Coreference Resolution

Working Title, First Draft: Content, not Wording

Patrick Kahardipraja, Olena Vyshnevska

1 Introduction

Coreference Resolution (CR) is a vital part of Natural Language Understanding and Natural Language Processing (NLP). It's applications include information extraction, question answering, summarization, machine translation, dialog systems etc. The task consists of mention detection and mention clustering. Some systems only focus on one part, while others propose end-to-end solutions.

The notion of mention – a span of text referring to some entity – is central to the task of CR. Mentions can be one of the three kinds: pronouns, named entity, and noun phrases. When several mentions *corefer* they refer to the same entity.

Since the seventies automated solutions for coreference resolutions have been researched (Woods, Kaplan, and Nash-webber 1972; Winograd 1972; Hobbs 1978). Some of the challenges central to the field include semantic and syntactic agreement between mentions; encoding non-local dependencies; paraphrasing of repetitions.

Since the introduction of pre-trained BERT model (Devlin et al. 2019) there has been a lot of research showing how the model outperforms previous state-of-the-art task-specific NLP models. cite some papers here Joshi et al. (2019) has introduced BERT models into coreference resolution tasks. However, the researchers have also encoded other information on top of BERT embeddings.

Our first research question is based on the recent research which shows that the BERT model itself encodes a significant amount of semantic and syntactic information. Therefore, we decided to test how well the span representations using BERT embeddings from the model by Joshi et al. (2019) finetuned on Ontonotes dataset (Pradhan et al. 2012) can encode coreference resolution information without any layers of additional information about a given sequence.

As a baseline, we used original pre-trained BERT embeddings. To encode a mention span, we used a convolutional layer and a self-attention layer.

Our second research question, which naturally follows from the first one, whether the span representations encoded with a pre-trained BERT model simply modelling local coreference? Will it be able to encode non-local coreference dependencies?

We found that the span represenations from within the coreference model by Joshi et al. (2019) reach *insert some numbers*. The baseline model is at *insert numbers*. This findings suggests that the pre-trained BERT model is a powerful tool for encoding information relevant to coreference resolution.

For our second research question (non-local VS local dependencies) we found that ... what we found for local non-locat

Our code is publicly available on GitHub. [1].

[1] https://github.com/alyonavyshnevska/bert_for_coreference_resolution

2 Related Work

2.1 Approaches to Coreference Resolution

Architectures for coreference resolution models are typically categorized as 1) mention-pair classifiers (Ng and Cardie 2002; Bengtson and Roth 2008), 2) mention-ranking classifiers (Durrett and Klein 2013; Wiseman, Rush, S. Shieber, et al. 2015; Clark and Manning 2016), 3) entity-level models (Haghighi and Klein 2010; Wiseman, Rush, and S. M. Shieber 2016), or 4) latent-tree models (Fernandes, Santos, and Milidiu 2012; Martschat and Strube 2015). Earlier solutions have been feature-based, while in the recent years neural classifiers have been particularly successful.

Nowadays a widely-adopted approach to coreference resolution are end-to-end models that perform mention detection, anaphoricity, and coreference jointly (Daniel Jurafsky 2019). Beforehand, the rule-based systems were in use. Notable mentions are work by Hobbs (1978), Lappin and Leass (1994). Hobbs (1978) invented a tree-search algorithm for identifying reference with robust results, which started a long series of syntax-based methods. Lappin and Leass (1994) combined syntactic and other features by assigning weights to those features and summing these up to score candidate mentions. Kennedy and Boguraev (1996) optimised the approach to avoid full syntactic parsing.

Yang et al. (2003) and Iida et al. (2003) helped establish mention-ranking approaches as influential solutions in the early 2000's. Ng (2005) pioneered the rise of end-to-end solutions. While rule-based systems have witnessed a short-lived revival in 2010s (Zhou, Ticrea, and Hovy 2004; Haghighi and Klein 2009), their struggles with semantic understanding for the models led to their eventual demise. The rise of neural architectures that dominated the NLP in the recent decade has inevitably established itself in coreference.

2.2 State of the Art Coreference Models

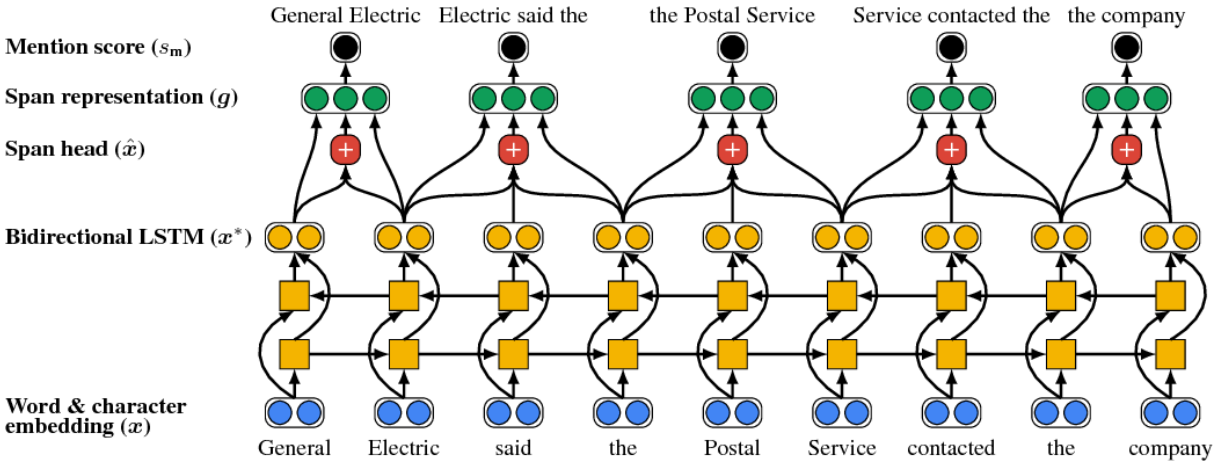

Figure 1: Figure 1: First step of the end-to-end coreference resolution model, which computes embedding representations of spans for scoring potential entity mentions. Low-scoring spans are pruned, so that only a manageable number of spans is considered for coreference decisions. In general, the model considers all possible spans up to a maximum width, but we depict here only a small subset. (Lee, He, Lewis, et al. 2017)

Lee, He, Lewis, et al. (2017) have proposed an span-ranking approach, which the authors describe as most similar to mention ranking, with reasoning over a larger space by detecting mentions and predicting coreference jointly in one end-to-end model.

The authors showed that from the space of all possible spans their model implicitly learns to produce meaningful mention candidates. A head-finding attention mechanism also learns a task-specific preference for head words, which has a strong correlation with head-word definitions in rule-based systems.

Building on to of the span-ranking neural architechture in Lee, He, Lewis, et al. (2017), Lee, He, and Zettlemoyer (2018) proposed a model that captures higher order interactions between spans in predicted clusters. It also alleviates the additional computational cost of higher-order inference due to a coarse-to-fine approach, which allows the model to at first compute a rough sketch of likely antecedents, and only at a later stage apply a more exhaustive inference. Hence, the coarse-to-fine approach expands the set of coreference links that the model is capable of learning.

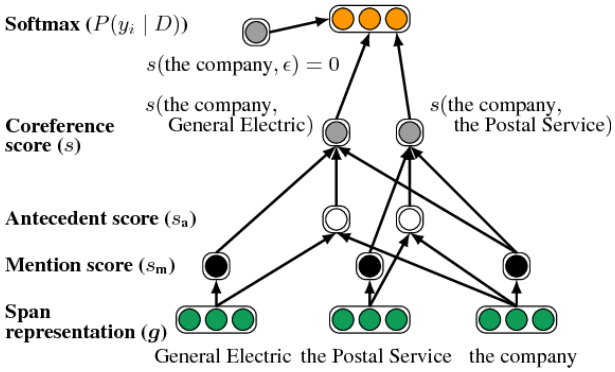

Figure 2: Second step of end-to-end model by Lee, He, Lewis, et al. (2017). Antecedent scores are computed from pairs of span representations. The final coreference score of a pair of spans is computed by summing the mention scores of both spans and their pairwise antecedent score.

To encode the span representations Lee, He, Lewis, et al. (2017) use bidirectional LSTM (Hochreiter and Schmidhuber 1997) to encode the lexical information of the inside and outside of each span. Lee, He, and Zettlemoyer (2018) additionaly use the newly published ELMo embedding representations by Peters et al. (2018) at the input to the LSTMs.

Joshi et al. (2019) took the architechture of Lee, He, and Zettlemoyer (2018) and improved the performence of the model further by replacing the LSTM-based encoder with the BERT transformer Devlin et al. (2019).

Devlin et al. (2019) made a significant break-through with their pre-trained BERT model. It can be finetuned with one additional output layer to create state-of-the-art models for a wide range of NLP tasks, such as question answering and language inference, without substantial taskspecific architecture modifications. BERT's training examples consist of 128 and 512 word pieces. Such passage-level training has been an important improvement over the previous methods. It helps the model learn dependencies over text sequences longer than one sentence, such as the previous state of the art model ELMo (Peters et al. 2018). Two model sized have been presented by Devlin et al. (2019): BERT-base(12 transformer blocks, hidden size 768, 12 self-attention heads, total Parameters=110M) and BERT-large (24 transformer blocks, hidden size 1024, 16 self-attention heads, total Parameters=340M).

Joshi et al. (2019) treat the first and last word-pieces in a mention (concatenated with the attended version of all word pieces in the span) as span representations. The researchers test both models with BERT-base and BERT-large transformers. With this change to the c2f-coref model the authors gain an additional 3.9% improvement on the OntoNotes compared to the already high results of Lee, He, and Zettlemoyer (2018). A quantitative analysis is shown in figure 3. A qualitative analysis done by the authors suggests that BERT-large (unlike BERT-base) is significantly better at distinguishing between related yet distinct entities or concepts.

2.3 Coreference Arc Prediction Task

Liu et al. (2019) found that frozen contextual represenations fed into linear models can show similar levels of performance as state of the art task-specific models on many NLP tasks. The authors

	MUC			B^3			$CEAF_{\phi_4}$			
	P	R	F1	P	R	F1	P	R	F1	Avg. F1
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
(Clark and Manning, 2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
(Wiseman et al., 2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Clark and Manning (2016)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
e2e-coref (Lee et al., 2017)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
c2f-coref (Lee et al., 2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
Fei et al. (2019)	**85.4**	77.9	81.4	**77.9**	66.4	71.7	70.6	66.3	68.4	73.8
EE (Kantor and Globerson, 2019)	82.6	**84.1**	83.4	73.3	**76.2**	74.7	72.4	71.1	71.8	76.6
BERT-base + c2f-coref (independent)	80.2	82.4	81.3	69.6	73.8	71.6	69.0	68.6	68.8	73.9
BERT-base + c2f-coref (overlap)	80.4	82.3	81.4	69.6	73.8	71.7	69.0	68.5	68.8	73.9
BERT-large + c2f-coref (independent)	84.7	82.4	**83.5**	76.5	74.0	**75.3**	**74.1**	**69.8**	**71.9**	**76.9**
BERT-large + c2f-coref (overlap)	85.1	80.5	82.8	77.5	70.9	74.1	73.8	69.3	71.5	76.1

Figure 3: OntoNotes: BERT improves the c2f-coref model on English by 0.9% and 3.9% respectively for base and large variants. (Joshi et al. 2019)

use probing models, also known as auxiliary or diagnostic classifiers (Shi, Padhi, and Knight 2016; Kádár, Chrupała, and Alishahi 2017) to analyse the linguistic information within contextual word representations. To test how the probing model performs in comparison to ELMo, the authors tested the models on a coreference arc prediction task, where the model predicts whether two mentions corefer. The Ontonotes dataset was used. To train the probing model the authors generate negative examples, where mentions that do not corefer are fed into the probing model alongside the mentions that do corefer. In our work we will also build our experiments similar to coreference arc prediction tasks suggested by the authors.

2.4 Probing Model

Tenney et al. (2019) introduced edge probing task design investigated and found that existing models trained on language modeling and translation produce strong representations for syntactic phenomena, but only offer comparably small improvements on semantic tasks over a non-contextual baseline. For coreference resolution, the models had to determining whether two spans of tokens refer to the same entity. In order to investigate how good the contextual word representation models can model long-range dependencies, the authors used a fixed-width convolutional layer on top of the word representations to build spans of word pieces. The authors concluded that using the CNN layers on top of BERT-large pretarined model has performed particularly well on difficult semantic tasks, such as coreference resolution. Hence, in out work we will also follow the approach of the authors to construct our baseline.

3 Approach

A benefit of using neural models for coreference resolution is their ability to use word embeddings to capture similarity between words, a property that many traditional feature-based models lack.

3.1 Span Representation and Long-Range Coreference

We attempt to investigate to what extent the span representations proposed using BERT (Devlin et al. 2019) embeddings in Joshi et al. (2019) can encode coreference information, and whether

it is able to encode non-local coreference phenomena or is it just simply modeling local coreference.

In order to analyse this, we consider 2 kinds of span representations: 1. BERT-based span representations finetuned on OntoNotes in Joshi et al. (2019) with first and last word-pieces (concatenated with the attention version of all word pieces in the span). 2. pre-trained BERT embeddings (not finetuned on OntoNotes) for all tokens within the mention span, which is then passed through a convolutional layer (with kernel width of 3 and 5) to incorporate the local context and followed by self-attention pooling operator to produce a fixed-length span representations. This is to model head words, inspired by approach from Tenney et al. (2019).

3.2 Arc Prediction Task

Both span representations will be then used as inputs for coreference arc prediction task Liu et al. 2019, where a probing model (in this case a simple FFNN) is used to predict coreference relations. The probing model is designed with limited capacity to focus on what information that can be extracted from the span representations. The probing model itself has a sigmoid output layer, which is trained to minimize binary cross entropy. Each negative samples (*w_entity*, *wb*) will be generated for every positive samples (*wa*, *wb*) where *wb* occurs after *wa* and *w_entity* is a token that occurs before *wb* and belong to a difference coreference cluster, to ensure a balanced data. By comparing the performance of the probing model using these two span representations, we can hypothesize to what extent that the proposed span representation in Joshi et al. (2019) can capture coreference information. We will also experiment with mention span separation distance to see how the probing model performs and whether if there is a degradation of accuracy and F1 score of the probing model with distant spans.

3.3 Data

For our experiments we use the English coreference resolution data from the CoNLL-2012 shared task based on the OntoNotes dataset (Pradhan et al. 2012). It consists of about one million words of newswire, magazine articles, broadcast news, broadcast conversations, web data and conversational speech data, and the New Testament. The dataset is split into 2802 training documents, 343 validation documents, and 348 test documents. On average, the training documents contain 454 words. The largest document contains a maximum of 4009 words. The main evaluation is the average F1 score of three metrics – *MUC*, B^3 and $CEAF_\phi 4$ on the test set according to the official CoNLL-2012 evaluation scripts. For our experiments we generate the positive and negative examples of coreferent expressions as training data for the probing model.

4 Experiments

4.1 Implementation and Hyperparameters

. We extract the span representations from the coreference model by Joshi et al. (2019). The model has been fine tuned on OntoNotes English data for 20 epochs using a dropout of 0.3, and learning rates of $1 * 10^{-5}$ and $2 * 10^{-4}$ with linear decay for the BERT parameters and the task parameters respectively. A batch size of 1 document has been used.

4.2 Baseline

5 Results

- some nice plots here
- some tables here

6 Discussion

- long long analysis of what we've seen
- generalisations, parallels
- what do this results tell us about coref and nlp in general
- discussion of why they are the way the are

We found that the model by Joshi et al. (2019) has performed very well. However, since the research is heavily based on the English language, it may be a phenomena particular to this language. The nominal declension in Ukrainian, for example, has seven cases; adjectives, pronouns have gender specific forms. Therefore, a similar analysis should be done for other languages before one can generalize the findings universally.

6.1 Future Work

- everything we don't have the time for
- mention other corpora that could be used for finetuning (Winogrande, GAP...)

7 Conclusion

- so what have we learned about coref in general and local dependencies in particular

References

Bengtson, Eric and Dan Roth (2008). "Understanding the Value of Features for Coreference Resolution". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pp. 294–303. URL: https://www.aclweb.org/anthology/D08-1031.

Clark, Kevin and Christopher D. Manning (2016). "Deep Reinforcement Learning for Mention-Ranking Coreference Models". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2256–2262. DOI: 10.18653/v1/D16-1245. URL: https://www.aclweb.org/anthology/D16-1245.

Daniel Jurafsky, James H. Martin (2019). *Speech and Language Processing. Draft of October2, 2019*. URL: https://web.stanford.edu/~jurafsky/slp3.

Devlin, Jacob et al. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Durrett, Greg and Dan Klein (2013). “Easy Victories and Uphill Battles in Coreference Resolution”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1971–1982. URL: https://www.aclweb.org/anthology/D13-1203.

Fernandes, Eraldo, Cicero dos Santos, and Roy Milidiu (2012). “Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution”. In: *Joint Conference on EMNLP and CoNLL - Shared Task*. Jeju Island, Korea: Association for Computational Linguistics, pp. 41–48. URL: https://www.aclweb.org/anthology/W12-4502.

Haghighi, Aria and Dan Klein (2009). “Simple Coreference Resolution with Rich Syntactic and Semantic Features”. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*. EMNLP 09. Singapore: Association for Computational Linguistics, pp. 1152–1161. ISBN: 9781932432633.

– (2010). “Coreference Resolution in a Modular, Entity-Centered Model”. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California: Association for Computational Linguistics, pp. 385–393. URL: https://www.aclweb.org/anthology/N10-1061.

Hobbs, Jerry R. (1978). “Resolving pronoun references”. In: *Lingua* 44.4, pp. 311–338. ISSN: 0024-3841. DOI: https://doi.org/10.1016/0024-3841(78)90006-2.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long Short-Term Memory”. In: *Neural Comput.* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/ne

Iida, Ryu et al. (2003). “Incorporating contextual cues in trainable models for coreference resolution”. In: *In Proceedings of the EACL Workshop on The Computational Treatment of Anaphora*. Citeseer.

Joshi, Mandar et al. (2019). “BERT for Coreference Resolution: Baselines and Analysis”. In: *Empirical Methods in Natural Language Processing (EMNLP)*.

Kádár, Akos, Grzegorz Chrupała, and Afra Alishahi (2017). “Representation of linguistic form and function in recurrent neural networks”. In: *Computational Linguistics* 43.4, pp. 761–780.

Kennedy, C. and B.K. Boguraev (1996). “Anaphora for everyone: Pronomial anaphora resolution without a paser.” In: pp. 113–118.

Lappin, S. and H. Leass (1994). “An algorithm for pronominal anaphora resolution”. In: vol. 20(4), pp. 535–561.

Lee, Kenton, Luheng He, Mike Lewis, et al. (2017). “End-to-end Neural Coreference Resolution”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 188–197. DOI: 10.18653/v1/D17-1018. URL: https://www.aclweb.org/anthology/D17-1018.

Lee, Kenton, Luheng He, and Luke Zettlemoyer (2018). “Higher-Order Coreference Resolution with Coarse-to-Fine Inference”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 687–692. DOI: 10.18653/v1/N18-2108. URL: https://www.aclweb.org/anthology/N18-2108.

Liu, Nelson F. et al. (2019). “Linguistic Knowledge and Transferability of Contextual Representations”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1073–1094. DOI: 10.18653/v1/N19-1112. URL: https://www.aclweb.org/anthology/N19-1112.

Martschat, Sebastian and Michael Strube (2015). “Latent Structures for Coreference Resolution”. In: *Transactions of the Association for Computational Linguistics* 3, pp. 405–418. DOI: 10.1162/tacl_a_00147. URL: https://www.aclweb.org/anthology/Q15-1029.

Ng, Vincent (2005). “Supervised Ranking for Pronoun Resolution: Some Recent Improvements”. In: *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*. AAAI'05. Pittsburgh, Pennsylvania: AAAI Press, pp. 1081–1086. ISBN: 157735236x.

Ng, Vincent and Claire Cardie (2002). “Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution”. In: *COLING 2002: The 19th International Conference on Computational Linguistics*. URL: https://www.aclweb.org/anthology/C02-1139.

Peters, Matthew E et al. (2018). “Deep contextualized word representations”. In: *Proceedings of NAACL-HLT*, pp. 2227–2237.

Pradhan, Sameer et al. (2012). “CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes”. In: *Joint Conference on EMNLP and CoNLL - Shared Task*. Jeju Island, Korea: Association for Computational Linguistics, pp. 1–40.

Shi, Xing, Inkit Padhi, and Kevin Knight (2016). “Does String-Based Neural MT Learn Source Syntax?” In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1526–1534. DOI: 10.18653/v1/D16-1159. URL: https://www.aclweb.org/anthology/D16-1159.

Tenney, Ian et al. (2019). “What do you learn from context? Probing for sentence structure in contextualized word representations”. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SJzSgnRcKX.

Winograd, T. (1972). *Understanding Natural Language*. Academic Press.

Wiseman, Sam, Alexander M. Rush, and Stuart M. Shieber (2016). “Learning Global Features for Coreference Resolution”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 994–1004. DOI: 10.18653/v1/N16-1114. URL: https://www.aclweb.org/anthology/N16-1114.

Wiseman, Sam, Alexander M. Rush, Stuart Shieber, et al. (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)*. Beijing, China: Association for Computational Linguistics, pp. 1416–1426. DOI: 10.3115/v1/P15-1137. URL: https://www.aclweb.org/anthology/P15-1137.

Woods, William A., Ronald M. Kaplan, and B. L. Nash-webber (1972). “The lunar sciences natural language information system: final report”. In:

Yang, Xiaofeng et al. (2003). “Coreference Resolution Using Competition Learning Approach”. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan: Association for Computational Linguistics, pp. 176–183.

Zhou, Liang, Miruna Ticrea, and Eduard Hovy (2004). “Multi-Document Biography Summarization”. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 434–441.