Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution

Simone Paolo Ponzetto and **Michael Strube**
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
http://www.eml-research.de/nlp

Abstract

In this paper we present an extension of a machine learning based coreference resolution system which uses features induced from different semantic knowledge sources. These features represent knowledge mined from WordNet and Wikipedia, as well as information about semantic role labels. We show that semantic features indeed improve the performance on different referring expression types such as pronouns and common nouns.

1 Introduction

The last years have seen a boost of work devoted to the development of machine learning based coreference resolution systems (Soon et al., 2001; Ng & Cardie, 2002; Yang et al., 2003; Luo et al., 2004, inter alia). While machine learning has proved to yield performance rates fully competitive with rule based systems, current coreference resolution systems are mostly relying on rather shallow features, such as the distance between the coreferent expressions, string matching, and linguistic form. However, the literature emphasizes since the very beginning the relevance of world knowledge and inference for coreference resolution (Charniak, 1973).

This paper explores whether coreference resolution can benefit from semantic knowledge sources. More specifically, whether a machine learning based approach to coreference resolution can be improved and *which phenomena* are affected by such information. We investigate the use of the WordNet and Wikipedia taxonomies for extracting *semantic similarity* and *relatedness* measures, as well as semantic parsing information in terms of *semantic role labeling* (Gildea & Jurafsky, 2002, SRL henceforth).

We believe that the lack of semantics in the current systems leads to a performance bottleneck. In order to correctly identify the discourse entities which are referred to in a text, it seems essential to reason over the lexical semantic relations, as well as the event representations embedded in the text. As an example, consider a fragment from the Automatic Content Extraction (ACE) 2003 data.

(1) But frequent visitors say that given the sheer weight of **the country**'s totalitarian ideology and generations of mass indoctrination, changing **this country**'s course will be something akin to turning a huge ship at sea. Opening **North Korea** up, even modestly, and exposing **people** to the idea that Westerners – and South Koreans – are not devils, alone represents an extraordinary change. [...] as **his people** begin to get a clearer idea of the deprivation **they** have suffered, especially relative to **their** neighbors. "**This** is **a society** that has been focused most of all on stability, [...]".

In order to correctly resolve the anaphoric expressions highlighted in bold, it seems that some kind of lexical semantic and encyclopedic knowledge is required. This includes that *North Korea* is a *country*, that *countries* consist of *people* and are *societies*. The resolution requires an encyclopedia (i.e. Wikipedia) look-up and reasoning on the content relatedness holding between the different expressions (i.e. as a path measure along the links of the WordNet and Wikipedia taxonomies). Event representations seem also to be important for coreference resolution, as shown below:

(2) A state commission of inquiry into the sinking of the Kursk will convene in Moscow on Wednesday, **the Interfax news agency** reported. **It** said that the diving operation will be completed by the end of next week.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 192–199,
New York, June 2006.

In this example, knowing that *the Interfax news agency* is the AGENT of the *report* predicate and *It* being the AGENT of *say* could trigger the (semantic parallelism based) inference required to correctly link the two expressions, in contrast to anchoring the pronoun to *Moscow*. SRL provides the semantic relationships that constituents have with predicates, thus allowing us to include such document-level *event descriptive information* into the relations holding between referring expressions (REs).

Instead of exploring different kinds of data representations, task definitions or machine learning techniques (Ng & Cardie, 2002; Yang et al., 2003; Luo et al., 2004) we focus on a few promising semantic features which we evaluate in a controlled environment. That way we try to overcome the plateauing in performance in coreference resolution observed by Kehler et al. (2004).

2 Related Work

Vieira & Poesio (2000), Harabagiu et al. (2001), and Markert & Nissim (2005) explore the use of WordNet for different coreference resolution subtasks, such as resolving bridging reference, *other*- and definite NP anaphora, and MUC-style coreference resolution. All of them present systems which infer coreference relations from a set of potential antecedents by means of a WordNet search. Our approach to WordNet here is to cast the search results in terms of semantic similarity measures. Their output can be used as features for a learner. These measures are not specifically developed for coreference resolution but simply taken 'off-the-shelf' and applied to our task without any specific tuning — i.e. in contrast to Harabagiu et al. (2001), who weight WordNet relations differently in order to compute the confidence measure of the path.

To the best of our knowledge, we do not know of any previous work using Wikipedia or SRL for coreference resolution. In the case of SRL, this layer of semantic context abstracts from the specific lexical expressions used, and therefore represents a higher level of abstraction than (still related) work involving predicate argument statistics. Kehler et al. (2004) observe no significant improvement due to predicate argument statistics. The improvement reported by Yang et al. (2005) is rather caused by their twin-candidate model than by the semantic knowledge. Employing SRL is closer in spirit to Ji et al. (2005), who explore the employment of the ACE 2004 relation ontology as a semantic filter.

3 Coreference Resolution Using Semantic Knowledge Sources

3.1 Corpora Used

To establish a competitive coreference resolver, the system was initially prototyped using the MUC-6 and MUC-7 data sets (Chinchor & Sundheim, 2003; Chinchor, 2001), using the standard partitioning of 30 texts for training and 20-30 texts for testing. Then, we moved on and developed and tested the system with the ACE 2003 Training Data corpus (Mitchell et al., 2003)[1]. Both the Newswire (NWIRE) and Broadcast News (BNEWS) sections where split into 60-20-20% document-based partitions for training, development, and testing, and later per-partition merged (MERGED) for system evaluation. The distribution of coreference chains and referring expressions is given in Table 1.

3.2 Learning Algorithm

For learning coreference decisions, we used a Maximum Entropy (Berger et al., 1996) model. This was implemented using the MALLET library (McCallum, 2002). To prevent the model from overfitting, we employed a tunable Gaussian prior as a smoothing method. The best parameter value is found by searching in the [0,10] interval with step value of 0.5 for the variance parameter yielding the highest MUC score F-measure on the development data.

Coreference resolution is viewed as a binary classification task: given a pair of REs, the classifier has to decide whether they are coreferent or not. The MaxEnt model produces a probability for each category y (coreferent or not) of a candidate pair, conditioned on the context x in which the candidate occurs. The conditional probability is calculated by:

$$p(y|x) = \frac{1}{Z_x}\left[\sum_i \lambda_i f_i(x, y)\right]$$

[1] We used the training data corpus only, as the availability of the test data is restricted to ACE participants. Therefore, the results we report cannot be compared directly with those using the official test data.

	BNEWS (147 docs – 33,479 tokens)				NWIRE (105 docs – 57,205 tokens)			
	#coref ch.	#pron.	#comm. nouns	#prop. names	#coref ch.	#pron.	#comm. nouns	#prop. names
TRAIN.	587	876	572	980	904	1,037	1,210	2,023
DEVEL	201	315	163	465	399	358	485	923
TEST	228	291	238	420	354	329	484	712
TOTAL	1,016	1,482	973	1,865	1,657	1,724	2,179	3,658
TOTAL (%)		34.3%	22.5%	43.2%		22.8%	28.8%	48.4%

Table 1: Partitions of the ACE 2003 training data corpus

where $f_i(x, y)$ is the value of feature i on outcome y in context x, and λ_i is the weight associated with i in the model. Z_x is a normalization constant. The features used in our model are all binary-valued feature functions (or indicator functions), e.g.

$$f_{\text{I\_SEMROLE}}(\text{ARG0/RUN}, \text{COREF}) = \begin{cases} 1 & \text{if candidate pair is coreferent and antecedent is the semantic argument ARG0 of predicate } \textit{run} \\ 0 & \text{else} \end{cases}$$

In our system, a set of pre-processing components including a POS tagger (Giménez & Màrquez, 2004), NP chunker (Kudoh & Matsumoto, 2000) and the *Alias-I LingPipe* Named Entity Recognizer[2] is applied to the text in order to identify the noun phrases, which are further taken as referring expressions (REs) to be used for instance generation. Therefore, we use automatically extracted noun phrases, rather than assuming perfect NP chunking. This is in contrast to other related works in coreference resolution (e.g. Luo et al. (2004), Kehler et al. (2004)).

Instances are created following Soon et al. (2001). We create a positive training instance from each pair of adjacent coreferent REs. Negative instances are obtained by pairing the anaphoric REs with any RE occurring between the anaphor and the antecedent. During testing each text is processed from left to right: each RE is paired with any preceding RE from right to left, until a pair labeled as coreferent is output, or the beginning of the document is reached. The classifier imposes a partitioning on the available REs by clustering each set of expressions labeled as coreferent into the same coreference chain.

3.3 Baseline System Features

Following Ng & Cardie (2002), our baseline system reimplements the Soon et al. (2001) system. The system uses 12 features. Given a potential antecedent RE$_i$ and a potential anaphor RE$_j$ the features are computed as follows[3].

(a) Lexical features

STRING_MATCH T if RE$_i$ and RE$_j$ have the same spelling, else F.

ALIAS T if one RE is an alias of the other; else F.

(b) Grammatical features

I_PRONOUN T if RE$_i$ is a pronoun; else F.

J_PRONOUN T if RE$_j$ is a pronoun; else F.

J_DEF T if RE$_j$ starts with *the*; else F.

J_DEM T if RE$_j$ starts with *this*, *that*, *these*, or *those*; else F.

NUMBER T if both RE$_i$ and RE$_j$ agree in number; else F.

GENDER U if either RE$_i$ or RE$_j$ have an undefined gender. Else if they are both defined and agree T; else F.

PROPER_NAME T if both RE$_i$ and RE$_j$ are proper names; else F.

APPOSITIVE T if RE$_j$ is in apposition with RE$_i$; else F.

(c) Semantic features

WN_CLASS U if either RE$_i$ or RE$_j$ have an undefined WordNet semantic class. Else if they both have a defined one and it is the same T; else F.

(d) Distance features

DISTANCE how many sentences RE$_i$ and RE$_j$ are apart.

[2] `http://alias-i.com/lingpipe`

[3] Possible values are U(nknown), T(rue) and F(alse). Note that in contrast to Ng & Cardie (2002) we interpret ALIAS as a lexical feature, as it solely relies on string comparison and acronym string matching.

3.4 WordNet Features

In the baseline system semantic information is limited to WordNet semantic class matching. Unfortunately, a WordNet semantic class lookup exhibits problems such as coverage, sense proliferation and ambiguity[4], which make the WN_CLASS feature very noisy. We enrich the semantic information available to the classifier by using semantic similarity measures based on the WordNet taxonomy (Pedersen et al., 2004). The measures we use include path length based measures (Rada et al., 1989; Wu & Palmer, 1994; Leacock & Chodorow, 1998), as well as ones based on information content (Resnik, 1995; Jiang & Conrath, 1997; Lin, 1998).

In our case, the measures are obtained by computing the similarity scores between the head lemmata of each potential antecedent-anaphor pair. In order to overcome the sense disambiguation problem, we factorise over all possible sense pairs: given a candidate pair, we take the cross product of each antecedent and anaphor sense to form pairs of synsets. For each measure WN_SIMILARITY, we compute the similarity score for all synset pairs, and create the following features.

WN_SIMILARITY_BEST the *highest* similarity score from all $\langle \text{SENSE}_{RE_i,n}, \text{SENSE}_{RE_j,m} \rangle$ synset pairs.

WN_SIMILARITY_AVG the *average* similarity score from all $\langle \text{SENSE}_{RE_i,n}, \text{SENSE}_{RE_j,m} \rangle$ synset pairs.

Pairs containing REs which cannot be mapped to WordNet synsets are assumed to have a null similarity measure.

3.5 Wikipedia Features

Wikipedia is a multilingual Web-based free-content encyclopedia[5]. The English version, as of 14 February 2006, contains 971,518 articles with 16.8 million internal hyperlinks thus providing a large coverage available knowledge resource. In addition, since May 2004 it provides also a taxonomy by means of the *category feature*: articles can be placed in one or more categories, which are further categorized to provide a category tree. In practice, the taxonomy is not designed as a strict hierarchy or tree of categories, but allows multiple categorisation schemes to co-exist simultaneously. Because each article can appear in more than one category, and each category can appear in more than one parent category, the categories do not form a tree structure, but a more general directed graph. As of December 2005, 78% of the articles have been categorized into 87,000 different categories.

Wikipedia mining works as follows (for an in-depth description of the methods for computing semantic relatedness in Wikipedia see Strube & Ponzetto (2006)): given the candidate referring expressions RE$_i$ and RE$_j$ we first pull the pages they refer to. This is accomplished by querying the page titled as the head lemma or, in the case of NEs, the full NP. We follow all redirects and check for disambiguation pages, i.e. pages for ambiguous entries which contain links only (e.g. *Lincoln*). If a disambiguation page is hit, we first get all the hyperlinks in the page. If a link containing the other queried RE is found (i.e. a link containing *president* in the *Lincoln* page), the linked page (*President of the United States*) is returned, else we return the first article linked in the disambiguation page. Given a candidate coreference pair RE$_{i/j}$ and the Wikipedia pages $P_{RE_{i/j}}$ they point to, obtained by querying pages titled as $T_{RE_{i/j}}$, we extract the following features:

I/J_GLOSS_CONTAINS U if no Wikipedia page titled $T_{RE_{i/j}}$ is available. Else T if the first paragraph of text of $P_{RE_{i/j}}$ contains $T_{RE_{j/i}}$; else F.

I/J_RELATED_CONTAINS U if no Wikipedia page titled as $T_{RE_{i/j}}$ is available. Else T if at least one Wikipedia hyperlink of $P_{RE_{i/j}}$ contains $T_{RE_{j/i}}$; else F.

I/J_CATEGORIES_CONTAINS U if no Wikipedia page titled as $T_{RE_{i/j}}$ is available. Else T if the list of categories $P_{RE_{i/j}}$ belongs to contains $T_{RE_{j/i}}$; else F.

GLOSS_OVERLAP the overlap score between the first paragraph of text of P_{RE_i} and P_{RE_j}. Following Banerjee & Pedersen (2003) we compute the score as $\sum_n m^2$ for n phrasal m-word overlaps.

[4] Following the system to be replicated, we simply mapped each RE to the first WordNet sense of the head noun.

[5] Wikipedia can be downloaded at `http://download.wikimedia.org/`. In our experiments we use the English Wikipedia database dump from 19 February 2006.

Additionally, we use the Wikipedia category graph. We ported the WordNet similarity path length based measures to the Wikipedia category graph. However, the category relations in Wikipedia cannot only be interpreted as corresponding to *is-a* links in a taxonomy since they denote meronymic relations as well. Therefore, the Wikipedia-based measures are to be taken as semantic relatedness measures. The measures from Rada et al. (1989), Leacock & Chodorow (1998) and Wu & Palmer (1994) are computed in the same way as for WordNet. Path search takes place as a depth-limited search of maximum depth of 4 for a least common subsumer. We noticed that limiting the search improves the results as it yields a better correlation of the relatedness scores with human judgements (Strube & Ponzetto, 2006). This is due to the high regions of the Wikipedia category tree being too strongly connected.

In addition, we use the measure from Resnik (1995), which is computed using an intrinsic information content measure relying on the hierarchical structure of the category tree (Seco et al., 2004). Given $P_{RE_{i/j}}$ and the lists of categories $C_{RE_{i/j}}$ they belong to, we factorise over all possible category pairs. That is, we take the cross product of each antecedent and anaphor category to form pairs of 'Wikipedia synsets'. For each measure WIKI_RELATEDNESS, we compute the relatedness score for all category pairs, and create the following features.

WIKI_RELATEDNESS_BEST the *highest* relatedness score from all $\langle C_{RE_i,n}, C_{RE_j,m} \rangle$ category pairs.

WIKI_RELATEDNESS_AVG the *average* relatedness score from all $\langle C_{RE_i,n}, C_{RE_j,m} \rangle$ category pairs.

3.6 Semantic Role Features

The last semantic knowledge enhancement for the baseline system uses SRL information. In our experiments we use the ASSERT parser (Pradhan et al., 2004), an SVM based semantic role tagger which uses a full syntactic analysis to automatically identify all verb predicates in a sentence together with their semantic arguments, which are output as PropBank arguments (Palmer et al., 2005). It is often the case that the semantic arguments output by the parser do not align with any of the previously identified noun phrases. In this case, we pass a semantic role label to a RE only when the two phrases share the same head. Labels have the form "$\text{ARG}_1\text{\_pred}_1 \ldots \text{ARG}_n\text{\_pred}_n$" for n semantic roles filled by a constituent, where each semantic argument label is always defined with respect to a predicate. Given such level of semantic information available at the RE level, we introduce two new features[6].

I_SEMROLE the semantic role argument-predicate pairs of RE_i.

J_SEMROLE the semantic role argument-predicate pairs of RE_j.

For the ACE 2003 data, 11,406 of 32,502 automatically extracted noun phrases were tagged with 2,801 different argument-predicate pairs.

4 Experiments

4.1 Performance Metrics

We report in the following tables the MUC score (Vilain et al., 1995). Scores in Table 2 are computed for all noun phrases appearing in either the key or the system response, whereas Tables 3 and 4 refer to scoring only those phrases which appear in both the key and the response. We therefore discard those responses not present in the key, as we are interested in establishing the upper limit of the improvements given by our semantic features. That is, we want to define a baseline against which to establish the contribution of the semantic information sources explored here for coreference resolution.

In addition, we report the accuracy score for all three types of ACE mentions, namely pronouns, common nouns and proper names. Accuracy is the percentage of REs of a given mention type correctly resolved divided by the total number of REs of the same type given in the key. A RE is said to be correctly resolved when both it and its direct antecedent are placed by the key in the same coreference class.

[6] During prototyping we experimented unpairing the arguments from the predicates, which yielded worse results. This is supported by the PropBank arguments always being defined with respect to a target predicate. Binarizing the features — i.e. do RE_i and RE_j have the same argument or predicate label with respect to their closest predicate? — also gave worse results.

	MUC-6			MUC-7		
original	R	P	F_1	R	P	F_1
Soon et al.	58.6	67.3	62.3	56.1	65.5	60.4
duplicated baseline	64.9	65.6	65.3	55.1	68.5	61.1

Table 2: Results on MUC

4.2 Feature Selection

For determining the relevant feature sets we follow an iterative procedure similar to the wrapper approach for feature selection (Kohavi & John, 1997) using the development data. The feature subset selection algorithm performs a hill-climbing search along the feature space. We start with a model based on all available features. Then we train models obtained by removing one feature at a time. We choose the worst performing feature, namely the one whose removal gives the largest improvement based on the MUC score F-measure, and remove it from the model. We then train classifiers removing each of the remaining features separately from the enhanced model. The process is iteratively run as long as significant improvement is observed.

4.3 Results

Table 2 compares the results between our duplicated Soon baseline and the original system. We assume that the slight improvements of our system are due to the use of current pre-processing components and another classifier. Tables 3 and 4 show a comparison of the performance between our baseline system and the ones incremented with semantic features. Performance improvements are highlighted in bold[7].

4.4 Discussion

The tables show that *semantic features improve system recall*, rather than acting as a ‘semantic filter’ improving precision. Semantics therefore seems to trigger a response in cases where more shallow features do not seem to suffice (see examples (1-2)).

Different feature sources account for different RE type improvements. WordNet and Wikipedia features tend to increase performance on common nouns, whereas SRL improves pronouns. WordNet features are able to improve by 14.3% and 7.7% the accuracy rate for common nouns on the BNEWS and NWIRE datasets (+34 and +37 correctly resolved common nouns out of 238 and 484 respectively), whereas employing Wikipedia yields slightly smaller improvements (+13.0% and +6.6% accuracy increase on the same datasets). Similarly, when SRL features are added to the baseline system, we register an increase in the accuracy rate for pronouns, ranging from 0.7% in BNEWS and NWIRE up to 4.2% in the MERGED dataset (+26 correctly resolved pronouns out of 620).

If semantics helps for pronouns and common nouns, it does not affect performance on proper names, where features such as string matching and alias suffice. This suggests that semantics plays a role in pronoun and common noun resolution, where surface features cannot account for complex preferences and semantic knowledge is required.

The best accuracy improvement on pronoun resolution is obtained on the MERGED dataset. This is due to making more data available to the classifier, as the SRL features are very sparse and inherently suffer from data fragmentation. Using a larger dataset highlights the importance of SRL, whose features are never removed in any feature selection process[8]. The accuracy on common nouns shows that features induced from Wikipedia are competitive with the ones from WordNet. The performance gap on all three datasets is quite small, which indicates the usefulness of using an encyclopedic knowledge base as a replacement for a lexical taxonomy.

As a consequence of having different knowledge sources accounting for the resolution of different RE types, the best results are obtained by (1) *combining features* generated *from different sources*; (2) *performing feature selection*. When combining different feature sources, we register an accuracy improvement on pronouns and common nouns, as well as an increase in F-measure due to a higher recall.

Feature selection always improves results. This is due to the fact that our full feature set is ex-

[7] All changes in F-measure are statistically significant at the 0.05 level or higher. We follow Soon et al. (2001) in performing a simple one-tailed, paired sample t-test between the baseline system’s MUC score F-measure and each of the other systems’ F-measure scores on the test documents.

[8] To our knowledge, most of the recent work in coreference resolution on the ACE data keeps the document source separated for evaluation. However, we believe that document source independent evaluation provides useful insights on the robustness of the system (cf. the CoNLL 2005 shared task cross-corpora evaluation).

	BNEWS						NWIRE					
	R	P	F_1	A_p	A_{cn}	A_{pn}	R	P	F_1	A_p	A_{cn}	A_{pn}
baseline	46.7	86.2	60.6	36.4	10.5	44.0	56.7	88.2	69.0	37.6	23.1	55.6
+WordNet	**54.8**	86.1	**66.9**	**36.8**	**24.8**	**47.6**	**61.3**	84.9	**71.2**	**38.9**	**30.8**	55.5
+Wiki	**52.7**	**86.8**	**65.6**	36.1	**23.5**	**46.2**	**60.6**	83.6	**70.3**	**38.0**	**29.7**	55.2
+SRL	**53.3**	85.1	**65.5**	**37.1**	**13.9**	**46.2**	**58.0**	**89.0**	**70.2**	**38.3**	**25.0**	**56.0**
all features	**59.1**	84.4	**69.5**	**37.5**	**27.3**	**48.1**	**63.1**	83.0	**71.7**	**39.8**	**31.8**	52.8

Table 3: Results on the ACE 2003 data (BNEWS and NWIRE sections)

	R	P	F_1	A_p	A_{cn}	A_{pn}
baseline	54.5	88.0	67.3	34.7	20.4	53.1
+WordNet	**56.7**	87.1	**68.6**	**35.6**	**28.5**	49.6
+Wikipedia	**55.8**	87.5	**68.1**	**34.8**	**26.0**	50.5
+SRL	**56.3**	**88.4**	**68.8**	**38.9**	**21.6**	51.7
all features	**61.0**	84.2	**70.7**	**38.9**	**29.9**	51.2

Table 4: Results ACE (merged BNEWS/NWIRE)

Feature set	F_1
baseline (Soon w/o DISTANCE)	58.4%
+WIKI_WU_PALMER_BEST	+4.3%
+J_SEMROLE	+1.8%
+WIKI_PATH_AVG	+1.2%
+I_SEMROLE	+0.8%
+WN_WU_PALMER_BEST	+0.7%

Table 5: Feature selection (BNEWS section)

tremely redundant: in order to explore the usefulness of the knowledge sources we included overlapping features (i.e. using *best* and *average* similarity/relatedness measures at the same time), as well as features capturing the same phenomenon from different point of views (i.e. using *multiple* measures at the same time). In order to yield the desired performance improvements, it turns out to be essential to filter out irrelevant features.

Table 5 shows the relevance of the best performing features on the BNEWS section. As our feature selection mechanism chooses the best set of features by removing them (see Section 4.2), we evaluate the contributions of the remaining features as follows. We start with a baseline system using all the features from Soon et al. (2001) that were not removed in the feature selection process (i.e. DISTANCE). We then train classifiers combining the current feature set with each feature in turn. We then choose the best performing feature based on the MUC score F-measure and add it to the model. We iterate the process until all features are added to the baseline system. The table indicates that all knowledge sources are relevant for coreference resolution, as it includes SRL, WordNet and Wikipedia features. The Wikipedia features rank high, indicating again that it provides a valid knowledge base.

5 Conclusions and Future Work

The results are somehow surprising, as one would not expect a community-generated categorization to be almost as informative as a well structured lexical taxonomy such as WordNet. Nevertheless Wikipedia offers promising results, which we expect to improve as well as the encyclopedia goes under further development.

In this paper we investigated the effects of using different semantic knowledge sources within a machine learning based coreference resolution system. This involved mining the WordNet taxonomy and the Wikipedia encyclopedic knowledge base, as well as including semantic parsing information, in order to induce semantic features for coreference learning. Empirical results show that coreference resolution benefits from semantics. The generated model is able to learn selectional preferences in cases where surface morpho-syntactic features do not suffice, i.e. pronoun and common name resolution. While the results given by using 'the free encyclopedia that anyone can edit' are satisfactory, major improvements can come from developing efficient query strategies – i.e. a more refined disambiguation technique taking advantage of the context in which the queries (e.g. referring expressions) occur.

Future work will include turning Wikipedia into an ontology with well defined taxonomic relations, as well as exploring its usefulness of for other NLP applications. We believe that an interesting aspect of Wikipedia is that it offers large coverage resources for many languages, thus making it a natural choice for multilingual NLP systems.

Semantics plays indeed a role in coreference resolution. But semantic features are expensive to compute and the development of efficient methods is required to embed them into large scale systems. Nevertheless, we believe that exploiting semantic knowledge in the manner we described will assist the research on coreference resolution to overcome the plateauing in performance observed by Kehler et al. (2004).

Acknowledgements: This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a KTF grant (09.003.2004). We thank Katja Filippova, Margot Mieskes and the three anonymous reviewers for their useful comments.

References

Banerjee, S. & T. Pedersen (2003). Extended gloss overlap as a measure of semantic relatedness. In *Proc. of IJCAI-03*, pp. 805–810.

Berger, A., S. A. Della Pietra & V. J. Della Pietra (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Charniak, E. (1973). Jack and Janet in search of a theory of knowledge. In *Advance Papers from the Third International Joint Conference on Artificial Intelligence, Stanford, Cal.*, pp. 337–343.

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

Chinchor, N. & B. Sundheim (2003). *Message Understanding Conference (MUC) 6.* LDC2003T13, Philadelphia, Penn: Linguistic Data Consortium.

Gildea, D. & D. Jurafsky (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Giménez, J. & L. Màrquez (2004). SVMTool: A general POS tagger generator based on support vector machines. In *Proc. of LREC '04*, pp. 43–46.

Harabagiu, S. M., R. C. Bunescu & S. J. Maiorano (2001). Text and knowledge mining for coreference resolution. In *Proc. of NAACL-01*, pp. 55–62.

Ji, H., D. Westbrook & R. Grishman (2005). Using semantic relations to refine coreference decisions. In *Proc. HLT-EMNLP '05*, pp. 17–24.

Jiang, J. J. & D. W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING)*.

Kehler, A., D. Appelt, L. Taylor & A. Simma (2004). The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proc. of HLT-NAACL-04*, pp. 289–296.

Kohavi, R. & G. H. John (1997). Wrappers for feature subset selection. *Artificial Intelligence Journal*, 97(1-2):273–324.

Kudoh, T. & Y. Matsumoto (2000). Use of Support Vector Machines for chunk identification. In *Proc. of CoNLL-00*, pp. 142–144.

Leacock, C. & M. Chodorow (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet. An Electronic Lexical Database*, Chp. 11, pp. 265–283. Cambridge, Mass.: MIT Press.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pp. 296–304.

Luo, X., A. Ittycheriah, H. Jing, N. Kambhatla & S. Roukos (2004). A mention-synchronous coreference resolution algorithm based on the Bell Tree. In *Proc. of ACL-04*, pp. 136–143.

Markert, K. & M. Nissim (2005). Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31(3):367–401.

McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit.*

Mitchell, A., S. Strassel, M. Przybocki, J. Davis, G. Doddington, R. Grishman, A. Meyers, A. Brunstain, L. Ferro & B. Sundheim (2003). *TIDES Extraction (ACE) 2003 Multilingual Training Data.* LDC2004T09, Philadelphia, Penn.: Linguistic Data Consortium.

Ng, V. & C. Cardie (2002). Improving machine learning approaches to coreference resolution. In *Proc. of ACL-02*, pp. 104–111.

Palmer, M., D. Gildea & P. Kingsbury (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Pedersen, T., S. Patwardhan & J. Michelizzi (2004). WordNet::Similarity – Measuring the relatedness of concepts. In *Companion Volume of the Proceedings of the Human Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 267–270.

Pradhan, S., W. Ward, K. Hacioglu, J. H. Martin & D. Jurafsky (2004). Shallow semantic parsing using Support Vector Machines. In *Proc. of HLT-NAACL-04*, pp. 233–240.

Rada, R., H. Mili, E. Bicknell & M. Blettner (1989). Development and application of a metric to semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of IJCAI-95*, Vol. 1, pp. 448–453.

Seco, N., T. Veale & J. Hayes (2004). An intrinsic information content metric for semantic similarity in WordNet. In *Proc. of ECAI-04*, pp. 1089–1090.

Soon, W. M., H. T. Ng & D. C. Y. Lim (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Strube, M. & S. P. Ponzetto (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proc. of AAAI-06*.

Vieira, R. & M. Poesio (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Vilain, M., J. Burger, J. Aberdeen, D. Connolly & L. Hirschman (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pp. 45–52.

Wu, Z. & M. Palmer (1994). Verb semantics and lexical selection. In *Proc. of ACL-94*, pp. 133–138.

Yang, X., J. Su & C. L. Tan (2005). Improving pronoun resolution using statistics-based semantic compatibility information. In *Proc. of ACL-05*, pp. 165–172.

Yang, X., G. Zhou, J. Su & C. L. Tan (2003). Coreference resolution using competition learning approach. In *Proc. of ACL-03*, pp. 176–183.