay
BOT: : *[FOLLOWUPBOT]* What about getting a second opinion from another doctor about your situation?.
USER: : why not
BOT: : *[FOLLOWUPBOT]* What does your doctor say about coming off of them and then trying to go back on.
USER: : i do not know
BOT: : *[MEDBOT]* Are you not able to fall asleep regularly or punctually ?
USER: : regularly
BOT: : *[EMPATHYBOT]* i hope things get better for you , and that you are able to enjoy your time with your family .
USER: : thank you
BOT: : *[EMPATHYBOT]* you are welcome , and i wish you the best of luck with the rest of your life . :)

Figure 3: Example Human-ComBot dialog

Figure 3 displays an example Human-Bot dialog using the COMBOT model which illustrates the interactions between the three composing bots: the EMPATHYBOT closes the conversation with social chit-chat, the FOLLOWUPBOT responds to the user turn and MEDBOT asks questions from the dialog tree whenever suitable.

The proportion of turns generated by each bot (cf. Figure 2d) varies from one dialog to another, illustrating the capacity of the ensemble model to adapt to various dialog users and contexts. We find that in 55% of the collected dialogs, a majority of turns (i.e., more than 33% of the turns) is generated by the EMPATHYBOT model; in 29% of the cases by the FOLLOWUPBOT and in 16% of the cases by the MEDBOT[10]

We also observe interesting dependencies and correlations. MEDBOT is triggered twice more often after FOLLOWUP-BOT (30 cases) than after EMPATHYBOT (12 cases) – this indicates that follow-up questions help to bring the user back to the questions contained in the dialog tree.

6 Conclusion

A qualitative analysis of the collected dialogs indicates several directions for further research.

Negation is often not recognised leading to interactions in which the model continues discussing a topic which was declared as irrelevant by the user. Another difficulty is knowing when to end the conversation. Long ones are good to complete the task, but bad for people who are ready to finish the conversation but feel forced to continue. To improve user engagement, a possibility would be to explore whether the information provided by sentiment analysers could be exploited to help maintain a positive interaction. By detecting polarity, it could also help improve negation handling.

Another key issue concerns the emotional impact of the dialog on the user. An interaction with the bot might highlight a health issue the user was not aware of resulting in increased user stress. In such a situation, a good policy would be to

[10] Since a COMBOT dialog has an average of 21 turns and only half of those are generated by the bot, this means that for 55% of the collected dialogs, the dialog contains more than 3 “social” dialog turns (turns generated by EMPATHYBOT). Similarly, 29% of the collected dialogs contain more than 3 follow-up turns (FOLLOWUPBOT) and 16% more than 3 task-based turns (MEDBOT).

provide the user with some notion of solution, some piece of information or advice which can help her face the situation and if possible, incite her to act to improve her health. Indeed, some of the dialogs collected with COMBOT show that users sometimes ask for help.

Here, a knowledge-based agent could be useful either to provide facts that are related to the topic at hand or to highlight the connections between facts that have been mentioned in the dialog.

Acknowledgements

We thank the anonymous reviewers for their feedback. We would like to acknowledge Farnaz Ghassemi for her help in developing the FOLLOWUPBOT. We gratefully acknowledge the support of the ALIAE company, the French National Center for Scientific Research, and the ANALGESIA Institute Foundation.

References

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot, 2020.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu Quoc, and V Le. Towards a Human-like Open-Domain Chatbot. Technical report.

Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. Two are Better than One: An Ensemble of Retrieval- and Generation-Based Dialog Systems. oct 2016. URL `http://arxiv.org/abs/1610.07149`.

Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. Strategy and policy learning for non-task-oriented conversational systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 404–412, Los Angeles, September 2016. Association for Computational Linguistics. doi:10.18653/v1/W16-3649. URL `https://www.aclweb.org/anthology/W16-3649`.

Ioannis Papaioannou, Amanda Cercas Curry, Jose L Part, Igor Shalyminov, Xinnuo Xu, Yanchao Yu, Ondřej Dušek, Verena Rieser, and Oliver Lemon. An ensemble model with ranking for social dialogue. *arXiv preprint arXiv:1712.07558*, 2017a.

Ioannis Papaioannou, Amanda Cercas Curry, Jose Part, Igor Shalyminov, Xu Xinnuo, Yanchao Yu, Ondrej Dusek, Verena Rieser, and Oliver Lemon. Alana: Social dialogue using an ensemble model and a ranker trained on user feedback. In *2017 Alexa Prize Proceedings*, 2017b.

Zhou Yu, Alan W. Black, and Alexander I. Rudnicky. Learning conversational systems that interleave task and non-task content. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4214–4220. AAAI Press, 2017. ISBN 9780999241103.

Nancie Gunson, Weronika Sieińska, Christopher Walsh, Christian Dondrup, and Oliver Lemon. It's good to chat? evaluation and design guidelines for combining open-domain social conversation with task-based dialogue in intelligent buildings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375863. doi:10.1145/3383652.3423889. URL `https://doi.org/10.1145/3383652.3423889`.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. Convert: Efficient and accurate conversational representations from transformers, 2019.

Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. Learning health-bots from training data that was automatically created using paraphrase detection and expert knowledge. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2020.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374, 2000. URL `https://www.aclweb.org/anthology/J00-3003`.

Jaime Carbonell and Jade Goldstein. Use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, pages 335–336, New York, New York, USA, 1998. ACM Press. ISBN 1581130155. doi:10.1145/3130348.3130369. URL `http://portal.acm.org/citation.cfm?doid=290941.291025`.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset, 2020.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:10.18653/v1/P18-1205. URL `https://www.aclweb.org/anthology/P18-1205`.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents, 2019.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1534. URL `https://www.aclweb.org/anthology/P19-1534`.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills, 2020.

Ilia Kulikov, Alexander H. Miller, Kyunghyun Cho, and Jason Weston. Importance of search and evaluation strategies in neural dialogue modeling, 2019.

Margaret Li, Jason Weston, and Stephen Roller. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*, 2019.

Mohsen Mesgar, Sebastian B ¨ Ucker, and Iryna Gurevych. A Neural Model for Dialogue Coherence Assessment. Technical report, 2019. URL `www.ukp.tu-darmstadt.de`.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. Better conversations by modeling, filtering, and optimizing for coherence and diversity. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. doi:10.18653/v1/d18-1432. URL `http://dx.doi.org/10.18653/v1/d18-1432`.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. Reinforcing Coherence for Sequence to Sequence Model in Dialogue Generation. Technical report, 2017. URL `http://ntcirstc.noahlab.com.hk/STC2/stc-cn.htm`.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D. Manning, and Curtis P. Langlotz. Biomedical and Clinical English Model Packages in the Stanza Python NLP Library. jul 2020. URL `http://arxiv.org/abs/2007.14640`.

Ö. Uzuner, B.R. South, S. Shen, and S.L. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

Workshop in recognition of Claire Gardent
Nancy, France, March 21, 2023

EXPLORING THE INFLUENCE OF DIALOG INPUT FORMAT FOR UNSUPERVISED CLINICAL QUESTIONNAIRE FILLING

Farnaz Ghassemi Toudeshki
ALIAE
IDMC
5 rue Jacques Villemaux
54000 Nancy
`farnaz.ghassemi@aliae.io`

Anna Liednikova
ALIAE
Université de Lorraine
5 rue Jacques Villemaux
54000 Nancy
`anna.liednikova@aliae.io`

Philippe Jolivet
ALIAE
5 rue Jacques Villemaux
54000 Nancy
`philippe.jolivet@aliae.io`

Claire Gardent
CNRS
615, rue du Jardin Botanique
54600 Villers-lès-Nancy
`claire.gardent@loria.fr`

ABSTRACT

In the medical field, we have seen the emergence of health-bots that interact with patients to gather data and track their state. One of the downstream application is automatic questionnaire filling, where the content of the dialog is used to automatically fill a pre-defined medical questionnaire. Previous work has shown that answering questions from the dialog context can successfully be cast as a Natural Language Inference (NLI) task and therefore benefit from current pre-trained NLI models. However, NLI models have mostly been trained on text rather than dialogs, which may have an influence on their performance. In this paper, we study the influence of content transformation and content selection on the questionnaire filling task. Our results demonstrate that dialog pre-processing can significantly improve the performance of zero-shot questionnaire filling models which take health-bots dialogs as input.

1 Introduction

Work on Question Answering (QA) and Machine Reading Comprehension (MRC) mostly focuses on wh-questions of arbitrary types (who, what, where etc.) whose answer can be found in text. The answer can be extractive where a short span of the text is identified as the answer [Pearce et al., 2021] or it can be abstractive where a free-form answer is generated from the question and some support document [Bauer et al., 2018].

Here, we focus instead on a QA setting where questions are restricted to polar (yes/no) and Agreement Likert Scale (ALS) questions and where answers are contained in a dialog rather than a paragraph text. As illustrated in Figure 1, this setting is useful for automatic questionnaire filling (AQF) in the medical field. Given a dialog between a patient and a health bot, the goal of automatic questionnaire filling is to answer a set of predefined questions from a medical questionnaire (here the Pain Beliefs and Perceptions Inventory (PBPI) questionnaire [Williams and Thorn, 1989]) based on the dialog content.

In previous work, Toudeshki et al. [2021] compared three ways of deriving answers to questions from dialogs: Natural Language Inference, Question Answering and Text Classification. For polar and ALS questions, they found that Natural Language Inference (NLI) performs best.

One possible limitation of their approach however is that they apply NLI models to dialogs while NLI models are trained on non-dialogic text.

Dialog
bot: What is the most difficult for you about your sleep ? **patient:** I have back pain that prevents me from sleeping. **bot:** I'm sorry to hear that. How long have you had back pain? **patient:** Since I've been working out, I've had constant back pain at night. **bot:** Do you think pain can last for long? **patient:** I think it will stop once I stop playing sports. **bot:** Should we let time fix the pain? **patient:** My doctor thinks that I need to get used to doing sports and that the pain will disappear after a while.
Questionnaire
(1) My pain is a temporary problem in my life. CQ: ☐No **Yes** ☐NA ALS: ☐Totally disagree ☐Rather disagree ☐Agree **Totally agree** ☐NA

Figure 1: An example of a dialog and a question from the PBPI Questionnaire, answered in CQ and ALS format

In this paper, we propose different ways of transforming and selecting dialog content before applying NLI to answer questions, and we analyse the impact of these operations on NLI-based questionnaire filling. Our hypothesis is that transforming the input dialog into a format closer to the text format on which NLI models are trained, should help these models perform better. Our experimental results confirm this hypothesis: it demonstrates that, in a zero-shot setting, transforming and selecting dialog content yields significant improvements over a baseline which takes the full dialog content as input.

2 Related work

We briefly situate our work with respect to three tasks which have similarities with Automatic Questionnaire Filling namely, Machine Reading Comprehension, Question Answering and Aspect-Based Sentiment Analysis (ABSA).

MRC/QA. Given a text and a question, MRC and QA models aim to derive the answer to that question from some input document [Zeng et al., 2020].

Similar to our approach, Ren et al. [2020] focus on filling in medical questionnaires consisting of polar questions about medical terms. However, in their case, the input to the model is a text (patient records) rather than a dialog. Furthermore, QA is modeled as a classification task which restricts the approach to a limited set of possible questions and answers. Finally, the questions are restricted to polar questions about terms whereas we consider polar and ALS questions about full sentences.

Recently, some work has focused on answering questions from dialogs rather than text. A simple approach for modeling a multi-turn dialog is to concatenate all turns [Zhang et al., 2019, Adiwardana et al., 2020]. However, for retrieval-based response selection, Zhang et al. [2018], Yuan et al. [2019] showed that turns-aware aggregation methods can achieve a better understanding of dialogs compared to considering all turns equally

Similarly for MRC on dialogs, turns-aware approach have been proposed which select turns in the conversation that are related to the input question: Zhang et al. [2021] uses embedding-based similarity to select such turns while Li et al. [2020] uses a pre-trained language model fine-tuned on NLI tasks. Their results showed that eliminating irrelevant turns effectively improves results. Our work extends on this work showing that both content selection and content transformation help improve MRC on dialogs.

Aspect-Based Sentiment Analysis. Aspect based sentiment analysis (ABSA) is the process of determining sentiment polarity for a specific aspect in a given context. An aspect term is generally a word or a phrase which describes an aspect of an entity [Jiang et al., 2019]. For instance, [Jang et al., 2021, Sun, 2022] investigate aspect-based sentiment analysis on user tweets related to COVID-19.

While AQF could be viewed as an ABSA task where each item should be labelled with one of three (polar question) or five (ALS question) sentiment value (agree, disagree, etc.), two key differences between ABSA and AQF is that (i) labels apply to sentences rather than aspect terms and (ii) contrary to these terms, the questions used in medical questionnaire can be very similar semantically (e.g., "Is your pain constant?" "Is your pain a temporary problem?") making it harder to extract the correct answer from the input dialog.

Closest to our work, Toudeshki et al. [2021] showed that pre-trained NLI models can be used to fill in questionnaires from dialogs in a zero-shot setting. We depart from their work in that we propose different ways of transforming and selecting dialog content and investigate how this impact zero-shot, dialog-based, automatic questionnaire filling.

3 Automatic Questionnaire Filling (AQF)

Task. Given a dialog D and a questionnaire Q, the Automatic Questionnaire Filling task consists in providing an answer a_i for each question $q_i \in Q$.

We address the task in a zero-shot setting (no training data). For evaluation, we provide a test set consisting of 100 dialogs and their associated questions and answers.

Questionnaire. We consider two types of questions: Closed Questions (CQ) and Agreement Likert Scale (ALS) questions. CQ have three possible answers (yes, no or Not Applicable, i.e. the dialog does not address the question) and ALS has five (totally disagree, rather disagree, agree, totally agree, NA). As illustrated in Figure 1, questions are reformulated as declarative statements with multiple choice answers. With the emergence of health-bots, AQF can

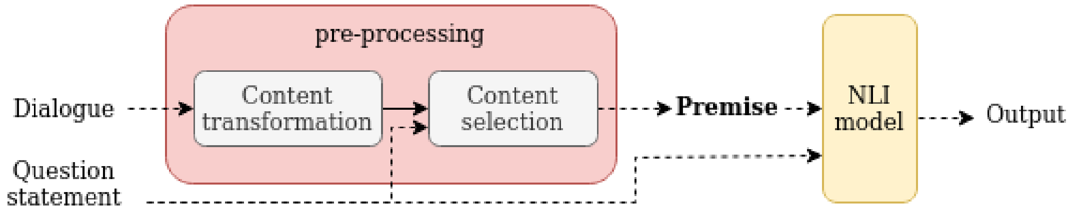


Figure 2: Dialog pre-processing schema

help transform human-bot dialogs into structured data which can be used by physicians to track patients condition. In particular, it can be used to fill in questionnaires such as the Pain Beliefs and Perceptions Inventory (PBPI, [Williams and Thorn, 1989]) questionnaire, which includes 16 questions and is standardly used in the context of clinical studies.

Collecting dialogs that include information for all of these questions is a difficult task, however. To facilitate data collection for the creation of the test set, we therefore decrease the number of questions by selecting five questions out of sixteen. Because the questions in the PBPI are often very similar, and knowing the answer to one of them allows deriving the answer to others, we chose questions that are semantically distinct from one another. The list of all PBPI questions is given in Appendix C and the five selected questions are indicated in bold.

Test Data. To evaluate our approach, we create a test set of 100 dialogs and their associated question/answer pairs.

The creation of the test data involves first, collecting human-bot dialogs and second, extracting answers to the PBPI questions from the collected dialogs.

Collecting Dialogs. We collect the dialogs using the Amazon Mechanical Turk platform and asking Turkers to interact with the ComBot health bot [Liednikova et al., 2021] while behaving as if they had chronic pain issues. To avoid Turkers introducing the PBPI questions verbatim in the dialog, they were given a list of topics to be mentioned rather than the questions themselves (See details in Appendix D). In this way, we ensure that the collected dialogs address the questions to be answered while encouraging their diversified paraphrasing during the conversation. Turkers received bonuses each time they mention a topic. Turkers were also given the ability to modify the bot utterance in order to redirect the conversation more easily: they could reject the current candidate in which case, the turn with the next highest confidences score would be displayed by the bot. More information about Turkers payments is provided in the Ethic section (Sec. A). Details of the instructions given to the Turkers and a screenshot of the annotation interface are given in the Appendix.

Identifying Question Answers. Two annotators with good English proficiency were asked to select the correct answer for each of the five selected questions based on each of the 100 collected dialogs. We computed agreement between the two annotators on all Q/A pairs and all 100 dialogs. The Kappa score is 0.94 for CQ and 0.86 for ALS question type. Thereafter, we used adjudication to decide on the final answer for all cases where the two annotators disagreed. The annotators were the first two authors of this paper.

The final test corpus consists of 100 dialogs, each associated with 10 questions (5 yes/no questions and 5 ALS questions) and their answers. Dialog length varies from 4 to 70 turns and from 47 to 593 tokens, with 17.1 turns and 218.7 tokens on average.

4 Approach

Following Toudeshki et al. [2021], we model question answering as an NLI task where the premise is derived from the dialog, the hypothesis from the question and the answer from the NLI result. Given a question and a dialog, our model, illustrated in Figure 2, answers the question in three steps as follows.

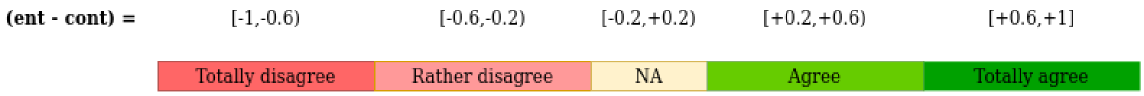


Figure 3: Map NLI scores to ALS answer types

Deriving an NLI Premise from the dialog. The NLI premise is derived from the input dialog using first, Content Transformation and second, Content Selection. As detailed in Section 5, we experiment with different ways of transforming and selecting content.

Deriving an NLI hypothesis from a question. To derive an NLI hypothesis from a question, we simply represent questions as statements (E.g., "I have pain regularly" instead of "Do you have pain regularly?"). Since the PBPI questionnaire questions are already in the form of a statement, we did not make any changes to them and used them as they are.

Deriving the answer. We use RoBERTa large [Liu et al., 2019] [1] fine-tuned on the MNLI dataset [Williams et al., 2018] to determine the entailment relation. We then derive the answer from the entailment relation between dialog and question as follows.

For Close Questions, we set the answer to "Yes" if NLI returns an entailment, "No" if it returns a contradiction and "NA" if it returns "neutral".

For ALS questions, we map the NLI result to agreement choices as follows. If "neutral" has the highest score, the answer is "NA". Else, the contradiction score is subtracted from the entailment score. The subtraction result lies in a range of (-1,1) which is uniformly divided into 5 segments corresponding to the 5 ALS answer types, as shown in figure 3.

5 NLI-oriented Dialog Pre-processing

We consider different ways of transforming and selecting dialog content.

We also study the impact of the NLI model used, comparing DeBERTa, the model used in Toudeshki et al. [2021], with RoBERTa [Liu et al., 2019], the model used in our approach.

The DeBERTa model [He et al., 2020][2] extends the BERT architecture with two innovative techniques: disentangled attention mechanism and an enhanced mask decoder. We compare AQF models with and without pre-processing and based on RoBERTa vs. DeBERTa, and find that whereas, when no pre-processing is applied, a DeBERTa model generally outperforms a RoBERTa-based model, the reverse is true when pre-processing is applied. This shows that while the improved DeBERTa-based, NLI model helps bridge the gap between dialog and text, explicit pre-processing still yields better results.

5.1 Content transformation

Null Transformation (CT_{null}) A null transformation baseline where we simply concatenate the turns of the input dialog. To encode the speaker information in each turn, the utterance is accompanied by the speaker role (patient/bot) at the beginning.

Summary (CT_{sum}) Pairs of adjacent turns are summarized, and the resulting summaries are concatenated. In this way, the input dialog is transformed into a sequence of two-turn summaries. We also tried summarizing the whole

[1] `https://huggingface.co/roberta-large-mnli`

[2] `https://github.com/microsoft/DeBERTa`

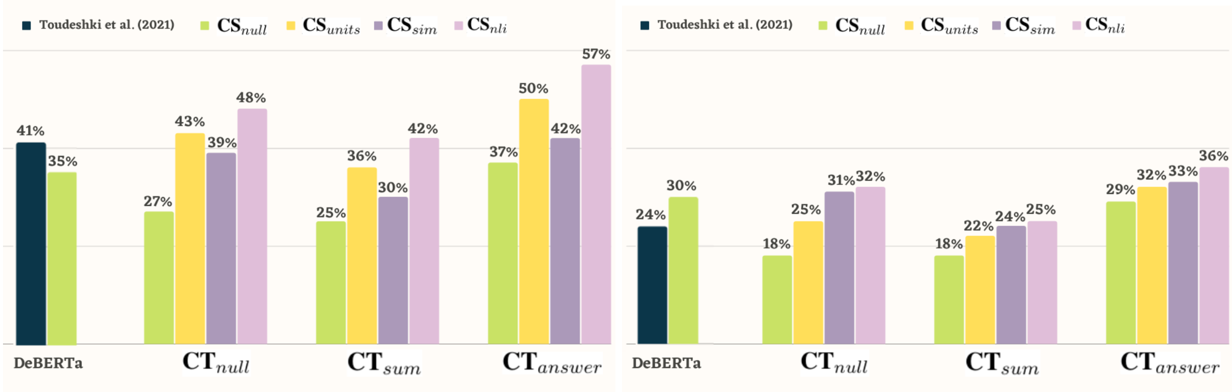


Figure 4: F1 macro average for Close Questions (on the left) and ALS questions (on the right) for the RoBERTa variant of our model. The two most left columns indicate the performance of [Toudeshki et al., 2021]'s model on their (dark blue) and our (light green) test set. The best results are obtained by the CT_{answer}, CS_{nli} model.

dialog in one go but found that applying summarization on each two turns rather than on the whole dialog gives better results. We use the **BART-large** model[3] [Lewis et al., 2020] fine-tuned on the News summarization corpus XSUM [Narayan et al., 2018] and on the dialog summarization corpus SAMSum [Gliwa et al., 2019]. The model achieves ROGUE-L score of 0.44 on SAMSum test set [4].

Long Answers (CT_{answer}) In information seeking dialog, adjacent turns often are question-answer pairs. Based on this observation, we map each pair of adjacent turns in the dialog into a single declarative sentence assuming that the first turn is a question (e.g., "Which drug did you take?"), the second is a short answer to that question (e.g., "Doliprane") and the sentence derived from the mapping is a long answer to the question (e.g., "I took Doliprane"). To learn this mapping, we fine-tune T5 [Raffel et al., 2019], a pre-trained encoder-decoder model, on two datasets of (question, incomplete answer, full answer) triples, one for wh- and one for yes-no (YN) questions. For wh-questions, we use 3,300 entries of the dataset consisting of (question, answer, declarative answer sentence) triples gathered by Demszky et al. [2018] using Amazon Mechanical Turk workers. For YN questions, we used the SAMSum corpus, [Gliwa et al., 2019] which contains short dialogs in chit-chat format. We created 1,100 (question, answer, full answer) triples by automatically extracting YN (question, answer) pairs from this corpus and manually associating them with the corresponding declarative answer. Data was splitted into train and test (9:1) and the fine-tuned model achieved 0.90 ROUGE-L score on the test set.

This fine-tuned model was applied to each two subsequent turns of the input dialogs, and the resulting declarative sentences were then concatenated to form the declarative transform of the whole dialog.

5.2 Content selection

The transformation operations described in the previous section yield sequences of dialog turns, two-turn summaries or full answers. We call these "input units" and consider three ways of pre-selecting the input units that will be used as premise when testing for entailment.

Null Content Selection (CS_{null}) A null content selection baseline where the premise is the concatenation of all the input units produced by the content transformation operations (dialog turns, sequence of two turn summaries, sequence of full form answers).

Unit-Based (CS_{units}). Each question is assessed against each input item. Given an input sequence I_n of length n, the answer a_i to a question q is then determined by aggregating the resulting entailment probabilities as follows:

- $a_i = NA$ if for all input items $i \in I_n$, the NA probability is highest.
- $a_i = Yes$ (resp. $a_i = No$) if for at least one item $i \in I_n$, the Yes (resp. No) probability is highest and the highest Yes (resp. No) probability is higher than the highest No (resp. Yes) probability.

[3] https://huggingface.co/Salesforce/bart-large-xsum-samsum
[4] https://paperswithcode.com/sota/abstractive-text-summarization-on-samsum

Two turns
bot: do you feel anxiety or stress during nights awakenings ? **patient:** I feel stressed during night awakenings although <u>I am not feeling guilty about being in pain</u>.
Generated summary Patient feels stressed during night awakenings although **he's not in pain**.

Table 1: An example of the summarization model performance on two subsequent turns, showing <u>missing</u> and **inconsistent** information in the output summary

Similarity (CS_{sim}). For each question q, we select a subset of input units that are semantically similar to q. We encode question and input units using SBERT[5] [Reimers and Gurevych, 2019] and compute *cosine similarity* for each (q, input unit) pair. We then select items whose similarity score is higher than 0.5, concatenate them and use the result as the NLI premise.

NLI (CS_{nli}). For each question q in the questionnaire, we select the input units that are related to q using the NLI model (RoBERTa-Large). Specifically, we select sentences which have an entailment or contradiction score higher than 0.5. All selected sentences are then concatenated to form the NLI premise.

5.3 Baseline and Comparison

Our baseline is the null method (CT_{null}+CS_{null}) i.e., the approach where question answering applies to the untransformed, unfiltered dialog. To compare our approach with Toudeshki et al. [2021], we also report the performance of their model on both their test set (10 dialogs) and on ours (100 dialogs).

6 Results

We evaluate our approach using macro and weighted F1 score.

6.1 How much does pre-processing help improve performance ?

Figure 4 shows the results for all combinations of our content transformation and selection methods[6].

Improvement over the baseline. Comparing our best model (CT_{answer}, CS_{nli}) with the no-preprocessing CT_{null}, CS_{null} baseline, we see (Figure 4) that pre-processing can multiply the macro and weighted F1 scores by two. The best pre-processing method combines a question+answer to sentence transformation (CT_{answer}) with the entailment-based content selection method (CS_{nli}).

Content transformation The CT_{answer} question+answer transform, which merges pairs of adjacent dialog turns into declarative statements, consistently yields the best results. A possible explanation is that this transform yields an input, a declarative sentence, which is consistent with the format of the training data used for NLI models.

Conversely, summarization (CT_{sum}) has the lowest performance. This could be due to errors such as hallucinations or omissions known to be produced by summarization systems Zhao et al. [2020]. Table 1 shows an example of such errors when applying the CS_{sum} transformation.

Content selection The NLI-based content selection method (CS_{nli}) consistently outperforms other content selection approaches. This is consistent with Toudeshki et al. [2021]'s findings that for automatic questionnaire filling in a medical setting, NLI models performed better on average on polar and ALS question types.

[5]`https://huggingface.co/sentence-transformers/paraphrase-distilroberta-base-v2`

[6]We first focus on the results of our RoBERTa based model and delay the comparison with DeBERTa based models to Section 6.4.

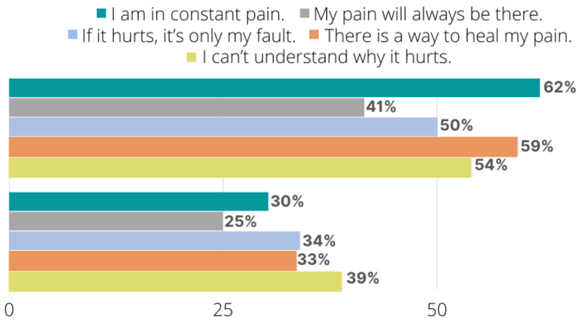


Figure 5: Break down of F1 macro average scores for each question based on out-performed model ($CT_{answer} + CS_{nli}$) results

	CQ					ALS						
support	NA 142	YES 228	NO 130	macro	weighted	NA 142	TD 54	RD 79	A 115	TA 110	macro	weighted
						CT_{null}						
CS_{null}	0.39	0.15	0.27	0.27	0.25	0.28	0.11	0.26	0.07	0.16	0.18	0.18
CS_{units}	0.33	0.48	0.46	0.43	0.43	0.33	0.25	0.02	0.07	0.58	0.25	0.25
CS_{sim}	0.52	0.55	0.10	0.39	0.42	0.54	0.07	0.09	**0.23**	0.60	0.31	0.36
CS_{nli}	0.34	0.60	0.48	0.48	0.50	0.34	0.29	0.08	0.21	**0.67**	0.32	0.34
						CT_{sum}						
CS_{null}	0.41	0.11	0.23	0.25	0.23	0.36	0.12	0.21	0.11	0.10	0.18	0.20
CS_{units}	0.32	0.33	0.43	0.36	0.35	0.32	0.23	0.06	0.02	0.44	0.22	0.23
CS_{sim}	0.49	0.40	0.02	0.30	0.32	0.51	0.00	0.05	0.21	0.46	0.24	0.30
CS_{nli}	0.37	0.43	0.46	0.42	0.42	0.31	0.28	0.10	0.09	0.48	0.25	0.26
						CT_{answer}						
CS_{null}	0.45	0.28	0.37	0.37	0.35	0.41	0.27	**0.27**	0.17	0.33	0.29	0.30
CS_{units}	0.40	0.59	0.51	0.50	0.52	0.41	0.29	0.17	0.16	0.57	0.32	0.33
CS_{sim}	**0.53**	0.60	0.13	0.42	0.46	**0.55**	0.10	0.20	**0.23**	0.59	0.33	**0.38**
CS_{nli}	0.45	**0.70**	**0.57**	**0.57**	**0.59**	0.42	**0.35**	0.16	**0.23**	0.65	**0.36**	**0.38**

Table 2: F1-Scores for RoBERTa for closed (CQ) and agreement Likert scale (ALS) question types; TD - totally disagree, RD - rather disagree, A - agree, TA - totally agree. CT: content transformation, CS: content selection.

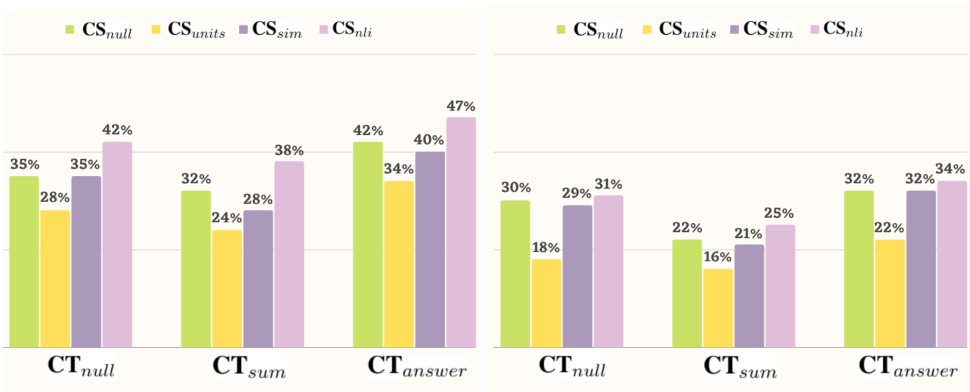


Figure 6: F1 macro average for the DeBERTa variant of our model on Closed Questions (CQ) on the left and Agreement Likert Scale (ALS) on the right. Test set of 100 dialogs with 10 questions each (5 yes/no questions and 5 ALS questions).

We also see that the second best performing content selection method varies depending on the question type. As CS_{unit} first filters question/item pairs with the highest probability, the method works well on CQ questions but struggles to handle more nuanced ALS questions, which leads to an overall drop in performance on ALS questions.

6.2 Impact of pre-processing on different question/answer types

Table 2 shows the results for all combinations of pre-processing steps for each question/answer type.

	CQ					ALS						
support	NA 142	YES 228	NO 130	macro	weighted	NA 142	TD 54	RD 79	A 115	TA 110	macro	weighted
CT_{null}												
CS_{null}	0.43	0.33	0.31	0.35	0.35	0.41	0.23	0.19	0.22	0.47	0.30	0.32
CS_{units}	0.15	0.29	0.40	0.28	0.28	0.15	0.21	0.07	0.00	0.45	0.18	0.17
CS_{sim}	0.51	0.45	0.09	0.35	0.37	**0.54**	0.03	0.05	**0.24**	0.60	0.29	0.35
CS_{nli}	0.34	0.51	0.40	0.42	0.43	0.29	0.23	0.17	0.21	**0.63**	0.31	0.32
CT_{sum}												
CS_{null}	0.40	0.29	0.26	0.32	0.31	0.37	0.16	0.11	0.16	0.31	0.22	0.25
CS_{units}	0.20	0.18	0.33	0.24	0.23	0.20	0.17	0.11	0.05	0.26	0.16	0.16
CS_{sim}	0.48	0.34	0.01	0.28	0.29	0.49	0.00	0.00	0.17	0.40	0.21	0.27
CS_{nli}	0.39	0.39	0.35	0.38	0.38	0.37	0.20	0.07	0.15	0.45	0.25	0.27
CT_{answer}												
CS_{null}	0.44	0.48	0.35	0.42	0.43	0.43	0.26	0.14	0.19	0.57	0.32	0.34
CS_{units}	0.19	0.45	0.39	0.34	0.36	0.19	0.20	0.07	0.10	0.55	0.22	0.23
CS_{sim}	**0.52**	0.51	0.16	0.40	0.42	0.53	0.09	0.15	0.20	0.61	0.32	**0.37**
CS_{nli}	0.40	**0.60**	**0.42**	**0.47**	**0.50**	0.41	**0.30**	**0.21**	0.16	**0.63**	**0.34**	0.36

Table 3: F1-Scores for DeBERTa for closed (CQ) and agreement Likert scale (ALS) question types; TD - totally disagree, RD - rather disagree, A - agree, TA - totally agree. CT: content transformation, CS: content selection.

Agreement answers (Yes, Totally agree) have the highest accuracy (about 70% in the best case) in both CQ and ALS questions, which suggests that the NLI model is better at confirming rather than rejecting a statement.

On *CQ questions*, various content selection methods have different impacts on each answer type.

CS_{sim} shows much lower (3-4 times lower) performance on 'No' class than on 'NA' or 'Yes', CS_{null} has higher accuracy for the 'NA' class than for 'Yes' or 'No' classes and CS_{nli} performs better on 'Yes' and 'No' answers than on 'NA'. Both CS_{nli} and CS_{units} gives the most balanced F1 distribution across classes.

For *ALS questions*, CS_{nli} and CS_{sim} show the best results. While the CS_{nli} model is best at identifying 'Totally agree' and 'Totally disagree' classes, CS_{sim} distinguishes well whether the answer is absent ('NA') or whether it belongs to the 'Totally agree' class.

Performance on ALS questions is always lower. This can be explained by choice of threshold that distinguishes classes 'Totally agree' and 'Agree' as well as 'Totally disagree' and 'Rather disagree'. As mentioned above, CS_{units} favorizes the extreme classes, which leads to a higher performance drop in comparison with CS_{sim} on ALS.

6.3 Break down of results for each question

Figure 5 presents the results of our best model (CT_{answer}+CS_{nli}) for each PBPI question separately.

The question "I am in constant pain." obtains highest score in CQ, while it performs poorly in ALS, demonstrating that the model is effective at detecting the presence of consistent pain but bad at predicting the level of agreement. The same behavior can be seen for the question "There is a way to heal my pain". On the other hand, for question "My pain will always be there" gets lowest score for both question types. The presence of the term "always" in the question turns it into a strong statement and consequently the model mostly rejects the statement unless it has been explicitly mentioned in the dialog.

6.4 Comparison with previous work and a different classifier (RoBERTa vs. DeBERTa)

Our model differs from previous work by Toudeshki et al. [2021] in two ways: it includes a pre-processing phase and uses the RoBERTa classifier, whereas Toudeshki et al. [2021] applies DeBERTa to the whole input dialog. We compare our model with (i) the same model using DeBERTa and (ii) Toudeshki et al. [2021]'s model both on their and our test set.

Comparison with previous work In Figure 4, the two columns on the far left show the performance of Toudeshki et al. [2021]'s model on two test sets: the test set they used (10 instances and 16 questions) and our test set (100 instances and 5 questions).

Unsurprisingly, Toudeshki et al. [2021]'s results vary with the test set: while they report F1 score of 41 for CQ and 24 for ALS questions on their test set, these change to 35 and 30 on ours.

We also see that Toudeshki et al. [2021]'s DeBERTa-based, no pre-processing model out-performs our RoBERTa-based, null-preprocessing model (CT_{null}, CS_{null}) on both test sets. We conjecture that this difference can be explained by DeBERTa's improved attention mechanism, which selects relevant information in the input dialog with respect to the hypothesis.

However, our best model outperforms Toudeshki et al. [2021]'s approach by 22 points F1 for CQ questions and 6 points for ALS questions, which indicates that pre-processing better helps bridge the gap between dialog and NLI-based QA.

DeBERTa vs. RoBERTa figure 6 and Table 3 show the result of our model when using DeBERTa instead of RoBERTa.

When using pre-processing, we see that the best RoBERTa model (CT_{answer}, CS_{nli}) outperforms the best DeBERTa model by 10 points F1 for CQ questions and 2 points for ALS questions.

Conversely, when no pre-processing is applied, the DeBERTa variant of our model outperforms the RoBERTa variant, which is consistent with the results discussed in the previous paragraph. For the DeBERTa variant, we observe that the CS_{null} baseline is no longer the lowest performing content selection approach, while results, but also the performance of CS_{units} and CS_{sim} becomes lower than the baseline (CS_{null}). This highlights the fact that the DeBERTa model performs better without weak content selection approaches. On the other hand, it can be seen that the impact of content selection and transformation approaches is significant in RoBERTa, although using a weaker classifier, and our model outperforms previous work. This shows that the proposed select-and-transform pre-processing approach improves results in both RoBERTa and DeBERTa, though this improvement is more significant in RoBERTa, suggesting that this latter model is more sensitive to the form and size of the input content.

7 Conclusion

In this paper, we studied how dialog pre-processing can impact the task of filling medical questionnaires based on patient-bot interactions. Our experimental results show that converting pairs of adjacent turns to declarative sentences and selecting input units based on their entailment relation with the question can significantly enhance performance.

A Ethics

Regarding Regulation (EU) 2017/745, described software is intended for general uses, even when used in a healthcare environment, it is intended for uses relating to lifestyle or well-being that do not constitute any a medical prediction and medical prognosis function without doctors validation or correction.

We gathered dialogs for experiments using Amazon Mechanical Turk. Because of the task's difficulty and estimated completion time, we set the initial reward at 1$. We assigned 0.5$ bonus for each key point mentioned by the user during the dialogue. If the user was successful in mentioning all five key points, he was awarded a bonus of 2.5$ in total.

B Experiment time estimation

The experiments were conducted with a laptop having Intel® Core™ i7-10610U CPU @ 1.80GHz * 8 and NVIDIA Quadro P520.

C Questionnaire

PBPI questionnaire statements are provided in table 4.

D Data Collection

Instructions used for data collection in Amazon Mechanical Turk and the interface are shown in figures 7, 8 and 9.

We requested the Turkers to converse with the heath-bot for at least 10 turns in total.

Id	Question
1	No one is able to tell me why it hurts.
2	I thought my pain could be healed, but now I'm not so sure.
3	There are times when it doesn't hurt.
4	My pain is difficult for me to understand.
5	**My pain will always be there.**
6	**I am in constant pain.**
7	**If it hurts, it's only my fault.**
8	I don't have enough information about my pain.
9	My pain is a temporary problem in my life.
10	I feel like I wake up with pain and fall asleep with it.
11	I am the cause of my pain.
12	**There is a way to heal my pain.**
13	I blame myself when it hurts.
14	**I can't understand why it hurts.**
15	One day, again, I won't have any pain at all.
16	My pain varies in intensity but it is always present with me.

Table 4: List of questions in PBPI questionnaire

Task Description

In this task, you are going to talk to a chatbot about health and quality of your life.

You are supposed to **play the role of a chronic pain patient**, and **share your pain with the bot**.
What is chronic pain? Doctors often define chronic pain as any pain that lasts for 3 to 6 months or more. Chronic pain can have real effects on day-to-day life and mental health.

It is very important that you **mention about all follwing key points during your conversation** (in a seamless way):

1. (Constantly/Temporarly) in pain
2. (Having/Losing) hope for getting healed
3. (Feeling/Not feeling) guiltiness that the pain is your fault
4. (Possibility/Impossibility) of healing
5. (Understanding/Not understanding) the reason of having pain

Try to **give an implicit and seamless reference to these keypoints** in the dialogue (with considering the flow of conversation). **Prevent using the same wording** in your messages.

**** BONUS ****

Playing the role of a chronic pain patient and mentioning each keypoint will get 0.5 $ bonus. By metioning all keypoints you will get 2.5 $ bonus (do not use the same wording).

note1: It is an information seeking conversation and **you are not expected to ask questions from the bot**.
note2: Wait for the bot message to be appeared completely, and then reply.
note3: When it is your time to reply, **reply only once**.

Figure 7: Instructions (part 1)

Lead the conversation

The chatbot is not developed to ask you explicitly about these key points. Therefore, you have to mention them creatively during the dialog flow.
To make it easier for you, we have given you the authority of **controling chatbot messages**. You can direct the conversation by **changing chatbot reply**. To do that, you can **click on the "next" button** (below the bot message) to change the chatbot utterance and if you found it good enough you can just continue the conversation.

Annotating each user reply

After you entered your answer, you will notice **5 checkpoints apear below your answer**. Each check point refers to each of the keypoints. **If one or mutliple keypoints have been mentioned in your answer (implicitly or explicitly) choose the related checkpoints**.
Please keep in my mind that you have to **fill check points before entering your next response**. They would be disabeled afterwards.

End the conversation

To end the conversation, you can click on **green "Submit" button**. But before that, **wiat for the bot message to be appeared completely, and then press the submit button**.
If you click on the button before reaching to the minimum number of turns (5 messages each user), you will receive an alert error message and be taken back to the conversation to complete the task.

Figure 8: Instructions (part 2)

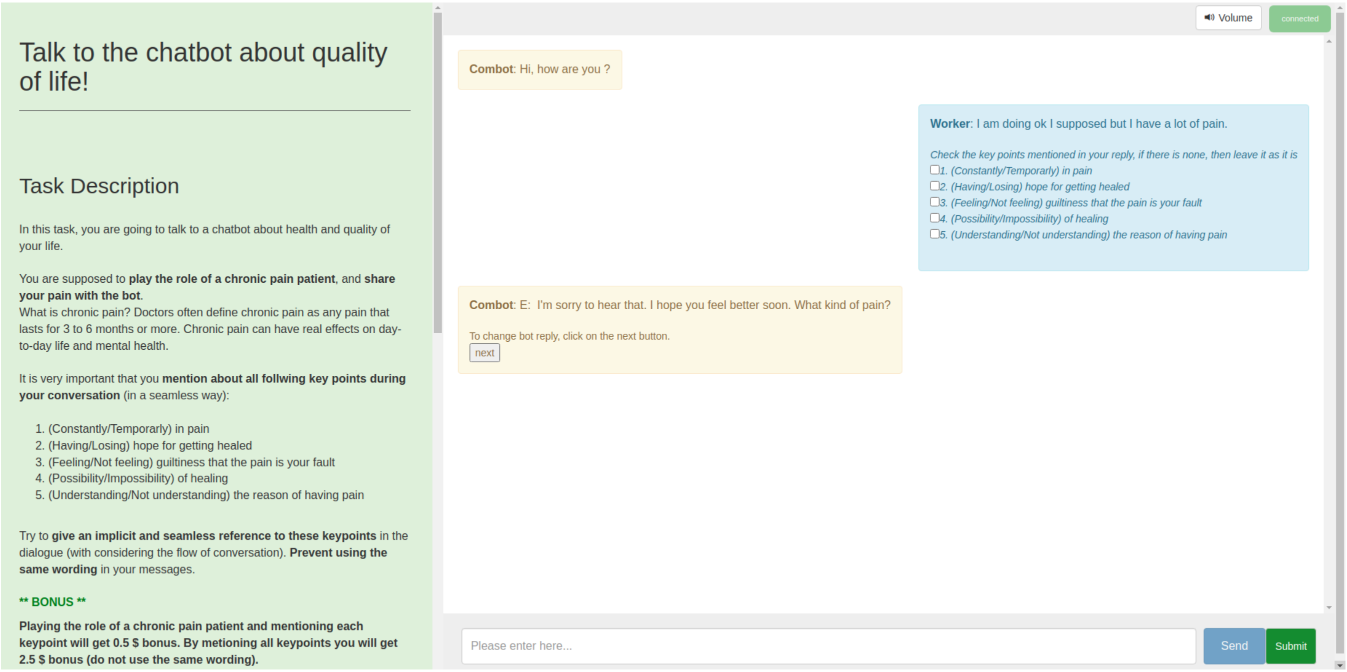


Figure 9: Interface

References

Kate Pearce, Tiffany Zhan, Aneesh Komanduri, and Justin Zhan. A comparative study of transformer-based language models on extractive question answering. *arXiv preprint arXiv:2110.03142*, 2021.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. Commonsense for generative multi-hop question answering tasks. *arXiv preprint arXiv:1809.06309*, 2018.

David A. Williams and Beverly E. Thorn. An empirical assessment of pain beliefs. *Pain*, 36(3):351–358, 1989. doi:10.1016/0304-3959(89)90095-x.

Farnaz Ghassemi Toudeshki, Philippe Jolivet, Alexandre Durand-Salmon, and Anna Liednikova. Zero-shot clinical questionnaire filling from human-machine interactions. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 51–62, 2021.

Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences*, 10(21):7640, 2020.

Jiangtao Ren, Naiyin Liu, and Xiaojing Wu. Clinical questionnaire filling based on question answering framework. *International Journal of Medical Informatics*, 141:104225, 2020.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DialoGPT: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. Modeling multi-turn conversation with deep utterance aggregation. *arXiv preprint arXiv:1806.09102*, 2018.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 111–120, 2019.

Zhuosheng Zhang, Junlong Li, and Hai Zhao. Multi-turn dialogue reading comprehension with pivot turns and knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1161–1173, 2021.

Junlong Li, Zhuosheng Zhang, and Hai Zhao. Knowledgeable dialogue reading comprehension on key turns. *arXiv preprint arXiv:2004.13988*, 2020.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285, 2019.

H Jang, E Rempel, D Roth, G Carenini, and NZ Janjua. Tracking COVID-19 discourse on twitter in North America: Topic modeling and aspect-based sentiment analysis. *Journal of Medical Internet Research*, 2021.

Mary Sun. *Natural Language Processing for Health System Messages: Deep Transfer Learning Approach to Aspect-Based Sentiment Analysis of COVID-19 Content*. PhD thesis, Harvard University, 2022.

Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. Gathering information and engaging the user combot: A task-based, serendipitous dialog model for patient-doctor interactions. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 21–29, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/N18-1101`.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2020.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.703. URL `https://www.aclweb.org/anthology/2020.acl-main.703`.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Dorottya Demszky, Kelvin Guu, and Percy Liang. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*, 2018.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `http://arxiv.org/abs/1908.10084`.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online, November 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.203. URL `https://aclanthology.org/2020.findings-emnlp.203`.