Search Space Pruning: A Simple Solution for Better Coreference Resolvers

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

There is a significant gap between the performance of a coreference resolution system on gold mentions and on system mentions. This gap is due to the large and unbalanced search space in coreference resolution when using system mentions. In this paper we show that search space pruning is a simple but efficient way of improving coreference resolvers. By incorporating our pruning method in one of the state-of-the-art coreference resolution systems, we achieve the best reported overall score on the CoNLL 2012 English test set. A version of our pruning method is available with the Cort coreference resolution source code.

1 Introduction

Coreference resolution is the task of clustering referring expressions in a text so that each resulting cluster represents an entity. It is a very challenging task in natural language processing and it is still far from being solved, i.e. the best reported overall CoNLL score on the CoNLL 2012 English test set is 63.39 (Wiseman et al., 2015).

Text spans referring to an entity are called mentions. Mentions are the primary objects in a coreference resolution system. As with most previous work on coreference resolution, we only consider mentions that are noun phrases. However, not all of the noun phrases are mentions. A noun phrase may not refer to any entity at all. The pronoun *it* in the sentence *it is raining* is an example of a non-referential noun phrase. Noun phrases which do refer to an entity (mentions) can be further divided into two categories: mentions referring to entities which only appear once in the discourse (i.e. singletons), and mentions realizing entities that have been referred to more than once in the text (i.e. coreferent mentions). Henceforth, we refer to both singletons and non-referential phrases as non-coreferent mentions. A large number of mentions that appear in a text are non-coreferent. For instance, more than 80% of mentions are singletons in the OntoNotes English development set (Marneffe et al., 2015).

The latent ranking model is the best performing model for coreference resolution to date (Wiseman et al., 2015; Martschat and Strube, 2015). If we use gold mentions, the latent ranking model of Martschat and Strube (2015) achieves an overall score of 80% on the CoNLL 2012 English test set. This result shows that once we have the ideal pruned search space, the ranking model with the current set of features is reasonably capable of finding corresponding entities of mentions. The substantial gap (17%) between the results of the gold mentions and system mentions implies that search space pruning is a promising direction for further improvements in coreference resolution.

Marneffe et al. (2015) examine different search space pruning methods that exist for coreference resolution. Among those, anaphoricity detection is the most popular method (e.g. Ng and Cardie (2002), Denis and Baldridge (2007), Ng (2009), Zhou and Kong (2009), Durrett and Klein (2013), Martschat and Strube (2015), Wiseman et al. (2015), Peng et al. (2015), and Lassalle and Denis (2015)), while singleton detection is a more recent method (Recasens et al., 2013; Ma et al., 2014; Marneffe et al., 2015).

Proceedings of NAACL-HLT 2016, pages 1005–1011,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

Anaphoricity detection examines whether a phrase is anaphoric. Singleton detection examines whether a phrase belongs to a coreference chain regardless of being anaphor or antecedent. Therefore, anaphoricity detection only prunes the search space of anaphors while singleton detection prunes the search space of both anaphors and antecedents.

Except for Clark and Manning (2015), all of the state-of-the-art coreference resolvers explicitly model anaphoricity detection (Martschat and Strube, 2015; Wiseman et al., 2015; Peng et al., 2015). Therefore, modeling search space pruning as singleton detection can provide additional information for the state-of-the-art coreference resolution systems.

In this paper we propose a simple but efficient singleton detection model. We first perform intrinsic evaluations and show that our simple model significantly improves the state-of-the-art results in singleton detection by a large margin. We then evaluate our singleton model extrinsically on coreference resolution showing that search space pruning improves different coreference resolution models.

2 Simple but Efficient Singleton Detection

In this section we show that pruning the coreference resolution search space is not a very difficult task. By using a simple set of features and a standard classifier, we achieve new state-of-the-art results for classifying coreferent and non-coreferent mentions.

Unlike Marneffe et al. (2015) who use both surface (i.e. part-of-speech and n-gram based) features and a large number (123) of carefully designed linguistic features, we select a simple and small set of shallow features:

1. lemmas of all words included in the mention;
2. lemmas of the two previous/next words before/after the mention;
3. part-of-speech tags of all words of the mention;
4. part-of-speech tags of the two previous/next words before/after the mention;
5. complete mention string;
6. length of the mention in words;
7. mention type (proper, nominal, pronominal);
8. whether the whole string of the mention appears again in the document;
9. whether the head of the mention appears again in the document.

We use an anchored SVM (Goldberg and Elhadad, 2007) with a polynomial kernel of degree two for classification. When only few features are available, anchored SVMs generalize much better than soft-margin-SVMs (Goldberg and Elhadad, 2009). In our experiments, we use a count threshold for discarding vary rare lexical features that occur fewer than 10 times.

Similar to Marneffe et al. (2015), we use three different configurations for evaluation. The *Surface* configuration only uses the shallow features. The *Combined* configuration uses the surface features plus the linguistic features introduced by Marneffe et al. (2015). The linguistic features of Marneffe et al. (2015) also include some pairwise combinations of the single features. Since our SVM with a polynomial kernel of degree two implicitly models feature pairs, we only include the single features in our *Combined* configuration. When removing mentions that are classified as non-coreferent during preprocessing, precision matters more than recall in order not to over prune coreferent mentions. To achieve higher precision, the *Confident* configuration uses high confidence predictions of SVM (i.e. classifying a mention as non-coreferent if the SVM output is less or equal to -1, and as coreferent if the output is greater or equal to +1). We use the same set of shallow features as *Surface* for *Confident*. However, Marneffe et al. (2015) use their combined feature set for *Confident*.

2.1 Results

Table 2 shows the results of our singleton detection model in comparison to that of Marneffe et al. (2015). We train our model on the CoNLL 2012 English training set and evaluate it on the development set using recall, precision, F1 measure and accuracy for both coreferent and non-coreferent mentions. Unlike Marneffe et al. (2015) that also use some gold annotations for their features, we extract all of our surface features from `'auto_conll'` files. Therefore, only predicted annotations are used.

The incorporation of linguistic features in Marneffe et al. (2015) improves the classification of both coreferent and non-coreferent mentions by about 1

			Non-Coreferent			Coreferent			
		#Features	R	P	F1	R	P	F1	Accuracy
Marneffe et al.	Surface	73,393	80.2	79.9	80.0	75.3	75.6	75.4	78.0
	Confident	73,516	56.0	89.8	69.0	48.2	90.7	62.9	52.2
	Combined	73,516	81.1	80.8	80.9	76.4	76.6	76.5	79.0
This work	Surface	8,331	89.37	87.08	88.21	80.32	83.59	81.92	85.73
	Confident	8,331	65.08	94.44	77.06	55.14	93.55	69.38	61.08
	Combined	8,446	89.48	87.16	88.30	80.45	83.76	82.07	85.85

Table 1: Results on the the CoNLL 2012 English development set.

percent in comparison to the *Surface* results. However, in our case, the linguistic features only improve the results by about 0.1 percent.

As the results show, by only using shallow features, we achieve a new state-of-the-art performance for singleton detection that improves the results of Marneffe et al. (2015) by a large margin for classifying both coreferent and non-coreferent mentions.

2.2 Error Analysis

For a singleton detector, precision errors (classifying a coreferent mention as non-coreferent) are more harmful than recall errors. If a coreferent mention is classified as non-coreferent, the recall of the coreference resolver that uses the singleton detector will decrease. On the other hand, recall errors only affect the singleton detector itself and not coreference resolvers.

The precision error ratios of our *Surface* and *Confident* systems for proper name (NAM), nominal (NOM) and pronominal (PRO) mentions are listed in Table 2. For each mention type, Table 2 also shows the precision error ratio by mention type related to the mentions that are first mentions of their corresponding entities. For example, in the *Confident* system 73.45% of the nominal mentions that are incorrectly classified as non-coreferent are first mentions of their corresponding entities. As can be seen, many of the precision errors in both *Surface* and *Confident* systems are errors in which the first mention of an entity is detected as non-coreferent. Detecting whether a mention will be referred to later, is indeed very hard and requires more context information. Features (8) and (9) from our feature set are designed to address the correct detection of the first mentions of entities to a limited degree. These features only address first mentions of entities that are referred to by later mentions with head or complete string match. More features considering properties of other mentions, rather than the examined mention itself, are required in order to improve the correct detection of the first mentions of entities.

		NAM	NOM	PRO
Surface	Error rate	23.17	70.61	6.22
	First mentions	57.68	65.54	20.19
Confident	Error rate	23.52	74.70	1.78
	First mentions	62.63	73.45	33.33

Table 2: Precision error ratio.

	NAM	NOM	PRO
Surface	30.48	34.07	35.45
Confident	23.65	58.25	13.50

Table 3: Recall error ratio.

Table 3 shows the ratio of recall errors for each mention type. For our *Surface* system, this ratio is more or less the same for different mention types. However, *Confident*'s main source of recall errors is the detection of non-coreferent nominal mentions.

2.3 Discussion

Our results significantly outperform the results of Marneffe et al. (2015) who use both surface features and a set of hand-engineered features targeting different linguistic phenomena related to the task. Our findings are mirrored by Durrett and Klein (2013)'s work on the coreference resolution task. Durrett and Klein (2013) show that a coreference resolution system that uses surface features can outperform those using hand-engineered linguistic features.

Linguistic features like syntactic nearness (on which Hobbs' algorithm (Hobbs, 1978) is based), morpho-syntactic and semantic agreement (e.g. number, gender and semantic class agreements), re-

		MUC			B^3			CEAF$_e$			Avg.
		R	P	F1	R	P	F1	R	P	F1	F1
Stanford	Baseline	64.58	63.65	64.11	49.53	55.21	52.22	53.06	44.82	48.59	54.97
	+Stanford Singleton	64.26	65.19	64.72	49.09	56.84	52.68	52.54	46.55	49.37	55.59
	+Preprocess Pruning	64.27	69.01	66.56	48.65	60.32	53.86	48.71	51.48	50.06	56.83
Cort	Pairwise	68.46	71.01	69.71	54.02	59.47	56.61	51.88	52.17	52.02	59.45
	+Preprocess Pruning	68.19	73.38	70.69	53.62	62.02	57.52	51.42	55.07	53.18	60.46
	Latent Ranking	68.55	77.22	72.63	54.64	66.78	60.11	52.85	60.3	56.33	63.02
	+Pruning Feature	68.81	78.37	73.28	55.46	66.9	60.65	52.07	62.23	56.7	63.54
Wiseman et al. (2015)		69.31	76.23	72.60	55.83	66.07	60.52	54.88	59.41	57.05	63.39

Table 4: Results on the English test set. All the improvements made by our singleton detection models are statistically significant.

cency, focus (Grosz and Sidner, 1986), and centering (Brennan et al., 1987) are examples of useful linguistic features for coreference resolution which have the additional benefit of being applicable to different languages. For example, Hobbs' algorithm and agreement features are being used successfully in the Stanford system (Lee et al., 2013). However, apart from features like these, a large number of linguistically motivated features have been proposed which either do not have a significant impact or are only applicable to a specific language or domain. Therefore, designing general linguistic features which provide information that is not captured by surface features deserve more attention in order to gain higher recall and better generalization.

We combine a simple set of surface features with a standard machine learning model that can handle a large number of surface features. This leads to a new state-of-the-art singleton detection with high precision that can easily be incorporated in a coreference resolution system for pruning non-coreferent mentions.

3 Pruning = Better Coreference Resolvers

In this section, we investigate the effect of search space pruning on coreference resolution. We choose the Stanford rule-based system (Lee et al., 2013) and the Cort[1] system (Martschat and Strube, 2015) as our baselines for coreference resolution. Wiseman et al. (2015) is the best performing coreference resolution system to date. However, we choose Cort as our learning-based baseline because Cort is a framework that allows evaluations on various coreference resolution models, i.e. ranking, antecedent trees, and pairwise. The pairwise model is the most commonly used model in coreference resolution, and latent ranking is the best performing model for coreference resolution to date (Wiseman et al., 2015; Martschat and Strube, 2015).

3.1 Results

Table 4 shows the results of integrating singleton detection into different coreference resolution approaches. We evaluate the systems on the CoNLL 2012 English test set using the MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_e$ (Luo, 2005) measures as provided by the CoNLL coreference scorer version 8.01 (Pradhan et al., 2014). According to the approximate randomization test (Noreen, 1989), all of the improvements made by our singleton detection module are statistically significant ($p < 0.05$).

Baseline shows the result of the Stanford system without using singleton detection. *+Stanford Singleton* is the result of the Stanford system including its singleton detection module (Recasens et al., 2013). *+Preprocess Pruning* is the result when our *Confident* model from Section 2 is used.

The singleton detection modules of Recasens et al. (2013) and Marneffe et al. (2015) are incorporated in the Stanford system in a heuristic way: if both anaphor and antecedent are classified as singleton, and none of them is a named entity, then those mentions will be disregarded. However, since our *Confident* model does have a high precision, we use it for removing all non-coreferent mentions in a preprocessing step. As shown in Table 4, our singleton detection improves the overall score of the *Baseline*

[1] http://github.com/smartschat/cort

system by about 2 percent on the test set.

Cort uses a perceptron for learning. Therefore, we use a perceptron in Cort while an anchored SVM would have performed slightly better. We also include all the additional features that are used in Cort for our Cort singleton detection model. SVM accuracy with surface features on the development set is about 0.1 percent better than that of the perceptron with Cort's additional features.

For the pairwise model, singleton detection is performed in a preprocessing step. The singleton detection module improves the overall performance of the pairwise model by about 1 percent on the test set.

The Cort latent model already performs search space pruning in the form of anaphoricity detection. Additional pruning of potential anaphors in the preprocessing step by the singleton model hurts the recall of the latent model. Therefore, we add the output of the singleton model as a new feature for both anaphor and antecedent. For obtaining these features for training, we split the training data into two halves and train a singleton perceptron separately on each half. The values of the singleton feature for the first half are computed based on the model that is trained on the second half, and vice versa. This way, the accuracy of singleton features on both training and testing is similar. If we would train the singleton model on the whole training data, we would overfit the model seriously. The values of the singleton feature would be very accurate on the training data, and the learner would overestimate the importance of this feature.

The new feature improves the overall performance of the latent ranking model by about 0.5 percent on the test set. This result is the best reported overall score for coreference resolution on the CoNLL 2012 English test set to date.

The singleton feature support is added to the Cort source code. It is available at http://github.com/smartschat/cort.

3.2 Discussion

Recent improvements in coreference resolution have been made by exploring more complex learning and inference strategies, a larger number of features, and joint processing. There are also technically viable solutions for improving the performance of a coreference resolver which do not work in practice. For instance, since coreference resolution is a set partitioning problem, entity-based models seem to be more suitable for coreference resolution than mention-pair models. However, entity-based models do not necessarily perform better than mention-pair models (e.g. Ng (2010) and Moosavi and Strube (2014)). The same is true for incorporating more semantic-level information in a coreference resolution system (e.g. Durrett and Klein (2013)).

In this paper, we show that coreference resolution can also simply be improved by performing search space pruning. The significant gap between the performance of the latent ranking model on gold mentions and on system mentions indicates that there is still room for further improvements in search space pruning.

4 Conclusions

We achieve new state-of-the-art results for singleton detection by only using shallow features and simple classifiers. We also show that search space pruning significantly improves different coreference resolution models. The substantial gap between the performance on gold mentions and on system mentions indicates that there is still plenty of room for further improvements in singleton detection. Therefore, search space pruning is a promising direction for further improvements in coreference resolution. The proposed singleton detector as a feature for coreference resolvers is implemented for the Cort coreference resolver. It is available with the Cort source code.

Acknowledgments

The authors would like to thank Sebastian Martschat for his help to use the proposed singleton detector in the Cort coreference resolver. We would also like to thank Mark-Christoph Müller, Benjamin Heinzerling, Mohsen Mesgar, Daraksha Parveen and Alex Judea for their helpful comments. 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

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the 1st International Conference on Language Resources and Evaluation,* Granada, Spain, 28–30 May 1998, pages 563–566.

Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics,* Stanford, Cal., 6–9 July 1987, pages 155–162.

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Beijing, China, 26–31 July 2015, pages 1405–1415.

Pascal Denis and Jason Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics,* Rochester, N.Y., 22–27 April 2007, pages 236–243.

Greg Durrett 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, Wash., 18–21 October 2013, pages 1971–1982.

Yoav Goldberg and Michael Elhadad. 2007. SVM model tampering and anchored learning: A case study in Hebrew NP chunking. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* Prague, Czech Republic, 23–30 June 2007, 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.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Jerry R. Hobbs. 1978. Resolving pronominal references. *Lingua*, 44:311–338.

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.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing,* Vancouver, B.C., Canada, 6–8 October 2005, pages 25–32.

Chao Ma, Janardhan Rao Doppa, J. Walker Orr, Prashanth Mannem, Xiaoli Fern, Tom Dietterich, and Prasad Tadepalli. 2014. Prune-and-score: Learning for greedy coreference resolution. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* Doha, Qatar, 25–29 October 2014, pages 2115–2126.

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

Nafise Sadat Moosavi and Michael Strube. 2014. Unsupervised coreference resolution by utilizing the most informative relations. In *Proceedings of the 25th International Conference on Computational Linguistics,* Dublin, Ireland, 23–29 August 2014, pages 644–655.

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.

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.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010, pages 1396–1411.

Eric W. Noreen. 1989. *Computer Intensive Methods for Hypothesis Testing: An Introduction.* Wiley, New York, N.Y.

Haoruo Peng, Kai-Wei Chang, and Dan Roth. 2015. A joint framework for coreference resolution and mention head detection. In *Proceedings of the 19th Conference on Computational Natural Language Learning,* Beijing, China, 30–31 July 2015, pages 12–21.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* Baltimore, Md., 22–27 June 2014, pages 30–35.

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.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pages 45–52, San Mateo, Cal. Morgan Kaufmann.

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.

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.