Unsupervised Coreference Resolution by Utilizing the Most Informative Relations

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

In this paper we present a novel method for unsupervised coreference resolution. We introduce a precision-oriented inference method that scores a candidate entity of a mention based on the most informative mention pair relation between the given mention entity pair. We introduce an informativeness score for determining the most precise relation of a mention entity pair regarding the coreference decisions. The informativeness score is learned robustly during few iterations of the expectation maximization algorithm. The proposed unsupervised system outperforms existing unsupervised methods on all benchmark data sets.

1 Introduction

Due to the advent of the internet, the world wide web, social media, the electronic distribution of information and new means of communication, the amount of text available in many different languages is rising. Natural language processing (NLP) is in charge of automatic processing this growing data. NLP research has mainly focused on English and very few other languages. Therefore there is a rich set of annotated corpora for linguistic analysis tasks for these languages. However, there are no such corpora for thousands of other languages. Since unsupervised methods do not require annotated data for learning a model, employing unsupervised methods has become a popular and important area of research in NLP.

In this paper, we propose a new precision oriented method for unsupervised coreference resolution. Our method evaluates the candidate entities of mentions based on the most precise relation of each mention and its candidate entity. Though we develop and evaluate our method for the English language, we intend to apply it to low resource languages in the future.

Common coreference resolution approaches rely on a combination of different features for each decision (for an overview over such approaches, see Ng (2010)). However, a few approaches break down this combination having precision in mind (Baldwin, 1997; Zhou and Su, 2004; Haghighi and Klein, 2009; Lee et al., 2013). The idea of starting with high precision knowledge is used in various NLP tasks including parsing (Borghesi and Favareto, 1982), word alignment (Brown et al., 1993), and named entity classification (Collins and Singer, 1999) with different names like "islands of reliability", "stepping stones", and "cautiousness". Lee et al. (2013) is a successful recent work that implements this idea as "sieve architecture". Lee et al. (2013) first decide on the basis of more precise features, and then they extend these decisions by using less precise features in later sieves. In this system less precise knowledge is used for extending the decisions made by high precision knowledge.

Our proposed inference method goes in the same direction but in a different way. The probability of each coreference decision is computed based on a single relation of a mention-entity. This single relation is the most precise relation that exist between the mention-entity. In contrast to Lee et al. (2013), our inference method will never take into account less precise relations if more precise ones are present. The relative precision of relations can be determined based on our linguistic intuition. If we would rely on linguistic intuition, our system would look much like Lee et al.'s (2013)'s system, except that it processes

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 644–655, Dublin, Ireland, August 23-29 2014.

all mentions in a single sieve, instead of iterating over all mentions for each input relation. However, it is not a trivial task to determine the relative importance of relations for each new relation, new domain, or new language. In this regard, we propose an informativeness score for automatically determining the relative precision of relations.

The informativeness score is computed based on the distinguishing power of relations among coreferring and non-coreferring mentions. We learn the informativeness score in an unsupervised way via few iterations of the Expectation Maximization (EM) algorithm. Overall, our inference method first finds the most precise relation that a mention has with its candidate entity based on the computed informativeness scores. It then computes the probability of joining the mention to the entity based on this best relation and its distribution among all candidate entities.

We empirically validate our approach on the OntoNotes and ACE data sets, showing that despite being entirely unsupervised, our system performs well on all benchmark data sets.

2 Related Work

Early coreference resolution systems were mainly rule-based systems (Lappin and Leass, 1994; Baldwin, 1997). The success of statistical approaches in different NLP tasks together with the availability of coreference annotated corpora (like MUC-6 (Chinchor and Sundheim, 2003) and MUC-7 (Chinchor, 2001)) facilitated a shift from deploying rule-based methods to machine learning approaches in coreference research in the 1990s.

The increasing importance of multilingual processing, brought the deployment of semi-supervised and unsupervised methods into attention for automatic processing of limited resource languages. There are several works which treat coreference resolution as an unsupervised problem (Cardie and Wagstaff, 1999; Angheluta et al., 2004; Haghighi and Klein, 2007; Ng, 2008; Poon and Domingos, 2008; Haghighi and Klein, 2009; Haghighi and Klein, 2010; Kobdani et al., 2011). We compare our results with the unsupervised systems of Haghighi and Klein (2007), Poon and Domingos (2008), Haghighi and Klein (2009), and Kobdani et al. (2011). The Haghighi and Klein (2010) approach is an almost unsupervised approach, and we do not include this system in our comparisons.

We use the expectation maximization algorithm for unsupervised learning. EM has been previously used for coreference resolution (Cherry and Bergsma, 2005; Ng, 2008; Charniak and Elsner, 2009). Cherry and Bergsma (2005) and Charniak and Elsner (2009) use EM for pronoun resolution, and Ng (2008) models coreference resolution as EM clustering. The model parameters of Ng (2008) are of the form $P(f_1, \ldots, f_k|C_{ij})$, where f_i is a feature, and C_{ij} corresponds to the coreference decision of two mentions m_i and m_j. These parameters along with the entity set, are two sets of unknown variables in Ng (2008). He computes the posterior probabilities of entities in the E-step, and determines the parameters from the N-best clustering (i.e. estimated entities) in the M-step. Ng (2008) starts from an initial guess about the entities and determines the parameters based on this initial guess (M-step). In order to compute the N-best clustering, Ng (2008) uses the Bell tree approach of Luo et al. (2004).

The informativeness scores of mention pair relations (Section 3.2.1) are our unknown parameters. Our inference method only requires the ranking of the informativeness scores (and not their exact values). Therefore, it is much easier to estimate the ranking of these parameters than parameters like $P(f_1, \ldots, f_k|C_{ij})$, and our search space for finding an optimized ranking of the informativeness scores is very small. Since it is easier to have an initial guess about the ranking of informativeness scores (rather than guessing an initial entity set), we start from an E-step with a random ranking.

In our experiments, EM converges very fast regardless of the initial state. Indeed, in the M-step, we use our new inference method for computing an estimation of entities. The use of the EM algorithm in our approach is discussed in more detail in Section 3.3.

3 Method Description

Our coreference resolution method is a mention-entity approach which works at mention-mention granularity for processing candidate entities. It estimates entities incrementally while processing the mentions.

For resolving each mention, our inference method scores all candidate entities. For scoring each candidate entity, it first finds the most informative mention-mention relation that exists between the mention and the candidate entity. It then computes the probability of joining the mention to the entity (i.e. the score of the candidate entity) based on the distribution of this relation among all candidate entities of the mention.

In order to find the best mention-mention relation of a mention and an entity, we introduce an informativeness score that scores mention pair relations based on their association with coreference links. This measure is a global measure, and it is computed based on the association analysis of the mention pair relations and coreference links on a whole entity set of all input documents.

We learn the informativeness score in an unsupervised way by using the EM algorithm. Inference is performed at each E-step of the EM iterations. At each E-step, the whole set of entities is constructed from scratch. The informativeness score of the input relations is computed in the M-step based on the estimated entities of the E-step.

3.1 Notations

Assume that M is a mention set of the input document, and each document consists of a set of entities E in which each entity contains one or more mentions of M. $R = \{r_1, \ldots, r_K\}$ is a set of input relations with the following property:

$$\forall r \in R: \; r(m,n) \in \{0,1\} \tag{1}$$

where m and n are two mentions and r can be any arbitrary relation between two mentions like having a specific feature-value (in which the feature can be a combinational feature), or a linguistic rule.

In order to capture the natural left-to-right ordering of mentions, $r(m,n)$ is zero when n is positioned after m in the input document.

3.2 Inference Method

The inference method processes mentions in the text from the beginning of a document to its end. Initially, each mention is in its own entity. For each mention $m \in M$, all partial entities that have been estimated so far (i.e. entities constructed while processing mentions which are positioned before m) are considered as candidate entities of m (i.e. E_m).

For each candidate entity u, the inference method first determines the best relation among all existing mention pair relations between m and u that can indicate a coreference link based on the informativeness score. We call this relation r_u:

$$r_u = \underset{r \in R}{\operatorname{argmax}} \; (IS(r) \times \max_{n \in u} r(m,n)) \tag{2}$$

where $IS(r)$ is the informativeness score of the r relation.

Apparently, when $IS(r) \times \max_{n \in u} r(m,n)$ is equal to zero, u will be removed from E_m.

After finding the most informative relation that exists between m and u (i.e. r_u), we compute the probability of joining m to u based on r_u as follows:

$$Pr[m \rightarrow u] = \frac{\sum_{n \in u} r_u(m,n)}{\sum_{v \in E_m} \sum_{x \in v} r_u(m,x)} \tag{3}$$

Equation 3 computes the local distribution of r_u among all entities belonging to E_m. After computing the probability of Equation 3 for all candidate entities, m will be joined to the $\hat{u}$ that has the highest probability:

$$\hat{u} = \underset{u \in E_m}{\operatorname{argmax}} \, Pr[m \rightarrow u] \tag{4}$$

In case of a tie condition ($\forall_{u,v \in E_m} \; Pr[m \rightarrow u] = Pr[m \rightarrow v]$), $\hat{u}$ will be the entity whose most informative relation is more precise than the most informative relation of the other candidates:

$$\hat{u} = \underset{u \in E}{\operatorname{argmax}} \, [\max_{r \in R} (IS(r) \times \max_{n \in u} r(n,m))] \tag{5}$$

After finding the best candidate entity of m, the method proceeds to find the best entity of the next mention, based on the new updated E.

A mention m will be left in its own entity in two cases: 1) when E_m is empty, and 2) when the value of $Pr[m \rightarrow \hat{u}]$ is below a predefined threshold. We consider this threshold equal to 0.5 in our experiments. This threshold indicates situations in which less than half of the occurrences of $r_{\hat{u}}$ exist between m and $\hat{u}$, and the others are spread among other entities. This entity can be extended while processing later mentions or it may remain as a singleton.

Please note that the inference method does not care about the exact values of $\{IS(r)\}$, and it only needs to have a ranking of the informativeness scores for the given relations in order to select the most informative one.

3.2.1 Informativeness Score

We want to score a set of given relations based on their discriminative power in making coreference decisions. From a statistical point of view, this can be expressed as to determine whether the existence of a relation indicates a coreference link or is due to chance. In this regard, we can examine the following two hypotheses:

$$\textbf{Hypothesis 0: } P(C = 1|r = 1) = p = P(C = 1|r = 0) \tag{6}$$

$$\textbf{Hypothesis 1: } P(C = 1|r = 1) = p_1 \neq p_2 = P(C = 1|r = 0) \tag{7}$$

where $C \in \{0, 1\}$ is a random variable for coreference decisions.

Hypothesis 0 (null hypothesis) formalizes independence (the coreference decisions are independent of relation r). Hypothesis 1 formalizes dependence, which in case $p_1 \gg p_2$ indicates a strong positive association between r and C. This is the pattern that we are interested in.

We use the G^2 log-likelihood ratio statistics for testing these hypotheses. The statistics was introduced to the NLP community by Dunning (1993), and is defined as follows:

$$-2 \log \lambda = 2 \cdot \log \frac{L(H1)}{L(H0)} \tag{8}$$

where $L(H)$ is the likelihood of a hypothesis based on observed data assuming a binomial probability distribution for the existence of r between coreferring mentions. Asymptotically, $-2 \log \lambda$ is χ^2 distributed with one degree of freedom.

Assuming that we have the whole set of entities of input documents, we can use the maximum likelihood estimator to compute p_1, p_2, and p as follows:

$$\begin{aligned}
p_1 &= \frac{\sum_{u \in E} \sum_{m \in u} \sum_{\substack{n \in u \\ n \neq m}} r(m, n)}{\sum_{x \in M} \sum_{\substack{y \in M \\ y \neq x}} r(x, y)} \\
p_2 &= \frac{\sum_{u \in E} \sum_{m \in u} \sum_{\substack{n \in u \\ n \neq m}} (1 - r(m, n))}{\sum_{x \in M} \sum_{\substack{y \in M \\ y \neq x}} (1 - r(x, y))} \\
p &= \frac{\sum_{u \in E} \sum_{m \in u} \sum_{\substack{n \in u \\ n \neq m}} 1}{\sum_{x \in M} \sum_{\substack{y \in M \\ y \neq x}} 1}
\end{aligned} \tag{9}$$

The log-likelihood ratio statistics can be used both for filtering out non-informative relations and for scoring the remaining relations. The filtering is done by comparing the value of $-2 \log \lambda$ to the desired threshold value obtained from the χ^2 table (15.0 in our experiments) and removing the relations that are not significant at the desired level.

Similar to Dunning (1993), the test statistics can be used as a measure for scoring. In our formulation, the test statistics scores given mention pair relations based on their association with coreference links in a way that more precise relations (relations that indicate a coreference link more strongly) will get a

higher score, and less precise relations (relations that are randomly spread among coreferring and non-coreferring mentions) will get a lower score.

The formulation of the log-likelihood ratio in Dunning (1993) is a two-tailed statistical test that if p_1 and p_2 significantly diverge from each other, the $-2\log\lambda$ would get a high value. However, as mentioned above, we are just interested in the cases that p_1 is much higher than p_2, because, otherwise, coreference links among the mentions which have the relation r in common are less frequent than expected.

Therefore, we use the one-sidedness condition as discussed by Kiss and Strunk (2006) for the log-likelihood test. In this case, a relation r is selected as an informative relation for coreference resolution when the $-2\log\lambda$ is larger than the desired threshold, and also $p_1 > p_2$:

$$IS(r) = \begin{cases} -2\log\lambda & \text{if } p_1 > p_2 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

We compute the values of $\{IS(r)\}$ based on entities of the whole set of input documents in order to have a global estimation of the associations in the input data. In order to have a domain- or genre-specific model, one should learn different $\{IS(r)\}$ for each different domain/genre. The domain/genre adaptation is discussed in more detail in the discussion part.

3.3 Learning Method

From what we have discussed so far, $\{IS(r)\}$ values and document entities (E) are two unknown sets of variables that we want to find. When $\{IS(r)\}$ is known, we can estimate entities by using the inference method described in Section 3.2. When the entities are known, we can compute the $\{IS(r)\}$ as described in Section 3.2.1. We can see that these two steps (i.e. determining entities and the informativeness scores), correspond to the E- and M-steps of the expectation maximization algorithm, respectively.

Expectation maximization is an iterative procedure for computing the maximum likelihood estimator of a parameter set when only a subset of data is available. The EM model involves some hidden variables (Z), observed data (X) and a set of unknown parameters (θ). In our modeling, the informativeness scores are the unknown parameters, the observed data is a set of relations corresponding to R, and entities are hidden variables.

In the M-step, the model estimates $\{IS(r)\}$ by using the association analysis of mention pair relations and coreference links over the entire entity set of the input documents. In the E-step, the algorithm performs the inference method of Section 3.2 and reconstructs the whole set of entities based on the given $\{IS(r)\}$ values. As mentioned before, the inference method only needs the ranking of the informativeness scores, and therefore different values of $\{IS(r)\}$ with similar ordering will lead to the same result. Our model starts from an initial E-step, in which the values of $\{IS(r)\}$ are ranked randomly. The iteration between the E- and M-steps continues until $\{IS(r)\}$ converges to steady values. The convergence and the initial state of the EM algorithm are discussed in more detail in the discussion part.

4 Experiments

4.1 Mention Pair Relations

Here is the list of pairwise relations that we use for common and proper nouns:

- *String match:* Two mentions have the same string after removing their post-modifiers.
- *Compatible head match:* Two mentions have the same head, and the pre-modifiers of the anaphor are a subset of the pre-modifiers of the antecedent.
- *Proper head match:* Two proper names have the same head, and they do not contain numeric or location pre-modifiers.
- *Substring:* All words of the anaphor appear in the antecedent (possibly in different order).
- *Acronym:* One mention is an acronym of the other.

For the ACE data, we use additionally the following relations:

- *Apposition:* Two mentions are in an apposition structure.
- *Demonym:* One mention is a name for a resident of a place that derives from the name of the place, and the other mention is the place name itself.
- *Predicate nominative:* The anaphor follows a linking verb and renames or describes the subject mention.
- *Role apposition:* The antecedent (with a noun head) is a modifier of a noun phrase whose head is the anaphor.

For the OntoNotes data sets, *Same speaker* (Lee et al., 2013) is the only feature for resolving pronouns. For the ACE data *Relative pronoun* (i.e. the anaphor is a relative pronoun that modifies the head of the antecedent) is also used. Pronouns, for which we do not have any feature, are linked to the nearest antecedent (based on the Hobbs distance) that currently belongs to a partial entity which is compatible with the pronoun. The compatibility is measured in terms of number, gender, person, animacy, and named entity label. This approach corresponds to the pronoun resolution strategy of the Stanford system.

The differences between the relations of the OntoNotes and ACE corpora is due to the fact that these two corpora have different annotation schemes. Some of the relations mentioned (e.g. *Apposition*) are considered as coreference relations only in the ACE data.

4.2 Data

We evaluate our method on the following data sets:

- **OntoNotes-Dev**: Development set of the OntoNotes data provided by the CoNLL2012 shared task (Pradhan et al., 2012). This data set consists of 303 documents.
- **OntoNotes-Test**: Test set of the OntoNotes data provided by the CoNLL2012 shared task (Pradhan et al., 2012). This data set consists of 322 documents.
- **ACE2004-nwire**: Newswire subset of the ACE 2004 data set consisting of 128 documents. This split of ACE2004 has been utilized in previous work (Poon and Domingos, 2008; Finkel and Manning, 2008; Haghighi and Klein, 2009; Lee et al., 2013).
- **ACE2004-Culotta-Test**: One of the test splits of the ACE 2004 data set that has been used in previous work (Culotta et al., 2007; Bengtson and Roth, 2008; Haghighi and Klein, 2009; Lee et al., 2013). This data set consists of 107 documents.
- **ACE2003-BNEWS**: BNEWS subset of the ACE 2003 data set utilized in Ng (2008) and Kobdani et al. (2011) consisting of 51 documents.
- **ACE2003-NWIRE**: NWIRE subset of the ACE 2003 data set utilized in Ng (2008) and Kobdani et al. (2011) consisting of 29 documents.

4.3 Preprocessing

The mention detection of the Stanford coreference system (Lee et al., 2013) is used for the OntoNotes data sets. We use the predicted information in the OntoNotes data sets for named entity labels, and syntactic roles. For experiments on the ACE data sets, gold mentions are used, so that comparison with previous work is possible. For preprocessing, the Stanford parser (Klein and Manning, 2003) and named entity recognizer (Finkel et al., 2005) are deployed.

We also use the singleton detection of the Stanford system (Recasens et al., 2013) for the OntoNotes data sets. When both mentions are detected as a singleton by the singleton detection module, the value of all their corresponding relations will be set to zero. In other words, $r(m, n)$ is set to zero when both n and m have been detected as a singleton. For examining the effect of the singleton detection module

		MUC			B^3			$CEAF_e$			Avg.
System		R	P	F1	R	P	F1	R	P	F1	F1
OntoNotes-Test											
Supervised	Berkeley	67.48	72.97	70.12	54.4	61.94	57.92	53.84	55.48	54.65	60.90
	IMS	65.23	70.10	76.58	49.41	60.69	54.47	51.34	49.14	50.21	57.42
Rule-based	Stanford	63.95	65.43	64.68	48.65	56.66	52.35	51.04	46.77	48.81	55.28
Unsupervised	This Work	65	64.27	64.64	49.96	55.35	52.52	51.82	46.66	49.11	55.42
OntoNotes-Dev											
Unsupervised	This Work	65.05	65.69	65.37	51.78	58.31	54.85	54.26	48.72	51.34	57.19
	− Singleton	65.44	63.83	64.62	52.26	56.29	54.2	54.63	46.45	50.21	56.34
	& Genre	65.09	65.7	65.39	51.84	58.31	54.89	54.26	48.75	51.36	57.21

Table 1: Experimental results on OntoNotes data sets.

in our inference method, we evaluate our system without this module. The result is shown in Table 1 (specified as "− Singleton"). The results of the Stanford system are also reported using the singleton detection module of Recasens et al. (2013).

	MUC			B^3		
System	R	P	F1	R	P	F1
ACE2003-NWIRE						
This Work	72.92	86.13	78.98	74.68	90.05	81.65
Haghighi07	44.7	55.5	49.5	-	-	-
Ng08	47.0	68.3	55.7	-	-	-
Kobdani11 (UNSEL)	68.6	64.8	66.6	73.6	61.5	67.0
ACE2003-BNEWS						
This Work	67.36	84.72	75.05	70.35	89.56	78.80
Haghighi07	56.8	68.3	62.0	-	-	-
Ng08	56.1	71.4	62.8	-	-	-
Kobdani11 (UNSEL)	65.0	69.5	67.1	65.9	70.2	68.0
ACE2004-nwire						
This Work	74.77	84.53	79.35	74.21	87.50	80.31
Haghighi07	62.3	66.7	64.2	-	-	-
Poon08	71.3	70.5	70.9	-	-	-
Haghighi09	75.09	77.0	76.5	74.5	79.4	76.9
ACE2004-Culotta-Test						
This Work	68.88	82.42	75.04	73.62	88.87	80.53
Haghighi09	77.7	74.8	79.6	78.5	79.6	79.0

Table 2: Comparison with other unsupervised systems on ACE data sets.

4.4 Results

We evaluate our proposed model with the most commonly used metrics for coreference resolution: for the OntoNotes data sets MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF (Luo, 2005) and their average F1 as used in the CoNLL 2011 and 2012 shared tasks; for the ACE data sets MUC and B^3. The experimental results for the OntoNotes and ACE data sets are presented in Tables 1 and 2, respectively.

On the OntoNotes test set, we compare our method with the three best publicly available coreference systems including the Berkeley system (Durrett and Klein, 2013), the IMS system (Björkelund and Farkas, 2012), and the Stanford system (Lee et al., 2013; Recasens et al., 2013). The Berkeley and IMS systems are both supervised approaches with a rich set of lexical features. At the other hand, the Stanford system is a deterministic system with a set of entity-level features that needs to go through all mentions for incorporating each of the input features. The Stanford system is the winner of the CoNLL2011 shared

OntoNotes-Dev
Same speaker > *Compatible head match* > *Substring* > *String match* > *Proper head match* > *Acronym*
ACE2004-nwire
Compatible head match > *Substring* > *Proper head match* > *String match* > *Demonym* > *Apposition* > *Same speaker* > *Role apposition* > *Relative pronoun* > *Acronym* > *Predicate nominative*

Table 3: The resulting ranking of informativeness scores on different data sets.

task. The IMS system is the 3rd best system on the CoNLL2012 shared task. The Berkeley system is a state-of-the-art supervised coreference system that outperforms both the Stanford and IMS systems. Despite being totally unsupervised and using pairwise features, the results of our system are on par with those of the Stanford system (according to the approximate randomization test, there is no significant difference). The comparison with this state-of-the-art rule based system (Lee et al., 2013), indicates the effectiveness of our coreference resolution approach, as it uses the same preprocessing modules and a simpler and smaller set of features. All results in the Table 1 are reported using the scorer-v7[1] of the CoNLL-2012 shared task (Pradhan et al., 2014).

On the ACE data sets, we compare our performance to those of the unsupervised systems mentioned in Section 2. As Table 2 shows, our method considerably outperforms other unsupervised systems on all data sets (except only for the MUC measure on the ACE2004-Culotta-Test data set).

5 Discussion

5.1 Informativeness Score

As discussed in Section 3.2.1, we determine the discriminative power of mention pair relations in coreference decisions based on the informativeness score (Equation 10), in which the statistical test is computed on the unsupervised estimated set of entities. The resulting ranking of the informativeness score for our input relations is presented in Table 3 on both OntoNotes and ACE data sets.

Another point that needs to be mentioned here is that we are currently using a set of simple and precise input relations. While using these input relations, the informativeness score cannot be efficiently used. The effectiveness of our informativeness score can be usefully assessed with complex relations (i.e. combinatorial features). However, learning of the informativeness scores for complex relations is not possible in a totally unsupervised configuration and one should at least use an informative initial state to guide the learning. We address this issue in our future work.

5.2 Domain/Genre Adaptation

The OntoNotes data set has seven genres regarding the type of text's sources: newswire (NW), broadcast news (BN), broadcast conversation (BC), magazine (MZ), telephone conversation (TC), web data (WB), pivot text (PT). Domain or genre adaptation is one of the current obstacles in language processing. In order to test the effect of genre adaptation in our approach, we try a variant of our approach in which the informativeness scores of the input relations (i.e. $\{IS(r)\}$) are learned separately for each genre. The results of this evaluation are presented in Table 1 by the name "& Genre".

As can be seen in Table 1, the genre-specific variant of our system is performing as well as the base version. This experiment indicates the robustness of our approach regarding the genre/domain adaptation. It can learn an appropriate approximation of the informativeness scores from a small amount of data (i.e. the data provided for a single genre instead of the data from all genres). The learned orderings of the informativeness scores for all genres are presented in Table 4.

When evaluated on each genre separately, the system has the best performance on PT, and the worst performance on the WB genre. The total ordering of genres based on the performance of our system is

[1] http://conll.cemantix.org/2012/software.html

Broadcast conversation, Web data
Same speaker > *Compatible head match* > *Substring* > *String match* > *Proper head match* > *Acronym*
Telephone conversation
Same speaker > *Compatible head match* > *Substring* > *String match* > *Proper head match*
Broadcast news, Newswire
Substring > *Compatible head match* > *String match* > *Proper head match* > *Same speaker* > *Acronym*
Pivot text
Same speaker > *Compatible head match* > *String match* > *Substring* > *Proper head match*
Magazine
Compatible head match > *Substring* > *String match* > *Proper head match* > *Same speaker* > *Acronym*

Table 4: The genre-specific ranking of informativeness scores.

as follow: PT, MZ, TC, BN, NW, BC, WB.

5.3 EM Initial State and Convergence

For the initial state of our EM algorithm, we need a ranking of the informativeness scores of the input relations. We try different initial states for the EM algorithm, from an informative ranking based on linguistic intuition about the precision of input relations to a misleading ranking (the informative order reversed). However, in all cases, the EM algorithm leads to the same ranking (as listed in Table 3). This indicates the robustness of our modeling.

It is more likely that a more precise relation will also get a higher value for its corresponding join probability of Equation 3, because it is unlikely that a precise relation connects a mention to several candidate entities. However, relations with low precision may connect a mention to several different entities, because they are spread over more different entities than relations with higher precision.

In our experiments, for all tested initial states, the model converges in 4 iterations on the OntoNotes data sets and 5 iterations on the ACE data sets.

5.4 Promising Alternative for the Stanford System

Our coreference resolution method is a self- contained approach, that does not need any external linguistic knowledge regarding the coreference relations. However, we can also consider a simple variant of this system in which a predefined ordering of features (based on linguistic intuition) is given, like the Stanford system. In this case, the EM algorithm will be no longer needed, and therefore, the algorithm resolves all mentions in a single iteration.

Therefore, this variant of our system can be considered as an efficient alternative to the Stanford system, that uses a simpler (pairwise instead of entity-based) and smaller (5 instead of 7 string matches) set of relations, and more importantly processes all mentions in a single iteration (instead of iterating over all mentions for each relation), and it still performs as well as its entity-based multi-sieve variant.

6 Conclusions

In this paper, we presented a new unsupervised coreference resolution method. We deploy a new precision-oriented inference method that decides about joining a mention to a candidate entity based on only the most informative mention pair relation that exists between the given mention entity pair. In order to determine the most informative relation of a mention and its candidate entity, we introduce an informativeness score for scoring mention-mention relations based on their global association with

coreference links. A relation whose existence strongly indicates a coreference link will get a high score, and a relation which is randomly spread among coreferring and non-coreferring mentions will get a low score. The informativeness score is robustly learned during a very few iterations of the EM algorithm.

Our proposed method performs well on all benchmark data sets. In the future we intend to apply this robust and efficient approach to new genres, domains, and also new languages.

Acknowledgments

The authors would like to thank Sebastian Martschat for his 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

Roxana Angheluta, Patrick Jeuniaux, Rudradeb Mitra, and Marie-Francine Moens. 2004. Clustering algorithms for noun phrase coreference resolution. In *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles,* Louvain La Neuve, Belgium, 10–12 March 2004, pages 60–70.

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.

Breck Baldwin. 1997. CogNIAC: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Text, Madrid, Spain, July 1997*, pages 38–45.

Eric Bengtson 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,* Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 294–303.

Anders Björkelund and Richárd Farkas. 2012. Data-driven multilingual coreference resolution using resolver stacking. In *Proceedings of the Shared Task of the 16th Conference on Computational Natural Language Learning,* Jeju Island, Korea, 12–14 July 2012, pages 49–55.

Luigi Borghesi and Chiara Favareto. 1982. Flexible parsing of discretely uttered sentences. In *Proceedings of the 9th International Conference on Computational Linguistics,* Prague, Czechoslovakia, 5–10 July 1982, pages 37–42.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Claire Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *Proceedings of the 1999 SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora,* College Park, Md., 21–22 June 1999, pages 82–89.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics,* Athens, Greece, 30 March – 3 April 2009, pages 148–156.

Colin Cherry and Shane Bergsma. 2005. An expectation maximization approach to pronoun resolution. *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 88–95.

Nancy Chinchor and Beth Sundheim. 2003. Message Understanding Conference (MUC) 6. LDC2003T13, Philadelphia, Penn: Linguistic Data Consortium.

Nancy Chinchor. 2001. Message Understanding Conference (MUC) 7. LDC2001T02, Philadelphia, Penn: Linguistic Data Consortium.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the 1999 SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora,* College Park, Md., 21–22 June 1999, pages 100–110.

Aron Culotta, Michael Wick, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. 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 81–88.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

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.

Jenny Rose Finkel and Christopher Manning. 2008. Enforcing transitivity in coreference resolution. In *Companion Volume to the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics,* Columbus, Ohio, 15–20 June 2008, pages 45–48.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* Ann Arbor, Mich., 25–30 June 2005, pages 363–370.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* Prague, Czech Republic, 23–30 June 2007, pages 848–855.

Aria Haghighi 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,* Singapore, 6–7 August 2009, pages 1152–1161.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity centered model. In *Proceedings of Human Language Technologies 2010: The Conference of the North American Chapter of the Association for Computational Linguistics,* Los Angeles, Cal., 2–4 June 2010, pages 385–393.

Tibor Kiss and Jan Strunk. 2006. Unsuperivsed multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Sapporo, Japan, 7–12 July 2003, pages 423–430.

Hamidreza Kobdani, Hinrich Schuetze, Michael Schiehlen, and Hans Kamp. 2011. Bootstrapping coreference resolution using word associations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics,* Portland, Oreg., 19–24 June 2011, pages 783–792.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

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, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the Bell Tree. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July 2004, pages 136–143.

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.

Vincent Ng. 2008. Unsupervised models for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 640–649.

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.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 650–659.

Sameer Pradhan, Alessandro Moschitti, and Nianwen Xue. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Shared Task of the 16th Conference on Computational Natural Language Learning,* Jeju Island, Korea, 12–14 July 2012, pages 1–40.

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 ACL 2014 Conference Short Papers,* Baltimore, Md., 22–27 June 2014. To appear.

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.

Guodong Zhou and Jian Su. 2004. A high-performance coreference resolution system using a constraint-based multi-agent strategy. In *the 16th International Conference on Computational Linguistics (COLING)*.