UBIU: A Language-Independent System for Coreference Resolution

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

Sandra Kübler
Indiana University
skuebler@indiana.edu

Abstract

We present UBIU, a language independent system for detecting full coreference chains, composed of named entities, pronouns, and full noun phrases which makes use of memory based learning and a feature model following Rahman and Ng (2009). UBIU is evaluated on the task "Coreference Resolution in Multiple Languages" (SemEval Task 1 (Recasens et al., 2010)) in the context of the 5th International Workshop on Semantic Evaluation.

1 Introduction

Coreference resolution is a field in which major progress has been made in the last decade. After a concentration on rule-based systems (cf. e.g. (Mitkov, 1998; Poesio et al., 2002; Markert and Nissim, 2005)), machine learning methods were embraced (cf. e.g. (Soon et al., 2001; Ng and Cardie, 2002)). However, machine learning based coreference resolution is only possible for a very small number of languages. In order to make such resources available for a wider range of languages, language independent systems are often regarded as a partial solution. To this day, there have been only a few systems reported that work on multiple languages (Mitkov, 1999; Harabagiu and Maiorano, 2000; Luo and Zitouni, 2005). However, all of those systems were geared towards predefined language sets.

In this paper, we present a language independent system that does require syntactic resources for each language but does not require any effort for adapting the system to a new language, except for minimal effort required to adapt the feature extractor to the new language. The system was completely developed within 4 months, and will be extended to new languages in the future.

2 UBIU: System Structure

The UBIU system aims at being a language-independent system in that it uses a combination of machine learning, in the form of memory-based learning (MBL) in the implementation of TiMBL (Daelemans et al., 2007), and language independent features. MBL uses a similarity metric to find the k nearest neighbors in the training data in order to classify a new example, and it has been shown to work well for NLP problems (Daelemans and van den Bosch, 2005). Similar to the approach by Rahman and Ng (2009), classification in UBUI is based on mention pairs (having been shown to work well for German (Wunsch, 2009)) and uses as features standard types of linguistic annotation that are available for a wide range of languages and are provided by the task.

Figure 1 shows an overview of the system. In preprocessing, we slightly change the formatting of the data in order to make it suitable for the next step in which language dependent feature extraction modules are used, from which the training and test sets for the classification are extracted. Our approach is untypical in that it first extracts the heads of possible antecedents during feature extraction. The full yield of an antecedent in the test set is determined after classification in a separate module. During postprocessing, final decisions are made concerning which of the mention pairs are considered for the final coreference chains.

In the following sections, we will describe feature extraction, classification, markable extraction, and postprocessing in more detail.

2.1 Feature Extraction

The language dependent modules contain finite state expressions that detect the heads based on the linguistic annotations. Such a language module requires a development time of approximately 1 person hour in order to adapt the regular expressions

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 96–99,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

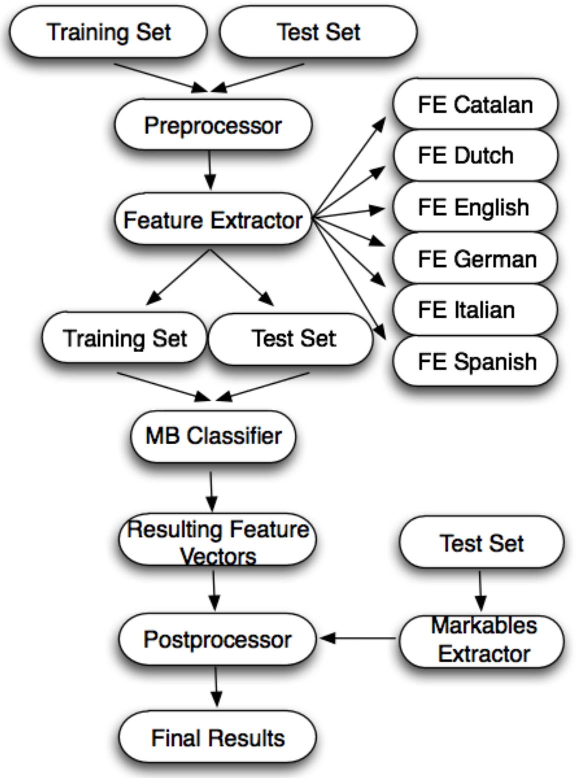

Figure 1: Overview of the system.

to the given language data (different POS tagsets, differences in the provided annotations). This is the only language dependent part of the system.

We decided to separate the task of finding heads of markables, which then serve as the basis for the generation of the feature vectors, from the identification of the scope of a markable. For the English sentence "Any details or speculation on who specifically, we don't know that at this point.", we first detect the heads of possible antecedents, for example "details". However, the decision on the scope of the markable, i.e. the decision between "details" or "Any details or speculation on who specifically" is made in the postprocessing phase.

One major task of the language modules is the check for cyclic dependencies. Our system relies on the assumption that cyclic dependencies do not occur, which is a standard assumption in dependency parsing (Kübler et al., 2009). However, since some of the data sets in the multilingual task contained cycles, we integrated a module in the preprocessing step that takes care of such cycles.

After the identification of the heads of markables, the actual feature extraction is performed. The features that were used for training a classifier (see Table 1) were selected from the feature pool

#	Feature Description
1	m_j - the antecedent
2	m_k - the mention to be resolved
3	Y if m_j is pron.; else N
4	Y if m_j is subject; else N
5	Y if m_j is a nested NP; else N
6	number - Sg. or Pl.
7	gender - F(emale), M(ale), N(euter), U(nknown)
8	Y if m_k is a pronoun; else N
9	Y if m_k is a nested NP; else N
10	semantic class – extracted from the NEs in the data
11	the nominative case of m_k if pron.; else NA
12	C if the mentions are the same string; else I
13	C if one mention is a substring of the other; else I
14	C if both mentions are pron. and same string; else I
15	C if both mentions are both non-pron. and same string; else I
16	C if both m. are pron. and either same pron. or diff. w.r.t. case; NA if at least one is not pron.; else I
17	C if the mentions agree in number; I if not; NA if the number for one or both is unknown
18	C if both m. are pron. I if neither
19	C if both m. are proper nouns; I if neither; else NA
20	C if the m. have same sem. class; I if not; NA if the sem. class for one or both m. is unknown
21	sentence distance between the mentions
22	concat. values for f. 6 for m_j and m_k
23	concat. values for f. 7 for m_j and m_k
24	concat. values for f. 3 for m_j and m_k
25	concat. values for f. 5 for m_j and m_k
26	concat. values for f. 10 for m_j and m_k
27	concat. values for f. 11 for m_j and m_k

Table 1: The pool of features for all languages.

presented by Rahman and Ng (2009). Note that not all features could be used for all languages. We extracted all the features in Table 1 if the corresponding type of annotation was available; otherwise, a null value was assigned.

A good example for the latter concerns the gender information represented by feature 7 (for possible feature values cf. Table 1). Let us consider the following two entries - the first from the German data set and the second from English:

1. Regierung Regierung Regierung NN NN cas=d|num=sg|gend=fem cas=d|num=sg|gend=fem 31 31 PN PN . . .
2. law _ law NN NN NN NN 2 2 PMOD PMOD . . .

Extracting the value from entry 1, where *gend=fem*, is straightforward; the value being *F*. However, there is no gender information provided in the English data (entry 2). As a result, the value for feature 7 is *U* for the closed task.

2.2 Classifier Training

Based on the features extracted with the feature extractors described above, we trained TiMBL. Then we performed a non-exhaustive parameter optimization across all languages. Since a full optimization strategy would lead to an unmanageable number of system runs, we concentrated on varying k, the number of nearest neighbors considered in classification, and on the distance metric.

Furthermore, the optimization is focused on language independence. Hence, we did not optimize each classifier separately but selected parameters that lead to best average results across all languages of the shared task. In our opinion, this ensures an acceptable performance for new languages without further adaptation. The optimal settings for all the given languages were k=3 with the Overlap distance and gain ratio weighting.

2.3 Markable Extraction

The markable extractor makes use of the dependency relation labels. Each syntactic head together with all its dependents is identified as a separate markable. This approach is very sensitive to incorrect annotations and to dependency cycles in the data set. It is also sensitive to differences between the syntactic annotation and markables. In the Dutch data, for example, markables for named entities (NE) often exclude the determiner, a nominal dependent in the dependency annotation. Thus, the markable extractor suggests the whole phrase as a markable, rather than just the NE.

During the development phase, we determined experimentally that the recognition of markables is one of the most important steps in order to achieve high accuracy in coreference resolution: We conducted an ablation study on the training data set. We used the *train* data as training set and the *devel* data as testing set and investigated three different settings:

1. Gold standard setting: Uses gold markable annotations as well as gold linguistic annotations (upper bound).
2. Gold linguistic setting: Uses automatically determined markables and gold linguistic annotations.
3. Regular setting: Uses automatically determined markables and automatic linguistic information.

S	Lang.	IM	CEAF	MUC	B^3	BLANC
1	Spanish	85.8	52.3	12.8	60.0	56.9
	Catalan	85.5	56.0	11.6	59.4	51.9
	English	96.1	68.7	17.9	74.9	52.7
	German	93.6	70.0	19.7	73.4	64.5
2	Spanish	61.0	41.5	11.3	42.4	48.7
	Catalan	60.8	40.5	9.6	41.4	48.3
	English	72.1	54.1	11.6	57.3	50.3
	German	57.7	45.5	12.2	45.7	44.3
3	Spanish	61.2	41.8	10.3	42.3	48.5
	Catalan	61.3	40.9	11.3	41.9	48.5
	English	71.9	54.7	13.3	57.4	50.3
	German	57.5	45.4	12.0	45.6	44.2

Table 2: Experiment results (as F1 scores) where IM is identification of mentions and S - Setting.

Note that we did not include all six languages: we excluded Italian and Dutch because there is no gold-standard linguistic annotation provided. The results of the experiment are shown in Table 2. From those results, we can conclude that the figures in Setting 2 and 3 are very similar. This means that the deterioration from gold to automatically annotated linguistic information is barely visible in the coreference results. This is a great advantage, since gold-standard data has always proved to be very expensive and difficult or impossible to obtain. The information that proved to be extremely important for the performance of the system is the one providing the boundaries of the markables. As shown in Table 2, the latter leads to an improvement of about 20%, which is observable in the difference in the figures of Setting 1 and 2. The results for the different languages show that it is more important to improve markable detection than the linguistic information.

2.4 Postprocessing

In Section 2.1, we described that we decided to separate the task of finding heads of markables from the identification of the scope of a markable. Thus, in the postprocessing step, we perform the latter (by the Markables Extractor module) as well as reformat the data for evaluation.

Another very important step during postprocessing is the selection of possible antecedents. In cases where more than one mention pair is classified as coreferent, only the pair with highest confidence by TiMBL is selected. Since nouns can be discourse-new, they do not necessarily have a coreferent antecedent; pronouns however, require an antecedent. Thus, in cases where all possible antecedents for a given pronoun are classified as not coreferent, we select the closest subject as antecedent; or if this heuristic is not successful, the antecedent that has been classified as not coreferent with the lowest confidence score (i.e. the highest distance) by TiMBL.

Lang.	S	IM	CEAF	MUC	B^3	BLANC
Catalan	G	84.4	52.3	11.7	58.8	52.2
	R	59.6	38.4	8.6	40.9	47.8
English	G	95.9	65.7	20.5	74.8	54.0
	R	74.2	53.6	14.2	58.7	51.0
German	G	94.0	68.2	21.9	75.7	64.5
	R	57.6	44.8	10.4	46.6	48.0
Spanish	G	83.6	51.7	12.7	58.3	54.3
	R	60.0	39.4	10.0	41.6	48.4
Italian	R	40.6	32.9	3.6	34.8	37.2
Dutch	R	34.7	17.0	8.3	17.0	32.3

Table 3: Final system results (as F1 scores) where IM is identification of mentions and S - Setting. For more details cf. (Recasens et al., 2010).

3 Results

UBIU participated in the closed task (i.e. only information provided in the data sets could be used), in the gold and regular setting. It was one of two systems that submitted results for all languages, which we count as preliminary confirmation that our system is language independent. The final results of UBIU are shown in Table 3. The figures for the identification of mentions show that this is an area in which the system needs to be improved. The errors in the gold setting result from an incompatibility of our two-stage markable annotation with the gold setting. We are planning to use a classifier for mention identification in the future.

The results for coreference detection show that English has a higher accuracy than all the other languages. We assume that this is a consequence of using a feature set that was developed for English (Rahman and Ng, 2009). This also means that an optimization of the feature set for individual languages should result in improved system performance.

4 Conclusion and Future Work

We have presented UBIU, a coreference resolution system that is language independent (given different linguistic annotations for languages). UBIU is easy to maintain, and it allows the inclusion of new languages with minimal effort.

For the future, we are planning to improve the system while strictly adhering to the language independence. We are planning to separate pronoun and definite noun classification, with the possibility of using different feature sets. We will also investigate language independent features and implement a markable classifier and a negative instance sampling module.

References

Walter Daelemans and Antal van den Bosch. 2005. *Memory Based Language Processing*. Cambridge University Press.

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.

Sanda M. Harabagiu and Steven J. Maiorano. 2000. Multilingual coreference resolution. In *Proceedings of ANLP 2000*, Seattle, WA.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Morgan Claypool.

Xiaoqiang Luo and Imed Zitouni. 2005. Multilingual coreference resolution with syntactic features. In *Proceedings of HLT/EMNLP 2005*, Vancouver, Canada.

Katja Markert and Malvina Nissim. 2005. Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31(3).

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of ACL/COLING 1998*, Montreal, Canada.

Ruslan Mitkov. 1999. Multilingual anaphora resolution. *Machine Translation*, 14(3-4):281–299.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of ACL 2002*, pages 104–111, Philadelphia, PA.

Massimo Poesio, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Vieira. 2002. Acquiring lexical knowledge for anaphora resolution. In *Proceedings of LREC 2002*, Las Palmas, Gran Canaria.

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

Marta Recasens, Lluís Màrquez, Emili Sapena, M.Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, Uppsala, Sweden.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Holger Wunsch. 2009. *Rule-Based and Memory-Based Pronoun Resolution for German: A Comparison and Assessment of Data Sources*. Ph.D. thesis, Universität Tübingen.