COREFERENCE RESOLUTION: DRAFT (ASSIGNMENT 1)

Matej Klemen
Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, 1000 Ljubljana
mk3141@student.uni-lj.si

Blažka Blatnik
Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, 1000 Ljubljana
bb3172@student.uni-lj.si

Martin Čebular
Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, 1000 Ljubljana
insert email here

March 29, 2020

ABSTRACT

Abstract content (to be added later).

1 Introduction

Coreference resolution is a task where the goal is to identify and group together all entity mentions that refer to a common entity in the text. Generally, the task can be thought of as a combination of mention detection and mention clustering and many approaches explicitly perform these two steps when doing coreference resolution. The mention detection step deals with the detection of all entities that refer to some entity in the text. Mention clustering then divides the entities into groups based on the entity they refer to.

In our work, we focus on coreference resolution for Slovene language, which has not been the subject of much research. For our experiments, we use the coref149 dataset [1]which is a corpus of 149 documents annotated with coreference information. We first evaluate an existing end-to-end approach and compare it to an approach performing the two steps of coreference resolution separately. Additionally, we modify the end-to-end approach, adding explicit information about additional morphological properties such as part of speech tags, and evaluate it.

The rest of the paper is structured as follows. In Section 2 we provide an overview of existing approaches to coreference resolution. In Section 3 we describe the used end-to-end approach to coreference resolution, as well as our modifications of it. In Section 4 we provide the results of our work, which we then discuss in Section 5. In Section 6 we summarize our work and provide some possible directions for further research.

2 Related Work

Most coreference resolution systems deal with two tasks: mention detection and mention clustering.

Traditionally, the task of mention detection was performed via analysis of parse trees and the use of heuristics [2]. The problem with hand-crafting rules is that they are language-specific and can be hard to define for less-researched languages.

The mention clustering task is where the approaches differ substantially. One approach is to treat it as a binary classification problem, where the goal is to determine whether two mentions are coreferent or not [3] [4]. The problem of this approach is that it treats all coreference candidates independently, so it cannot choose the most probable candidate when multiple valid ones exist. A different way to do mention clustering, which solves this problem, is with mention

ranking [5]. In this approach, candidates for coreference are scored in some way and the best scoring candidate is proclaimed as the coreferent mention. The benefit of this approach is that it does not consider candidates in isolation, but rather in comparison to other mentions. An approach which takes this even further is the entity-mention approach [5]. Here, the models are trained to determine whether the currently considered mention belongs to some preceding coreference cluster [6]. In our work, we make use of the span ranking approach, which is a modification of the mention ranking approach.

Recently, an end-to-end approach was introduced [7], where the two steps are combined and learned together using deep neural networks. This approach considers all spans of tokens up to specific length as candidates for coreference. The spans are then scored in isolation and as mention pairs to produce a final coreference score, which is used in the span ranking coreference resolution framework. Because the approach only considers pairs of mentions when scoring candidates, it can produce globally inconsistent clusters. An approach by Lee et al. [8] solves this by iteratively refining the obtained coreference clusters.

The end-to-end approach was further researched and improved upon, for example by using more sophisticated contextualized embeddings [9], but is the case in most of the other areas in nature language processing, the research is mostly focused on the English language. Some examples of research done for other languages include a Lithuanian rule-based approach [10] and approaches for Polish [11] and Basque [12] that use neural networks.

3 Methods

4 Results

5 Discussion

6 Conclusion

References

[1] Slavko Žitnik and Marko Bajec. Coreference resolution for slovene on annotated data from coref149. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 6(1):37–67, Jun. 2018.

[2] Giuseppe Attardi, Maria Simi, and Stefano Dei Rossi. TANL-1: Coreference resolution by parse analysis and similarity clustering. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 108–111. Association for Computational Linguistics, July 2010.

[3] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

[4] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[5] Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China, July 2015. Association for Computational Linguistics.

[6] Xiaofeng Yang, Jian Su, GuoDong Zhou, and Chew Lim Tan. An NP-cluster based approach to coreference resolution. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 226–232, Geneva, Switzerland, aug 23–aug 27 2004. COLING.

[7] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[8] Kenton Lee, Luheng He, and Luke Zettlemoyer. Higher-order coreference resolution with coarse-to-fine inference. *CoRR*, abs/1804.05392, 2018.

[9] Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. BERT for coreference resolution: Baselines and 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 5803–5808, Hong Kong, China, November 2019. Association for Computational Linguistics.

[10] Voldemaras Žitkus, Rita Butkienė, Rimantas Butleris, Rytis Maskeliunas, Robertas Damasevicius, and Marcin Woźniak. Minimalistic approach to coreference resolution in lithuanian medical records. *Computational and Mathematical Methods in Medicine*, 2019:1–14, 03 2019.

[11] Bartłomiej Nitoń, Paweł Morawiecki, and Maciej Ogrodniczuk. Deep neural networks for coreference resolution for polish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[12] Gorka Urbizu, Ander Soraluze, and Olatz Arregi. Deep cross-lingual coreference resolution for less-resourced languages: The case of basque. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 35–41, Minneapolis, USA, June 2019. Association for Computational Linguistics.