Coreference resolution for Slovene language: draft

Matej Klemen, Blažka Blatnik, Martin Čebular
University of Ljubljana
Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana
{mk3141, bb3172, mc0239}@student.uni-lj.si

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 assume the mentions are already detected and focus only on mention clustering. Specifically, we focus on Slovene language, which has not yet been the subject of much research. For our experiments, we use the coref149 dataset (Žitnik and Bajec, 2018), which is a corpus of 149 documents annotated with coreference information. We implement a baseline approach for mention clustering using a linear regression scorer and various handcrafted features. We then try to improve its performance using neural network based approaches. Our goal is to see if the process of designing linguistic features, which requires a lot of domain knowledge, can be partially or fully replaced by models leveraging various types of embeddings. We evaluate our approaches on the dataset and compare their strengths and weaknesses.

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 and evaluate two approaches to coreference resolution, as well as our modifications of them. 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 (Attardi et al., 2010). 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 (Soon et al., 2001) (Ng and Cardie, 2002). 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 (Wiseman et al., 2015). 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 (Wiseman et al., 2015). Here, the models are trained to determine whether the currently considered mention belongs to some preceding coreference cluster (Yang et al., 2004). 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 to coreference resolution was introduced (Lee et al., 2017), 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. (Lee et al., 2018) 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 (Joshi et al., 2019), but as 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 (Žitkus et al., 2019) and approaches for Polish (Nitoń et al., 2018) and Basque (Urbizu et al., 2019) that use neural networks.

3 Methods

3.1 Baseline model

Our baseline model is a mention pair scorer based on linear regression and handcrafted features. Scores are obtained for every antecedent candidate appearing in the document and then normalized using softmax function. The candidate with the highest score is chosen as the model prediction. In training phase we minimize the average cross-entropy of the ground truth antecedents. We optimize our model using stochastic gradient descent. For constructing the features, we enrich our dataset with additional metadata from the ssj500k (Krek et al., 2019) dataset.

The features we use in our baseline model are based on already-proven ones, reported in existing literature (Žitnik et al., 2014). They are described in Table 1. Categorical features are encoded into binary ones using one-hot encoding.

Table 1: Features used in our baseline model.

Feature	**Description**
string_match	exact match for pronouns or match in lemmas
same_sentence	are both mentions in same sentence
same_gender	one-hot encoded vector for values: same gender, different gender, can't determine
same_number	one-hot encoded vector for values: match in number, don't match in number, can't determine
is_appositive	both mentions have noun-related tag and previous mention is followed by comma
is_alias	one mention is a subset of another
is_prefix	one mention is prefix of another
is_suffix	one mention is suffix of another
is_reflexive	one mention is followed by another that is reflexive pronoun
jaro_winkler_dist	similarity value between two mentions according to Jaro-Winkler metric

4 Results

We split the dataset into a training, validation and test set in ratio 70%:15%:15%. The validation set is used to select the best hyperparameters for our model as well as for regularization. When the validation loss stops decreasing, the training process is interrupted and the model is evaluated on the test set.

For model evaluation MUC (Vilain et al., 1995), BCubed (Bagga and Baldwin, 1998), CEAFe (Luo et al., 2004) and CoNLL 2012 scores are used. For each metric, three numbers are reported: precision, recall and F1 score. CoNLL 2012 metric is defined as the average of the other three metrics. Besides numerical metrics we also develop and use a visualization of the model predictions, shown in Figure 1. It shows a comparison between ground truth coreference clusters and the predicted clusters of our model. This helps us understand how our model clusters mentions and where it fails.

The results are shown in Table 2.

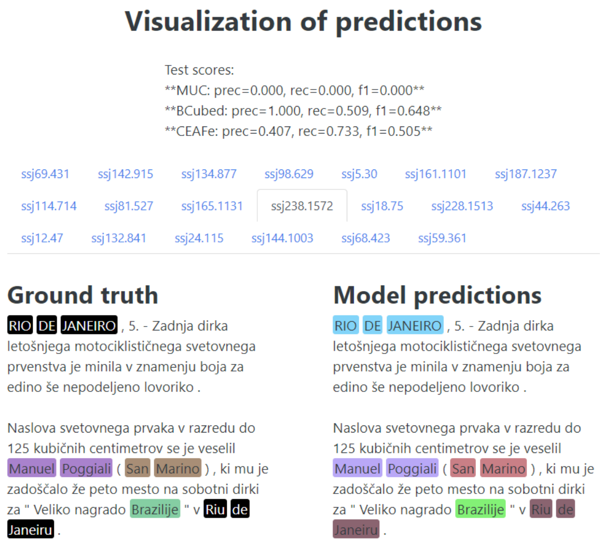

Figure 1: Visual representation of predictions next to ground truth.

Table 2: Results of our approaches. Each metric consists of precision, recall and F1 score.

Model	MUC	B3	CEAFe
Baseline	0.570; 0.850; 0.666	0.983; 0.660; 0.773	0.755; 0.345; 0.445

Todo: results of baseline model, compared with results from two basic approaches (all in one cluster, every mention in separate cluster) and in the end improved with NN model.

5 Discussion

6 Conclusion

References

Giuseppe Attardi, Maria Simi, and Stefano Dei Rossi. 2010. 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.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. 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. Association for Computational Linguistics.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2019. Training corpus ssj500k 2.2. Slovenian language resource repository CLARIN.SI.

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

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

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 135. Association for Computational Linguistics.

Vincent Ng and Claire Cardie. 2002. 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. Association for Computational Linguistics.

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

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

Gorka Urbizu, Ander Soraluze, and Olatz Arregi. 2019. 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. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. 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. Association for Computational Linguistics.

Xiaofeng Yang, Jian Su, GuoDong Zhou, and Chew Lim Tan. 2004. 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. COLING.

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

Slavko Žitnik and Marko Bajec. 2018. Coreference resolution for Slovene on Annotated Data from coref149. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 6(1):37–67.

Slavko Žitnik, Lovro Šubelj, and Marko Bajec. 2014. Skipcor: Skip-mention coreference resolution using linear-chain conditional random fields. *PLOS ONE*, 9(6):1–14.