# Use Generalized Representations, But Do Not Forget Surface Features

**Nafise Sadat Moosavi** and **Michael Strube**
Heidelberg Institute for Theoretical Studies gGmbH
Schloss-Wolfsbrunnenweg 35
69118 Heidelberg, Germany
{nafise.moosavi|michael.strube}@h-its.org

## Abstract

Only a year ago, all state-of-the-art coreference resolvers were using an extensive amount of surface features. Recently, there was a paradigm shift towards using word embeddings and deep neural networks, where the use of surface features is very limited. In this paper, we show that a simple SVM model with surface features outperforms more complex neural models for detecting anaphoric mentions. Our analysis suggests that using generalized representations and surface features have different strength that should be both taken into account for improving coreference resolution.

## 1 Introduction

Coreference resolution is the task of finding different mentions that refer to the same entity in a given text. Anaphoricity detection is an important step for coreference resolution. An anaphoricity detection module discriminates mentions that are coreferent with one of the previous mentions. If a system recognizes mention $m$ as non-anaphoric, it does not need to make any coreferent links for the pairs in which $m$ is the anaphor.

The current state-of-the-art coreference resolvers (Wiseman et al., 2016; Clark and Manning, 2016a; Clark and Manning, 2016b), as well as their anaphoricity detection modules, use deep neural networks, word embeddings and a small set of features describing surface properties of mentions. While it is shown that this small set of features has significant impact on the overall performance (Clark and Manning, 2016a), their use is very limited in the state-of-the-art systems in comparison to the embedding features.

In this paper, we first introduce a new neural model for anaphoricity detection that considerably outperforms the anaphoricity detection of the state-of-the-art coreference resolver, i.e. deepcoref introduced by Clark and Manning (2016a). However, we show that a simple SVM model that is adapted from our coreferent mention detection approach (Moosavi and Strube, 2016), significantly outperforms the more complex neural models. We show that the SVM model also generalizes better than the neural model on a new domain other than the CoNLL dataset.

## 2 Discriminating Mentions for Coreference Resolution

The recognition of various categories of mentions can be beneficial for coreference resolution. The detection of the following categories is most common in the literature: (1) non-referential, (2) discourse-old, and (3) coreferent mentions. One can also discriminate other categories of mentions like mentions that are unlikely to be antecedents or discourse-new mentions (Uryupina, 2009). However, they are not common in comparison to the above categories.

### 2.1 Non-Referential Mentions

*Non-referential* mentions do not refer to an entity. These mentions only fill a syntactic position. For instance, "it" in "it is raining" is a non-referential mention. The approaches proposed by Evans (2001), Müller (2006), Bergsma et al. (2008), Bergsma and Yarowsky (2011) are examples of detecting non-referential cases of the pronoun *it*. Byron and Gegg-Harrison (2004) present a more general approach for detecting non-referential noun phrases.



*Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017), co-located with EACL 2017*, pages 1–7,
Valencia, Spain, April 4, 2017. 

### 2.2 Discourse-Old Mentions

Each mention can be assessed from the point of view of the discourse model (Prince, 1992). According to the discourse model, a mention may be new, old or inferable. Mentions which introduce a new entity into the discourse are *discourse-new* mentions. A discourse-new mention may be a singleton or it may be the first mention of a coreference chain. For instance, The first "Plato" in Example 2.1 is a *discourse-new* mention.

**Example 2.1.** *Plato* was a philosopher in Classical Greece. *This philosopher* is the founder of the Academy in Athens. *Plato* died at the age of 81.

A *discourse-old* mention refers to an entity that is already evoked in the discourse. Except for first mentions of coreference chains, other coreferent mentions are *discourse-old*. For instance, "this philosopher" and the second "Plato" in Example 2.1 are *discourse-old* mentions.

A mention is *inferable* if the hearer can infer the identity of the mention from another entity that has already been evoked in the discourse. "the windows" in Example 2.2 is an *inferable* mention.

**Example 2.2.** I walked into *the room*. *The windows* were all open.

The detection of discourse-old mentions is commonly referred to as *anaphoricity detection* (e.g. Zhou and Kong (2009), Ng (2009), Wiseman et al. (2015), Lassalle and Denis (2015), inter alia) while the task of anaphoric mention detection, based on its original definition, is of no use for coreference resolution. Mentions whose interpretations do not depend on previous mentions are called *non-anaphoric* mentions (van Deemter and Kibble, 2000). For example, both "Plato"s in Example 2.1 are non-anaphoric.

For consistency with the coreference literature, we refer to the task of discourse-old mention detection as anaphoricity detection.

Currently, all the state-of-the-art coreference resolvers learn anaphoricity detection jointly with coreference resolution (Wiseman et al., 2015; Wiseman et al., 2016; Clark and Manning, 2016a). The approaches proposed by Ng and Cardie (2002), Ng (2004), Ng (2009), Zhou and Kong (2009), Uryupina (2009) are examples of independent anaphoricity detection approaches.

### 2.3 Coreferent Mentions

Marneffe et al. (2015) discriminate mentions as coreferent vs. non-coreferent. Coreferent mentions are those mentions that appear in a coreference chain. A non-coreferent mention therefore can be a non-referential noun phrase or a referential noun phrase whose entity is only mentioned once (i.e. singleton). The proposed approaches of Recasens et al. (2013), Marneffe et al. (2015), and Moosavi and Strube (2016) discriminate mentions for coreference resolution this way.

## 3 Anaphoricity Detection Models

Anaphoricity detection is the most common approach for discriminating mentions for a coreference resolver. All of the state-of-the-art coreference resolvers use anaphoricity detection. In this paper, we compare three different anaphoricity detection approaches: two approaches using neural networks and word embeddings, and one using an SVM model and surface features. Clark and Manning (2016a) introduce the first neural model. Since Clark and Manning (2016a) train their anaphoricity model jointly with the coreference model, we refer to this model as the joint model. We introduce a new anaphoricity detection model as the second neural model using a Long-Short Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997). The third approach is adapted from our state-of-the-art coreferent mention detection (Moosavi and Strube, 2016).

### 3.1 Joint Model

As one of the neural models for anaphoricity detection, we consider the anaphoricity module of deep-coref[1], the state-of-the-art coreference resolution system introduced by Clark and Manning (2016a). This model has three layers for encoding different types of information regarding a mention. The first layer encodes the word embeddings of the head, first, last, two previous/following words, and the syntactic parent of the mention. The second layer encodes the averaged word embeddings of the five previous/following words, all words of the mention, sentence words, and document words. The third layer encodes the following features of a mention: type, length, position and whether it is embedded in another mention. The outputs of these three layers are combined into one vector and then get passed through a network with two hidden layers. This anaphoricity model is trained

[1] Available at `https://github.com/clarkkev/deep-coref`

jointly with the deep-coref coreference model.

### 3.2 LSTM Model

In this section we propose a new neural model for anaphoricity detection. Apart from the properties of the mention itself, we consider a limited number of surrounding words. We first generalize the context of a mention by removing the mention from the context and replacing it with a special placeholder. In our experiments, we consider the 10 previous and following words of a mention. We concatenate the mention tokens and the head token to the generalized word sequence. We separate the head and mention tokens in the concatenated sequence using two different placeholders.

The word embeddings of the above sequence are encoded using a bidirectional LSTM. LSTMs show convincing results on generating meaningful representations for various NLP tasks (e.g. Sutskever et al. (2014) and Vinyals et al. (2014)).

We also incorporate a set of surface features that contains (1) mention type (proper, nominal (definite, indefinite), pronouns (*he, I, it, she, they, we, you*)), (2) string match in the text, (3) string match in the previous context, (4) head match in the text, (5) head match in the previous context, (6) contains tokens of another mention, (7) contains tokens of a previous mention, (8) contained in another mention, (9) contained in a previous mention, and (10) embedded in another mention. These features are concatenated with the output of the bidirectional LSTM and get passed through one more layer that generates the output.

We also experiment with a more complex model including two different LSTMs for encoding mentions and their surrounding words. We consider longer sequences of previous words and an attention mechanism for processing the long sequence. However, the performance did not improve upon the LSTM model while it considerably increased the training time.

#### 3.2.1 Implementation Details

Hyperparameters are tuned on the CoNLL 2012 development set. We minimize the cross entropy loss using gradient-based optimization and the Adam update rule (Kingma and Ba, 2014). We use minibatches of size 50. A dropout (Hinton et al., 2012) with a rate of 0.3 is applied to the output of LSTM. We initialize the embeddings with the 300-dimensional Glove embeddings (Pennington et al., 2014). The size of LSTM's hidden layer is set to 128. The model is trained in only one epoch.

### 3.3 SVM Model

Our SVM model introduced in Moosavi and Strube (2016), achieves state-of-the-art results for coreferent mention detection. This model uses the following set of features: lemmas and POS tags of all words of a mention, lemmas and POS tags of the two previous/following words, mention string, mention length, mention type (proper, nominal, pronoun, list), string match in the text, and head match in the text. We use a similar SVM model for anaphoricity detection. In addition to the features we used for coreferent mention detection, we also add the following features for anaphoricity detection: string match in the previous context, head match in the previous context, mention words are contained in another mention, mention words are contained in a previous mention, mention contains words of another mention, mention contains words of a previous mention. Similar to Moosavi and Strube (2016), we use an anchored SVM (Goldberg and Elhadad, 2007) with a polynomial kernel of degree two and remove feature-values that occur less than 10 times. The use of an anchored SVM with pruning helps the model to generalize better on new domains (Goldberg and Elhadad, 2009).

## 4 Performance Evaluation

We evaluate the anaphoricity models on the CoNLL 2012 dataset. It is worth noting that all of the examined anaphoricity detectors in this section use the same mention detection module and results are reported using system detected mentions. The performance of the mention detection module is of crucial importance for anaphoricity detection. Therefore, it is important that the compared anaphoricity detectors use the same mention detection.

| | Non-Anaphoric | | | Anaphoric | | |
|---|---|---|---|---|---|---|
| | R | P | F1 | R | P | F1 |
| joint | - | - | - | 81.81 | 77.18 | 79.43 |
| LSTM | 90.71 | 92.64 | 91.66 | 85.00 | 81.48 | 83.20 |
| LSTM* | 90.51 | 87.31 | 88.88 | 72.64 | 78.64 | 75.52 |
| SVM | 92.42 | 92.61 | 92.51 | 84.66 | 84.30 | 84.48 |

Table 1: Results on the CoNLL 2012 test set.

The LSTM model that is described in Section 3.2 is denoted as *LSTM* in Table 1. In order to investigate the effect of the used surface

features, we also report the results of the LSTM model without using these features (*LSTM*$^*$).

The following observations can be drawn from the results of Table 1: (1) our LSTM model outperforms the joint model while using less features and being trained independently, (2) the results of the LSTM$^*$ model is considerably lower than those of LSTM, especially for recognizing anaphoric mentions, and (3) the simple SVM model outperforms the neural models in detecting both anaphoric and non-anaphoric mentions.

### 4.1 Generalization Evaluation

In order to investigate the generalization on new domains, we evaluate the LSTM and SVM models on the WikiCoref dataset (Ghaddar and Langlais, 2016). The WikiCoref dataset is annotated according to the same annotation guideline as that of CoNLL. Therefore, it is an appropriate dataset for performing out-of-domain evaluations when CoNLL is used for training. For the experiments of Table 2, all models are trained on the CoNLL 2012 training data and tested on the WikiCoref dataset.

The word dictionary that is used for the LSTM model is built based on the CoNLL 2012 training data. All words that are not included in this dictionary are treated as out of vocabulary words with randomly initialized word embeddings. We further improve the performance of LSTM on WikiCoref, by adding the words from the WikiCoref dataset into its dictionary. The LSTM model trained with this extended dictionary is denoted as *LSTM*$^\dagger$ in Table 2. *LSTM*$^\dagger$ results are still lower than those of the SVM model while SVM does not use any information from the test dataset. Pruning rare lexical features from the training data along the incorporation of part of speech tags, which are far more generalizable than lexical features, could explain the generalizability of the SVM model on the new domain.

| | Non-Anaphoric | | | Anaphoric | | |
|---|---|---|---|---|---|---|
| | R | P | F1 | R | P | F1 |
| LSTM | 95.53 | 89.88 | 92.62 | 69.50 | 84.58 | 76.31 |
| LSTM$^\dagger$ | 93.25 | 92.78 | 93.01 | 79.41 | 80.57 | 79.99 |
| SVM | 93.83 | 93.05 | 93.43 | 80.11 | 82.07 | 81.08 |

Table 2: Results on the WikiCoref dataset.

## 5 Analysis Based on Mention Types

We analyze the output of the LSTM and SVM models on the CoNLL 2012 test set to see how well they perform for different types of mentions. As can be seen from Table 3, there is not much difference between the performance of LSTM and SVM for recognizing anaphoric pronouns. SVM detects anaphoric proper names better while LSTM is better at recognizing anaphoric common nouns.

We also analyze the output of LSTM$^*$. As can be seen, the incorporation of surface features does not affect the detection of anaphoric pronouns very much while it mainly affects the detection of anaphoric proper names by about 24 percent.

In order to see whether the same pattern holds for coreference resolution, we compare the recall and precision errors of the best coreference system that only uses surface features, i.e. cort (Martschat and Strube, 2015) with singleton features (Moosavi and Strube, 2016) [2], and the state-of-the-art deep coreference resolver, i.e. deep-coref (Clark and Manning, 2016a). The comparison of the errors for the CoNLL 2012 test set is shown in Table 4. We use the error analysis tool of cort introduced by Martschat and Strube (2014) for the results of Table 4. As can be seen from Table 4, while deep-coref is significantly better than cort for resolving common nouns and specially pronouns, its result does not go far beyond that of cort when it comes to resolving proper names.

| | Anaphoric | | | | | |
|---|---|---|---|---|---|---|
| | R | P | F1 | R | P | F1 |
| | Proper names | | | Common nouns | | |
| LSTM | 79.49 | 82.31 | 80.88 | 62.96 | 65.04 | 63.99 |
| LSTM$^*$ | 47.60 | 70.09 | 56.69 | 46.30 | 57.75 | 51.40 |
| SVM | 83.80 | 85.71 | 84.74 | 52.46 | 71.98 | 60.69 |
| | Pronouns | | | Other | | |
| LSTM | 94.67 | 85.60 | 89.91 | 29.11 | 63.88 | 40.00 |
| LSTM$^*$ | 92.67 | 86.01 | 89.22 | 10.13 | 34.78 | 15.69 |
| SVM | 95.59 | 86.29 | 90.71 | 32.91 | 76.47 | 46.02 |

Table 3: Anaphoricity results for each mention type on the CoNLL 2012 test set.

## 6 Discussion

In this paper we analyze the effect of surface features for anaphoricity detection, which is a small but an important step for coreference resolution.

[2] Available at https://github.com/ns-moosavi/cort/tree/singleton_feature

| | Name | Noun | Pronoun |
|---|---|---|---|
| | #Recall Errors | | |
| deep-coref | 1110 | 1499 | 1537 |
| cort | 1145 | 1638 | 1655 |
| | #Precision Errors | | |
| deep-coref | 713 | 672 | 1162 |
| cort | 738 | 747 | 1736 |

Table 4: Coreference error analysis.

Our analysis shows that surface features, as it was known, are important. Based on our results, the effects of incorporating surface properties and generalized representations are different for different types of mentions. These results suggest that apart from a unified model, we should consider different models or at least different features for processing different types of mentions and do not put all the burden on a single model to learn the differences. The works by Lassalle and Denis (2013) and Denis and Baldridge (2008) are examples of models in which distinct models have been used for various types of mentions. Besides, our analysis shows the importance of surface features for proper names. Word embeddings are very useful for capturing semantic relatedness. A coreference resolver that uses word embeddings has a great advantage in better resolution of common nouns and pronouns. However, the use of surface features in current state-of-the-art coreference resolvers is very limited. Before going towards using more sophisticated knowledge sources, there are still easy victories that can be achieved by incorporating more generalizable surface properties, especially for proper names.

## Acknowledgments

The authors would like to thank Kevin Clark for his help with the deep-coref software and Mark-Christoph Müller for his helpful comments. We would also like to thank the four anonymous reviewers for their detailed comments on an earlier draft of the paper. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a Heidelberg Institute for Theoretical Studies PhD. scholarship.

## References

Shane Bergsma and David Yarowsky. 2011. NADA: A robust system for non-referential pronoun detection. In *Anaphora Processing and Applications. Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium,* Lisbon, Portugal, 6-7 October 2011, pages 12–23. Springer, Heidelberg.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Distributional identification of non-referential pronouns. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* Columbus, Ohio, 15–20 June 2008, pages 10–18.

Donna Byron and Whitney Gegg-Harrison. 2004. Eliminating non-referring noun phrases from coreference resolution. In *Proceedings the Discourse Anaphora and Reference Resolution Conference*, pages 21–26.

Kevin Clark and Christopher D. Manning. 2016a. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany, August. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016b. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas, November. Association for Computational Linguistics.

Pascal Denis and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 660–669.

Richard Evans. 2001. Applying machine learning toward an automatic classification of it. *Literary and Linguistic Computing*, 16(1):45–57.

Abbas Ghaddar and Philippe Langlais. 2016. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, 05/2016.

Yoav Goldberg and Michael Elhadad. 2007. SVM model tampering and anchored learning: a case study in Hebrew NP chunking. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 224–231.

Yoav Goldberg and Michael Elhadad. 2009. On the role of lexical features in sequence labeling. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* Singapore, 6–7 August 2009, pages 1142–1151. Association for Computational Linguistics.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Emmanuel Lassalle and Pascal Denis. 2013. Improving pairwise coreference models through feature space hierarchy learning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Sofia, Bulgaria, 4–9 August 2013, pages 497–506.

Emmanuel Lassalle and Pascal Denis. 2015. Joint anaphoricity detection and coreference resolution with constrained latent structures. In *Proceedings of the 29th Conference on the Advancement of Artificial Intelligence,* Austin, Texas, 25–30 January 2015, pages 2274–2280.

Marie-Catherine de Marneffe, Marta Recasens, and Christopher Potts. 2015. Modeling the lifespan of discourse entities with application to coreference resolution. *Journal of Artificial Intelligent Research*, 52:445–475.

Sebastian Martschat and Michael Strube. 2014. Recall error analysis for coreference resolution. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* Doha, Qatar, 25–29 October 2014, pages 2070–2081.

Sebastian Martschat and Michael Strube. 2015. Latent structures for coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:405–418.

Nafise Sadat Moosavi and Michael Strube. 2016. Search space pruning: A simple solution for better coreference resolvers. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* San Diego, Cal., 12–17 June 2016, pages 1005–1011.

Christoph Müller. 2006. Automatic detection of nonreferential *it* in spoken multi-party dialog. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics,* Trento, Italy, 3–7 April 2006. 49-56.

Vincent Ng and Claire Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics,* Taipei, Taiwan, 24 August – 1 September 2002, pages 730–736.

Vincent Ng. 2004. Learning noun phrase anaphoricity to improve coreference resolution. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July 2004, pages 151–158.

Vincent Ng. 2009. Graph-cut-based anaphoricity determination for coreference resolution. In *Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics,* Boulder, Col., 31 May – 5 June 2009, pages 575–583.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* Doha, Qatar, 25–29 October 2014, pages 1532–1543.

Ellen F. Prince. 1992. The ZPG letter: Subjects, definiteness, and information-status. In W.C. Mann and S.A. Thompson, editors, *Discourse Description. Diverse Linguistic Analyses of a Fund-Raising Text*, pages 295–325. John Benjamins, Amsterdam.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Atlanta, Georgia, 9–14 June 2013, pages 627–633.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, pages 3104–3112.

Olga Uryupina. 2009. Detecting Anaphoricity and Antecedenthood for Coreference Resolution. *Procesamiento del Lenguaje Natural*, 42.

Kees van Deemter and Rodger Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637.

Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2014. Grammar as a foreign language. *arXiv preprint arXiv:1412.7449*.

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 (Volume 1: Long Papers),* Beijing, China, 26–31 July 2015, pages 1416–1426.

Sam Wiseman, Alexander M. Rush, and Stuart 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, Cal., 12–17 June 2016. To appear.

Guodong Zhou and Fang Kong. 2009. Global learning of noun phrase anaphoricity in coreference resolution via label propagation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* Singapore, 6–7 August 2009, pages 978–986.