Improving on Recall Errors for Coreference Resolution

Bishan Yang
Department of Computer Science
Cornell University
bishan@cs.cornell.edu

Claire Cardie
Department of Computer Science
Cornell University
cardie@cs.cornell.edu

Abstract

We present a pairwise learning method that aims to improve on recall errors for noun phrase coreference resolution. We first show the weakness of a widely-used state-of-the-art coreference resolution system – Stanford's rule-based sieve system on grouping proper names and common nouns. We then design a feature-based classifier and an embedding-based ranker that are tailored to model mention reference relations for proper names and common nouns. Experiments show that a combination of these two learning models allows for better recall performance while maintaining precision, and it provides promising improvement over the Stanford's sieve system on the CoNLL 2011 data set.

1 Introduction

Noun phrase coreference resolution, which deals with identifying and grouping noun phrases that refer to the same discourse entity, is an essential component for systems that extract and integrate information from natural language text. It is a challenging task due to its complex nature: accurate coreference decisions require exploiting syntactic, semantic and discourse cues as well as world knowledge. Prior work on noun phrase coreference resolution mainly falls into two categories: rule-based methods (e.g. [2, 7, 9]) that rely on high-precision deterministic cues such as string matching and grammatical rules, and learning-based methods (e.g. [14, 17, 6]) that encode rich syntactic and semantic features using supervised learning models. A recent study on coreference resolution error analysis [11] showed that state-of-the-art coreference systems, both rule-based and learning-based, suffer from recall errors when resvolving proper names and common nouns since the systems mainly rely on overlapping patterns of surface forms. For example,

> **Mickey Mouse's new home**, settling on Chinese land for the first time, has captured worldwide attention. There's only one month left before the opening of **Hong Kong Disneyland** on September 12.

Because there is no string overlap, most state-of-the-art systems will miss the coreference link between *Mickey Mouse's new home* and *Hong Kong Disneyland*. However, the information that *Mickey Mouse* is related to *Disneyland* can help to determine that the phrases are coreferent.

In this work, we propose learning-based techniques that aim to tackle such recall errors. We start by analyzing weaknesses in a widely-used state-of-the-art noun phrase coreference resolution system — Stanford's rule-based "Sieve" system — focusing on errors that involve proper names and common nouns. [1] To improve on these errors, we develop a learning-based model that aim to estimate the co-referential relation between two noun phrases. The model is a combination of a feature-based

[1] We do not consider pronoun resolution in this work. We believe that it requires a specialized learning model as pronouns appear to exhibit different reference patterns from other types of nominal mentions.

classifier with rich semantic and syntactic features and an embedding-based ranker that ranks noun phrase pairs according to their coreference compatibility. We show that each model provides complementary strengths and that combining the two can result in recall improvement while maintaining precision. Finally, we integrate our model into the Sieve system and demonstrate promising overall performance improvements on the CoNLL 2011 coreference resolution data set.

2 Background and Motivation

It is widely known that string-matching rules typically contribute substantially to the performance of modern coreference resolution systems [18, 19, 13]. To gain insight into which factors beyond surface-level matches play an important role, we examine (binary) coreference decisions for noun phrase pairs where exact string matching and head matching do not apply.

In particular, we employ the Sieve system [9], a rule-based model that consists of multiple precision-oriented "sieves" that merge coreferent mentions into the same equivalence class in multiple passes, and evaluate it on the development set from the CoNLL 2011 shared task [16]. We use the predicted annotations for POS, parses, NER and speaker tags provided with the CoNLL data. We consider two evaluation settings, one that takes gold-standard *mentions* as input — these comprise all and only those noun phrases involved in coreference relationships in the document ("Singleton" noun phrases are removed) — and one in which a system predicts which noun phrases should be considered mentions of entities referred to more than once in the discourse.

			Number of recall errors w.r.t. mention type		
	Precision	Recall	Prop-Prop	Nom-Nom	Prop-Nom/Nom-Prop
STANFORDSIEVE (Gold)	89.4	6.1	963	1730	2760
STANFORDSIEVE (Predicted)	90.8	7.0	707	1023	1942

Table 1: Evaluation of the Sieve system on mention pair classification using mention pairs with no exact string or head match. Precision and Recall are measured with respect to the positive class. *Prop* indicates proper names and *Nom* indicates common nouns.

In both settings, we select mention pairs from each document by filtering from all possible ordered pairs of noun phrases those that (a) involve a pronoun or (b) could be resolved using string match or head match. We label each remaining mention pair as positive if its mentions refer to the same discourse entity and as negative otherwise. This results in $5,808$ $(3,936)$ positive mention pairs and $251,527$ $(170,048)$ negative mention pairs in the gold-standard (predicted) mention setting.

We run the Sieve system under each mention-pair setting and report precision and recall for the positive class in Table 1. (Performance for the negative class is not very informative as the majority of mention pairs are not coreferent.) It shows that the Sieve system provides high precision but extremely low recall on the selected mention pairs. This implies that the majority of the correct coreference links are not predicted. By examining the recall errors on gold mentions, we found that most errors are due to the lack of semantic and commonsense knowledge, e.g. missing the link between "the park" and "Disney", between "the interruption" and "a pause", and between "my kid" and "my daughter". The errors also exhibit rich forms of semantic compatibility, e.g. entity type agreement, synonymy, hypernymy and world knowledge. A few hand-coded rules can hardly capture such variety of semantic information and thus a learning-based approach is needed.

3 Pairwise Learning

We explore two pairwise learning models: a feature-based model and an embedding-based model, and also their combination for mention reference prediction. Training data consists of positive mention pairs (coreferent) and negative mention pairs (not coreferent) constructed from each training document. Note that we only consider mention pairs with no exact string match or head match. We expect a model that is specific to mention pairs of this type to perform better than a general model trained on all possible mention pairs.

3.1 Feature-based Model

The feature-based model is a binary classifier that learns to distinguish coreferent mention pairs from non-coreferent mention pairs using a rich set of features. We consider features introduced in [3] which capture syntax (e.g. mention types (i.e. proper or nominal) and grammatical roles) as well as semantics (e.g. Wordnet categories and entity types); the head noun pair (to capture some surface information); and the shortest dependency path (both lexicalized and unlexicalized) between the pair of head nouns (to capture longer-distance syntactic relations).

Training employs the standard logistic regression objective that maximizes the log likelihood of the positive mention pairs. We trained the model using the training set of CoNLL2011 [2] and evaluated it on the development set as in Section 2. Table 2 shows the results.

We can see that the feature-based model provides a large improvement on recall but precision sacrifices compared to the Sieve baseline. This indicates that the model succeeds in retrieving coreference links that are missed by deterministic rules, however it introduces many spurious links that harm the precision. For example, it mistakenly link *the Iranian Government* with *Spain* even *Iran* and *Spain* are two completely different countries.

			Number of recall errors w.r.t. mention type		
	Precision	Recall	Prop-Prop	Nom-Nom	Prop-Nom/Nom-Prop
FEATURELR (Gold)	83.9	50.8	385	912	1561
FEATURELR (Predicted)	82.1	52.5	275	548	1052

Table 2: Evaluation of the feature-based model (FEATURELR) on mention pair classification

3.2 Embedding-based Model

The feature-based model relies on highly-sparse features and are prone to overfitting. We introduce an embedding-based model that represents each mention as a low-dimensional vector, and learns a scoring function that rank a pair of mentions according to their compatibility in a referential relation.

We construct the vector representations of mentions using $d = 300$-dimensional word vectors pre-trained on large-scale Wikipedia text released by the *word2vec* tool [12]. Specifically, for each mention m, we construct a vector $v_m \in R^{2d}$ by concatenating the average of the constituting word vectors of the mention and the head noun vector. The compatibility score for a mention pair (m_1, m_2) is defined via the weighted Eucleadian Distance of the vector representations:

$$g(m_1, m_2) = -(v_{m_1} - v_{m_2})^T W (v_{m_1} - v_{m_2})$$

where $W \in R^{2d \times 2d}$ is a parameter matrix.

The parameter matrix can be learned via a ranking objective [3] that encourages referential mentions to have higher compatibility score than any non-referential mentions. More formally, for each mention m, denote the set of its referential mentions as $A = \{a_i\}$ and the set of its non-referential mentions as $A' = \{a'_i\}$, we minimize the margin ranking loss [8]

$$L_m = \sum_{a \in A} \sum_{a' \in A'} \max\{g(m, a') - g(m, a) + 1, 0\}$$

Training is performed by mini-batch stochastic gradient descent with Adagrad [5]. In our experiments, we constrain W to be a diagonal matrix, since it provides similar performance to learning a full matrix and learning is much faster. At prediction time, we find a threshold T using the development set such that for each mention pair (m_1, m_2), if $g(m_1, m_2) \geq T$ then they are coreferential otherwise they are not.

Note that our ranking objective is very different from the antecedent ranking objective used in the coreference literature [4]: our model considers the coreferential relation between mentions to be

[2] We filter negative pairs that appear in a window of more than two sentences to make the training set more balanced.

[3] We also experimented with the classification objective with cross-entropy error but found that the ranking objective provides better empirical performance for the embedding-based model on mention pair classification.

symmetric, ignoring the mention ordering; also, it allows more than one candidates to corefer with the considered mention while the traditional ranking model selects only one best antecedent for each mention.

	Precision	Recall	Number of recall errors w.r.t. mention type		
			Prop-Prop	Nom-Nom	Prop-Nom/Nom-Prop
EMBEDDING	34.8	55.8	400	447	1663
COMBINED	88.0	32.0	643	1072	2133

Table 3: Evaluation of the embedding-based model (EMBEDDING) and the combined model (COMBINED) on mention pair classification

We show results of the embedding-based model on mention pair classification in Table 3 under the gold mention setting (the results under the predicted mention setting demonstrate a similar trend). We can see that the embedding model reduces the recall errors but provides poor precision. The error reduction comes from capturing the semantics of mentions via low-dimensional word vectors. The poor precision indicates that mention compatibility alone is not sufficient for predicting mention references. To take advantage of the strengths of both feature-based and embedding-based models, we combine their predictions by linking two mentions only if both models predict them to be coreferent. This can correct many precision errors made by the feature-based model. For example, it corrects the error of linking *New Democracy* to *the republic* made by the feature-based model. Overall the combined model (COMBINED) provides recall improvement while maintaining precision.

4 Coreference Results

Now we evaluate the effect of the pairwise model on the end results of coreference resolution. We integrated the model into the Sieve system by adding an additional cluster-merging sieve that merges coreferent mentions according to the model output. We conducted experiments on the CoNLL2011 development set and test set and report the results using MUC [20], B^3 [1], CEAF (CEAF$_e$ and CEAF$_m$) [10] as well as the average F1 computed using the latest version of the official CoNLL scorer [15]. We consider STANFORDSIEVE as our baseline and compare it to its two extensions based on the feature-based model and the combined model.

We show the results in table 4. We found that in general, the learning-based models significantly improve recall in MUC and B^3 over the rule-based baseline. This further confirms that pairwise learning allows for more coverage of the coreferential mentions which are missed by the deterministic rules. Consistent with our pairwise classification results, the combined model improves recall while exhibiting a small drop on precision. The feature-based model provides better recall with a large drop on precision. Overall, the combined model provides the best F1 scores in various metrics, outperforming the Sieve baseline under all evaluation settings.

	MUC			B^3			CEAF$_e$	CEAF$_m$	Avg
	P	R	F1	P	R	F1	F1	F1	F1
	CoNLL 2011 Development Set (Gold Mentions)								
STANFORDSIEVE	**89.6**	69.6	78.4	**86.4**	58.8	70.0	68.7	71.3	72.1
+FEATURELR	87.4	**77.8**	**82.3**	75.5	**70.1**	72.7	69.7	70.6	73.8
+COMBINED	88.8	75.1	81.4	81.9	66.2	**73.2**	**70.8**	**73.1**	**74.6**
	CoNLL 2011 Test Set (Gold Mentions)								
STANFORDSIEVE	**89.1**	70.2	78.6	**83.7**	57.3	68.0	66.7	68.6	70.5
+FEATURELR	87.8	**77.5**	**82.3**	76.2	**67.7**	71.7	68.6	70.0	73.2
+COMBINED	88.7	74.8	81.2	81.7	65.8	**72.9**	**70.5**	**72.9**	**74.4**
	CoNLL 2011 Test Set (Predicted Mentions)								
STANFORDSIEVE	60.2	58.5	59.4	**51.8**	45.3	48.3	46.1	52.6	51.6
+FEATURELR	61.2	**63.4**	**62.3**	48.9	**51.8**	50.3	46.8	54.0	53.3
+COMBINED	61.2	62.0	61.6	50.4	50.6	**50.5**	**48.1**	**55.8**	**54.0**

Table 4: Coreference results on CoNLL 2011 development set and test set

5 Conclusion

In this paper, we aim to improve coreference resolution by reducing recall errors for proper names and common nouns. We propose a pairwise model for predicting mention references that composes of a feature-based classifier and an embedding-based ranker which allows for better recall performance while maintaining precision. Experiments show that our model can also lead to promising improvement over a strong coreference baseline on the CoNLL 2011 data set.

Our study shows that improving recall while maintaining precision is challenging for coreference resolution. For future work, we would like to work on a joint framework for integrating classic feature indicators, low-dimensional mention embeddings, and semantic knowledge from existing knowledge bases, for improving both precision and recall on mention reference prediction.

Acknowledgement

This work was supported in part by DARPA-BAA-12-47 DEFT grant #12475008 and NSF grant BCS-0904822.

References

[1] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–6. Citeseer, 1998.

[2] Breck Baldwin. Cogniac: high precision coreference with limited knowledge and linguistic resources. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45, 1997.

[3] Eric Bengtson and Dan Roth. Understanding the value of features for coreference resolution. In *EMNLP*, pages 294–303, 2008.

[4] Pascal Denis and Jason Baldridge. Specialized models and ranking for coreference resolution. In *EMNLP*, pages 660–669, 2008.

[5] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

[6] Greg Durrett and Dan Klein. Easy victories and uphill battles in coreference resolution. In *EMNLP*, pages 1971–1982, 2013.

[7] Aria Haghighi and Dan Klein. Simple coreference resolution with rich syntactic and semantic features. In *EMNLP*, pages 1152–1161, 2009.

[8] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. *Advances in neural information processing systems*, pages 115–132, 1999.

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

[10] Xiaoqiang Luo. On coreference resolution performance metrics. In *EMNLP*, pages 25–32, 2005.

[11] Sebastian Martschat and Michael Strube. Recall error analysis for coreference resolution. In *EMNLP*, pages 2070–2081, 2014.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[13] Vincent Ng. Supervised noun phrase coreference research: The first fifteen years. In *ACL*, pages 1396–1411, 2010.

[14] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *ACL*, pages 104–111, 2002.

[15] Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. Scoring coreference partitions of predicted mentions: A reference implementation. In *ACL*, pages 22–27, 2014.

[16] Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *CoNLL*, pages 1–27, 2011.

[17] Altaf Rahman and Vincent Ng. Supervised models for coreference resolution. In *EMNLP*, pages 968–977, 2009.

[18] W.M. Soon, H.T. Ng, and C.Y. Lim. Corpus-based learning for noun phrase coreference resolution. In *EMNLP*, pages 285–291, 1999.

[19] Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. Coreference resolution with reconcile. In *ACL*, pages 156–161, 2010.

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