UBIU: A Robust System for Resolving Unrestricted Coreference

Desislava Zhekova
University of Bremen
zhekova@uni-bremen.de

Sandra Kübler
Indiana University
skuebler@indiana.edu

Abstract

In this paper, we discuss the application of UBIU to the CoNLL-2011 shared task on "Modeling Unrestricted Coreference" in OntoNotes. The shared task concentrates on the detection of coreference not only in noun phrases but also involving verbs. The information provided for the closed track included WordNet as well as corpus generated number and gender information. Our system shows no improvement when using WordNet information, and the number information proved less reliable than the information in the part of speech tags.

1 Introduction

Coreference Resolution is the process of identifying the linguistic expressions in a discourse that refer to the same real world entity and to divide those expressions into equivalence classes that represent each discourse entity. For this task, a deeper knowledge of the discourse is often required. However, such knowledge is difficult to acquire. For this reason, many systems use superficial information such as string match. The CoNLL shared task on "Modeling Unrestricted Coreference in OntoNotes" (Pradhan et al., 2011) presents challenges that go beyond previous definitions of the task. On the one hand, mention extraction is part of the task while many previous approaches assumed gold standard mentions. On the other hand, coreference is not restricted to noun phrases, verbs are also included. Thus, in *Sales of passenger cars grew 22%. The strong growth followed year-to-year increases.*, the verb *grew* has an identity relation with the noun phrase *The strong growth*.

The system that we used for the shared task is the memory-based machine learning system UBIU (Zhekova and Kübler, 2010). We describe the most important components of the system in section 2. The system was originally developed for robust, multilingual coreference resolution, and thus had to be adapted to this shared task. We investigate the quality of our mention extraction in section 2.1 and the quality of the features used in the classifier in section 2.2. In section 3, we present UBIU's results on the development set, and in section 4, UBIU's final results in the shared task.

2 UBIU

UBIU (Zhekova and Kübler, 2010) was developed as a multilingual coreference resolution system. A robust approach is necessary to make the system applicable for a variety of languages. For this reason, we use a machine learning approach to classify mention pairs. We use TiMBL (Daelemans et al., 2007), a memory-based learner (MBL) that labels the feature vectors from the test set based on the k nearest neighbors in the training instances. We chose TiMBL since MBL has been shown to work well with small training sets. A non-exhaustive parameter optimization on the development set led us to use the *IB1* algorithm, similarity is computed based on weighted overlap, the relevance weights are computed using gain ratio and the number of nearest neighbors is set to $k = 3$ (for a description of the algorithm and parameters cf. (Daelemans et al., 2007)). The classifier is preceded by a mention extractor, which identifies possible mentions, and a feature extractor. The latter creates a feature vector for each possible pair of a potentially coreferring

mention and all possible antecedents in a context of 3 sentences. Another important step is to separate singleton mentions from coreferent ones since only the latter are annotated in OntoNotes. Our markable extractor overgenerates in that it extracts all possible mentions, and only after classification, the system can decide which mentions are singletons. We investigate the performance of the mention and feature extraction modules in more detail below.

2.1 Mention Extraction

UBIU's mention extractor uses part-of-speech (POS), syntactic, and lemma information provided in the OntoNotes data set to detect mentions. The module defines a mention for each noun phrase, based on syntactic information, as well as for all possessive pronouns and all proper nouns, based on their POS tags. Since for the shared task, verbs are also potentially coreferent, we included a mention for each of the verbs with a predicate lemma. An example of the output of the mention extraction module is shown in table 1. Each mention is numbered with an individual number and thus still represents a distinct entity. Since singleton mentions are not annotated in the OntoNotes data set, mentions without coreference relations after classification need to be removed from the answer set, which can only be performed after coreference resolution when all coreferent pairs are identified. For this reason, the markable extractor is bound to overgenerate. The latter can clearly be seen when the mention extraction output is compared to the provided gold mentions (cf. the last column in table 1).

We conducted a simple experiment on the development data in order to gain insight into the performance of the mention extraction module. Using the scorer provided by the shared task, we evaluated the output of the module, without performing coreference resolution and without removing singleton mentions. This led to a recall of 96.55 % and a precision of 18.55%, resulting in an F-score of 31.12. The high recall shows that the system is very reliable in finding mentions with the correct boundaries. However, since we do not remove any singletons, UBIU overgenerates and thus the system identified a considerable number of singletons, too. Nevertheless, the fact that UBIU identified 96.55% of all mentions shows that the performance of the mention extraction module is close to optimal.

#	Word	POS	Parse bit	ME output	Gold
0	Devastating	VBG	(TOP(NP(NP*	(1)|(2|(3	-
1	Critique	NN	*)	3)	-
2	of	IN	(PP*	-	-
3	the	DT	(NP*	(4	(32
4	Arab	JJ	*	-	-
5	World	NN	*))	4)	32)
6	by	IN	(PP*	-	-
7	One	CD	(NP(NP*)	(5|(6)	-
8	of	IN	(PP*	-	-
9	Its	PRP$	(NP*	(7)|(8	(32)
10	Own	JJ	*))))))	8)|5)|2)	-

Table 1: The output of the mention extractor for a sample sentence.

2.2 Feature Extraction

Feature extraction is the second important subtask for the UBIU pipeline. Since mentions are represented by their syntactic head, the feature extractor uses a heuristic that selects the rightmost noun in a noun phrase. However, since postmodifying prepositional phrases may be present in the mention, the noun may not be followed by a preposition. For each mention, a feature vector is created for all of its preceding mentions in a window of 3 sentences. After classification, a filter can optionally be applied to filter out mention pairs that disagree in number, and another filter deletes all mentions that were not assigned an antecedent in classification. Note that the number information was derived from the POS tags and not from the number/gender data provided by the shared task since the POS information proved more reliable in our system.

Initially, UBIU was developed to use a wide set of features (Zhekova and Kübler, 2010), which constitutes a subset of the features described by Rahman and Ng (2009). For the CoNLL-2011 shared task, we investigated the importance of various additional features that can be included in the feature set used by the memory-based classifier. Thus, we conducted experiments with a base set and an extended feature set, which makes use of lexical semantic features.

Base Feature Set Since the original feature set in Zhekova and Kübler (2010) contained information that is not easily accessible in the OntoNotes data set (such as grammatical functions), we had to restrict the feature set to information that can be derived solely from POS annotations. Further infor-

#	Feature Description
1	m_j - the antecedent
2	m_k - the mention to be resolved
3	Y if m_j is a pronoun; else N
4	number - S(ingular) or P(lural)
5	Y if m_k is a pronoun; else N
6	C if the mentions are the same string; else I
7	C if one mention is a substring of the other; else I
8	C if both mentions are pronominal and are the same string; else I
9	C if the two mentions are both non-pronominal and are the same string; else I
10	C if both mentions are pronominal and are either the same pronoun or different only w.r.t. case; NA if at least one of them is not pronominal; else I
11	C if the mentions agree in number; I if they disagree; NA if the number for one or both mentions cannot be determined
12	C if both mentions are pronouns; I if neither are pronouns; else NA
13	C if both mentions are proper nouns; I if neither are proper nouns; else NA
14	sentence distance between the mentions

Table 2: The pool of features used in the base feature set.

mation as sentence distance, word overlap etc. was included as well. The list of used features is shown in table 2.

Extended Feature Set Since WordNet information was provided for the closed setting of the CoNLL-2011 shared task, we also used an extended feature set, including all features from the base set along with additional features derived from WordNet. The latter features are shown in table 3.

2.3 Singletons

In section 2.1, we explained that singletons need to be removed after classification. However, this leads to a drastic decrease in system performance for two reasons. First, if a system does not identify a coreference link, the singleton mentions will be removed from the coreference chains, and consequently, the system is penalized for the missing link as well as for the missing mentions. If singletons are included, the system will still receive partial credit for them from all metrics but MUC. For this reason, we investigated filtered and non-filtered results in combination with the base and the extended feature sets.

3 Results on the Development Set

The results of our experiment on the development set are shown in table 4. Since the official scores of the shared task are based on an average of MUC,

#	Feature Description
15	C if both are nouns and m_k is hyponym of m_j; I if both are nouns but m_k is not a hyponym of m_j; NA otherwise
16	C if both are nouns and m_j is hyponym of m_k; I if both are nouns but m_j is not a hyponym of m_k; NA otherwise
17	C if both are nouns and m_k is a partial holonym of m_j; I if both are nouns but m_k is not a partial holonym of m_j; NA otherwise
18	C if both are nouns and m_j is a partial holonym of m_k; I if both are nouns but m_j is not a partial holonym of m_k; NA otherwise
19	C if both are nouns and m_k is a partial meronym of m_j; I if both are nouns but m_k is not a partial meronym of m_j; NA otherwise
20	C if both are nouns and m_j is a partial meronym of m_k; I if both are nouns but m_j is not a partial meronym of m_k; NA otherwise
21	C if both are verbs and m_k entails m_j; I if both are verbs but m_k does not entail m_j; NA otherwise
22	C if both are verbs and m_j entails m_k; I if both are verbs but m_j does not entail m_k; NA otherwise
23	C if both are verbs and m_k is a hypernym of m_j; I if both are verbs but m_k is not a hypernym of m_j; NA otherwise
24	C if both are verbs and m_j is a hypernym of m_k; I if both are verbs but m_j is not a hypernym of m_k; NA otherwise
25	C if both are verbs and m_k is a troponym of m_j; I if both are verbs but m_k is not a troponym of m_j; NA otherwise
26	C if both are verbs and m_j is a troponym of m_k; I if both are verbs but m_j is not a troponym of m_k; NA otherwise

Table 3: The features extracted from WordNet.

B^3, and CEAFE, we report these measures and their average. All the results in this section are based on automatically annotated linguistic information. The first part of the table shows the results for the base feature set (UBIU$_B$), the second part for the extended feature set (UBIU$_E$). We also report results if we keep all singletons (& Sing.) and if we filter out coreferent pairs that do not agree in number (& Filt.). The results show that keeping the singletons results in lower accuracies on the mention and the coreference level. Only recall on the mention level profits from the presence of singletons. Filtering for number agreement with the base set has a detrimental effect on mention recall but increases mention precision so that there is an increase in F-score of 1%. However, on the coreference level, the effect is negligible. For the extended feature set, filtering results in a decrease of approximately 2.0% in mention precision, which also translates into lower coreference scores. We also conducted an experiment in which we filter before classification (& Filt. BC), following a more standard approach. The reasoning

	IM			MUC			B^3			CEAFE			Average
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	F_1
UBIU$_B$	62.71	38.66	47.83	30.59	24.65	27.30	67.06	62.65	64.78	34.19	40.16	36.94	43.01
UBIU$_B$ & Sing.	95.11	18.27	30.66	30.59	24.58	27.26	67.10	62.56	64.75	34.14	40.18	36.92	42.97
UBIU$_B$ & Filt.	61.30	40.58	48.83	29.10	25.77	27.33	64.88	64.63	64.76	35.38	38.74	36.98	43.02
UBIU$_B$ & Filt. BC	61.33	40.49	48.77	28.96	25.54	27.14	64.95	64.48	64.71	35.23	38.71	36.89	42.91
UBIU$_E$	62.72	39.09	48.16	30.63	24.94	27.49	66.72	62.76	64.68	34.19	39.90	36.82	43.00
UBIU$_E$ & Sing.	95.11	18.27	30.66	29.87	20.96	24.64	69.13	57.71	62.91	32.28	42.24	36.59	41.38
UBIU$_E$ & Filt.	63.01	36.62	46.32	28.65	21.05	24.27	68.10	58.72	63.06	32.91	41.53	36.72	41.35
Gold ME	100	100	100	38.83	82.97	52.90	39.99	92.33	55.81	66.73	26.75	38.19	48.97

Table 4: UBIU system results on the development set.

is that the training set for the classifier is biased towards not assuming coreference since the majority of mention pairs does not have a coreference relation. Thus filtering out non-agreeing mention pairs before classification reduces not only the number of test mention pairs to be classified but also the number of training pairs. However, in our system, this approach leads to minimally lower results, which is why we decided not to pursue this route. We also experimented with instance sampling in order to reduce the bias towards non-coreference in the training set. This also did not improve results.

Contrary to our expectation, using ontological information does not improve results. Only on the mention level, we see a minimal gain in precision. But this does not translate into any improvement on the coreference level. Using filtering in combination with the extended feature set results in a more pronounced deterioration than with the base set.

The last row of table 4 (Gold ME) shows results when the system has access to the gold standard mentions. The MUC and B^3 results show that the classifier reaches an extremely high precision (82.97% and 92.33%), from which we conclude that the coreference links that our system finds are reliable, but it is also too conservative in assuming coreference relations. For the future, we need to investigate undersampling the negative examples in the training set and more efficient methods for filtering out singletons.

4 Final Results

In the following, we present the UBIU system results in two separate settings: using the test set with automatically extracted mentions (section 4.1) and using a test set with gold standard mentions, including singletons (section 4.2). An overview of all systems participating in the CoNLL-2011 shared task and their results is provided by Pradhan et al. (2011).

4.1 Automatic Mention Identification

The final results of UBIU for the test set without gold standard mentions are shown in the first part of table 5. They are separated into results for the coreference resolution module based on automatically annotated linguistic information and the gold annotations. Again, we report results for both the base feature set (UBIU$_B$) and the extended feature set using WordNet features (UBIU$_E$). A comparison of the system results on the test and the development set in the UBIU$_B$ setting shows that the average F-score is considerably lower for the test set, 40.46 vs. 43.01 although the quality of the mentions remains constant with an F-score of 48.14 on the test set and 47.83 on the development set.

The results based on the two data sets show that UBIU's performance improves when the system has access to gold standard linguistic annotations. However, the difference between the results is in the area of 2%. The improvement is due to gains of 3-5% in precision for MUC and B^3, which are counteracted by smaller losses in recall. In contrast, CEAFE shows a loss in precision and a similar gain in recall, resulting in a minimal increase in F-score.

A comparison of the results for the experiments with the base set as opposed to the extended set in 5 shows that the extended feature set using WordNet information is detrimental to the final results averaged over all metrics while it led to a slight improvement on the mention level. Our assumption is that while in general, the ontological information is useful, the additional information may be a mixture of relevant and irrelevant information. Mihalcea (2002) showed for word sense disambiguation that

		IM			MUC			B^3			CEAFE			Average
		R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1	F_1
Automatic Mention Identification														
auto	UBIU$_B$	67.27	37.48	48.14	28.75	20.61	24.01	67.17	56.81	61.55	31.67	41.22	35.82	40.46
	UBIU$_E$	67.49	37.60	48.29	28.87	20.66	24.08	67.14	56.67	61.46	31.57	41.21	35.75	40.43
gold	UBIU$_B$	65.92	40.56	50.22	31.05	25.57	28.04	64.94	62.23	63.56	33.53	39.08	36.09	42.56
	UBIU$_E$	66.11	40.37	50.13	30.84	25.14	27.70	65.07	61.83	63.41	33.23	39.05	35.91	42.34
Gold Mentions														
auto	UBIU$_B$	67.57	58.66	62.80	34.14	40.43	37.02	54.24	71.09	61.53	39.65	33.73	36.45	45.00
	UBIU$_E$	69.19	57.27	62.67	33.48	37.15	35.22	55.47	68.23	61.20	38.29	34.65	36.38	44.27
gold	UBIU$_B$	67.64	58.75	62.88	34.37	40.68	37.26	54.28	71.18	61.59	39.69	33.76	36.49	45.11
	UBIU$_E$	67.72	58.66	62.87	34.18	40.40	37.03	54.30	71.04	61.55	39.64	33.78	36.47	45.02

Table 5: Final system results for the coreference resolution module on automatically extracted mentions on the gold standard mentions for the base and extended feature sets.

memory-based learning is extremely sensitive to irrelevant features. For the future, we are planning to investigate this problem by applying forward-backward feature selection, as proposed by Mihalcea (2002) and Dinu and Kübler (2007).

4.2 Gold Mentions

UBIU was also evaluated in the experimental setting in which gold mentions were provided in the test set, including singletons. The results of the setting using both feature sets are reported in the second part of table 5. The results show that overall the use of gold standard mentions results in an increase of the average F-score of approx. 4.5%. Where mention quality and MUC are concerned, gold standard mentions have a significant positive influence on the average F-score. For B^3 and CEAFE, however, there is no significant change in scores. The increase in performance is most noticeable in mention identification, for which the F-score increases from 48.14 to 62.80. But this improvement in mention identification has a smaller effect on the overall coreference system performance leading to a 5% increase of results.

In contrast to the gold mention results in the development set, we see lower precision values in the test set. This is due to the fact that the test set contains singletons. Detecting singletons reliably is a difficult problem that needs further investigation.

5 Conclusion and Future Work

In the current paper, we presented the results of UBIU in the CoNLL-2011 shared task. We showed that for a robust system for coreference resolution such as UBIU, automatically annotated linguistic data is sufficient for mention-pair based coreference resolution. We also showed that ontological information as well as filtering non-agreeing mention pairs leads to an insignificant improvement of the overall coreference system performance. The treatment of singletons in the data remains a topic that requires further investigation.

References

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2007. TiMBL: Tilburg memory based learner – version 6.1 – reference guide. Technical Report ILK 07-07, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University.

Georgiana Dinu and Sandra Kübler. 2007. Sometimes less is more: Romanian word sense disambiguation revisited. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2007*, Borovets, Bulgaria.

Rada Mihalcea. 2002. Instance based learning with automatic feature selection applied to word sense disambiguation. In *Proceedings of the 19th International Conference on Computational Linguistics, COLING'02*, Taipeh, Taiwan.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 Shared Task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL)*, Portland, Oregon.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of EMNLP 2009*, Singapore.

Desislava Zhekova and Sandra Kübler. 2010. UBIU: A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 96–99, Uppsala, Sweden.