Large Corpus-based Semantic Feature Extraction for Pronoun Coreference

Shasha Liao
Dept. of Computer Science
New York University
liaoss@cs.nyu.edu

Ralph Grishman
Dept. of Computer Science
New York University
grishman@cs.nyu.edu

Abstract

Semantic information is a very important factor in coreference resolution. The combination of large corpora and 'deep' analysis procedures has made it possible to acquire a range of semantic information and apply it to this task. In this paper, we generate two statistically-based semantic features from a large corpus and measure their influence on pronoun coreference. One is contextual compatibility, which decides if the antecedent can be used in the anaphor's context; the other is role pair, which decides if the actions asserted of the antecedent and the anaphor are likely to apply to the same entity. We apply a semantic labeling system and a baseline coreference system to a large corpus to generate semantic patterns and convert them into features in a MaxEnt model. These features produce an absolute gain of 1.5% to 1.7% in resolution accuracy (a 6% reduction in errors). To understand the limitations of these features, we also extract patterns from the test corpus, use these patterns to train a coreference model, and examine some of the cases where coreference still fails. We also compare the performance of patterns extracted from semantic role labeling and syntax.

1 Introduction

Coreference resolution is the task of determining whether two phrases refer to the same entity. Coreference is critical to most NLP tasks, yet even the sub-problem of pronoun coreference remains very challenging. In principle, we need several types of information to identify the right antecedent. First, number and gender agreement constraints can narrow the candidate set. If multiple candidates remain, we would next use some sequence or syntactic features, like position, word, word salience and discourse focus. For example, whether an antecedent is in subject position might be helpful because the subject is more likely to be referred to; or an entity that has been referred to repeatedly is more likely to be referred to again. However, these features do not suffice to pick the correct antecedent, and sometimes similar syntactic structures might have quite different coreference solutions. For example, for the following two sentences:

(1) *The terrorist shot a 13-year-old boy;* ***he*** *was arrested after the attack.*

(2) *The terrorist shot a 13-year-old boy;* ***he*** *was fatally wounded in the attack.*

it is likely that "he" refers to "*terrorist*" in (1) and "*boy*" in (2). However, we cannot get the right antecedent using the features we mentioned above because the examples share the same antecedent words and syntactic structure. People can still resolve these correctly because "*terrorist*" is more likely to be arrested than "*boy*", and because the one shooting is more likely to be arrested than the one being shot.

In such cases, semantic constraints and preferences are required for correct coreference resolution. Methods for acquiring and using such knowledge are receiving increasing attention in

Proceedings of the Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010), pages 60–68,
Beijing, August 2010

recent work on anaphora resolution. Dagan and Itai (1990), Bean and Riloff (2004), Yang and Su (2007), and Ponzetto and Strube (2006) all explored this task.

However, this task is difficult because it requires the acquisition of a large amount of semantic information. Furthermore, there is not universal agreement on the value of these semantic preferences for pronoun coreference. Kehler et al. (2004) reported that such information did not produce apparent improvement in overall pronoun resolution.

In this paper, we will extract semantic features from a semantic role labeling system instead of a parse tree, and explore whether pronoun coreference resolution can benefit from such knowledge, which is automatically extracted from a large corpus. We studied two features: the contextual compatibility feature which has been demonstrated to work at the syntactic level by previous work; and the role pair feature, which has not previously been applied to general domain pronoun co-reference. In addition, to obtain a rough upper bound on the benefits of our approach and understand its limitations, we conducted a second experiment in which the semantic knowledge is extracted from the evaluation corpus.

We will use the term *mention* to describe an individual referring phrase. For most studies of coreference, mentions are noun phrases and may be headed by a name, a common noun, or a pronoun. We will use the term *entity* to refer to a set of coreferential mentions.

2 Related Work

Contextual compatibility features have long been studied for pronoun coreference: Dagan and Itai (1990) proposed a heuristics-based approach to pronoun resolution. It determined the preference of candidates based on predicate-argument frequencies.

Bean and Riloff (2004) present a system, which uses contextual role knowledge to aid coreference resolution. They used lexical and syntactic heuristics to identify high-confidence coreference relations and used them as training data for learning contextual role knowledge. They got substantial gains on articles in two specific domains, terrorism and natural disasters.

Yang et al. (2005) use statistically-based semantic compatibility information to improve pronoun resolution. They use corpus-based and web-based extraction strategies, and their work shows that statistically-based semantic compatibility information can improve coreference resolution.

In contrast, Kehler et al. (2004) claimed that the contextual compatibility feature does not help much for pronoun coreference: existing learning-based approaches already performed well; such statistics are simply not good predictors for pronoun interpretation; data is sparse in the collected predicate-argument statistics.

The role pair feature has not been studied for general, broad-domain pronoun co-reference, but it has been used for other tasks: Pekar (2006) built pairs of 'templates' which share an 'anchor' argument; these correspond closely to our role pairs. Association statistics of the template pairs were used to acquire verb entailments. Abe et al. (2008) looked for pairs appearing in specific syntactic patterns in order to acquire finer-grained event relations. Chambers and Jurafsky (2008) built narrative event chains, which are partially ordered sets of events related by a common protagonist. They use high-precision hand-coded rules to get coreference information, extract predicate arguments that link the mentions to verbs, and link the arguments of the coreferred mentions to build a verb entailment model.

Bean and Riloff (2004) used high-precision hand-coded rules to identify coreferent mention pairs, which are then used to acquire role pairs that they refer to as *Caseframe Network* features. They use these features to improve coreference resolution for two domain-specific corpora involving terrorism and natural disasters. Their result raises the natural question as to whether the approach (which may capture domain-specific pairs such as “kidnap—release” in the terrorism domain) can be successfully extended to a general news corpus. We address this question in the experiments reported here.

3 Corpus Analysis

In order to extract semantic features from our large training corpus, we apply a sequence of analyzers. These include name tagging, parsing, a baseline coreference analyzer, and, most important, a semantic labeling system that can generate the logical grammatical and predicate-argument representation automatically from a

parse tree (Meyers et al. 2009). We use semantic labeling because it provides more general and meaningful patterns, with a "deeper" analysis than parsed text. The output of the semantic labeling is the dependency representation of the text, where each sentence is a graph consisting of nodes (corresponding to words) and arcs. Each arc captures up to three relations between two words: (1) a SURFACE relation, the relation between a predicate and an argument in the parse of a sentence; (2) a LOGIC1 (grammatical logical) relation which regularizes for lexical and syntactic phenomena like passive, relative clauses, and deleted subjects; and (3) a LOGIC2 (predicate-argument) relation corresponding to relations in PropBank and NomBank. It is designed to be compatible with the Penn TreeBank (Marcus et al., 1994) framework and therefore, Penn TreeBank-based parsers, while incorporating Named Entities, PropBank, and NomBank.

Because nouns and verbs provide the most relevant contexts and capture the events in which the entities participate, we generate *semantic patterns* (triples) only for those arcs with verb or noun heads. We use the following relations:

- Logic2 relations: We use in particular the Arg0 relation (which corresponds roughly to *agent*) and Arg1 relation (which corresponds roughly to *patient*).
- Logic1 relations: We use in particular the Sbj and Obj relations, representing the logical subject and object of a verb (regularizing passive, relative clauses, deleted subjects)
- Surface relations: T-pos relation is particularly used, which captures the head noun – determiner relation for possessive constructs such as "bomber's attack" and "his responsibility".

For example, for the sentence:

`John is hit by Tom's brother.`

we generate the semantic patterns

<Arg1 hit John>
<Arg0 hit brother>
<T-pos brother Tom>

We apply this labeling system to all the data we use, and to generate the semantic pattern, we take first its predicate-argument role; if that is null, we take its logical grammatical role; if both are null, we take its surface role.

To reduce data sparseness, all inflected words are changed to their base form (e.g. "attackers"→"attacker"). All names are replaced by their ACE types (person, organization, location, etc.). Only patterns with noun arguments are extracted because we only consider noun phrases as possible antecedents.

4 Semantic Features

4.1 Contextual Compatibility Patterns

Pronouns, especially neutral pronouns ("it", "they"), carry little semantics of their own, so examining the compatibility of the context of a pronoun and its candidate antecedents is a good way to improve antecedent selection. Specifically, we want to determine whether the predicate, which is applied to the anaphor, can be applied to the antecedents. We take the semantic pattern with the anaphor in third position. Then, each candidate antecedent is substituted for the anaphor to see if it is suitable for the context. For example, consider the sentence

The company issued a statement that it bought G.M.

which would generate the semantic patterns

<Arg0 issue company>
<Arg1 issue statement>
<Arg0 buy it>
<Arg1 buy Organization>

(here "G.M" is a name of type *organization* and so is replaced by the token Organization). The relevant context of the anaphor is the semantic pattern <Arg0 buy it>. Suppose there are two candidate antecedents for "it": "company" and "statement". We would generate the two semantic patterns <Arg0 buy company> and <Arg0 buy statement>. Assuming <Arg0 buy company> is more highly ranked than <Arg0 buy statement>, we can infer that the anaphor is more likely to refer to "company". (We describe the specific metric we use for ranking below, in section 4.3.) As further examples consider:

(3) *The suspect's lawyer, Chifumu Banda, told the court **he** had advised Chiluba not to appear in court Friday.*

(4) *Foreign military analysts said it would be highly unusual for an accident to kill a whole submarine crew and* ***they*** *suggested possible causes to a disaster…*

For (3), if we know that a lawyer is more likely to give advice than a suspect, we could link "*he*" to "*lawyer*" instead of "*suspect*" in the first sentence. For (4), if we know that analysts are more likely to "*suggest*" than crew, we can link "*they*" to "analysts" in the second sentence.

4.2 Role Pair Patterns

The role pair pattern is a new feature in general pronoun co-reference. The original intuition for introducing it into coreference is that there are pairs of actions involving the same entity that are much more likely to occur together than would be true if one assumed statistical independence. The second action may be a rephrasing or elaboration of the first, or the two might be actions that are part of a common 'script'. For example:

(5) *Prime Minister Mahathir Mohamad sacked the former deputy premier in 1998, who was sentenced to a total of 15 years in jail after being convicted of corruption and sodomy.* ***He*** *was released after four years because….*

(6) *The robber attacked the boy with a knife;* ***he*** *was bleeding heavily and died in the hospital the next day.*

For (5), if we know that the person who was sentenced is more likely to be released than the person who sacked others, we would know "*he*" refers to "*deputy premier*" instead of "*prime minister*". And in (6), because someone being attacked is more likely to die than the attacker, we can infer that "*he*" refers to "*boy*".

To acquire such information, we need to identify those pairs of predicates which are likely to apply to the same entity. We collect this data from a large corpus. The basic process is: apply a baseline coreference system to produce mentions and entities for a large corpus. For every entity, record the predicates for every mention, and then the pairs of predicates for successive mentions within each entity.

Although the performance of the baseline coreference is not very high, and individual documents may yield many idiosyncratic pairs, we can gather many significant role pairs by collecting statistics from a large corpus and filtering out the low frequency patterns; this process can eliminate much of the noise due to coreference errors.

Here is an example of the extracted role pairs involving "attack":

Arg0 attack *x* ↔	Obj volley *x*
	Arg0 bombard *x*
	Obj barrage *x*
	Arg0 snatch *x*
	Sbj attack *x*
	Arg0 pound *x*
	Obj reoccupy *x*
	Arg1 halt *x*
	Arg0 assault *x*
	Arg1 bombard *x*

Table1. Top 10 role pairs associated with "Arg0 attack x"

4.3 Contextual Compatibility Scores

To properly compare the patterns involving alternative candidate antecedents, we need to normalize the raw frequencies first. We followed Yang et al. (2005)'s idea, which normalizes the pattern frequency by the frequency of the candidates, and use a relative score that is normalized by the maximum score of all its candidates:

$$CompScore(P_{context,Cand}) = \frac{CompFreq(P_{context,Cand})}{Max_{Ci \in Set(cands)} CompFreq(P_{context,Ci})}$$

and

$$CompFreq(P_{context,Cand}) = \frac{freq(P_{context,Cand})}{freq(Cand)}$$

where $P_{context,Cand}$ is the contextual compatibility pattern built from the context of the pronoun and the base form of the candidate.

In contrast to Yang's work, which used contextual compatibility on the *mention* level, we consider the contextual compatibility of an *entity* to an anaphor: we calculate the contextual information of all the mentions and choose the one with highest score as the contextual compatibility score for this entity[1]:

[1] Note that all the mentions in the entity are generated by the overall coreference system. Also, the ACE entity type of names is determined by the system. No key annotations are considered in the entire coreference phase.

$$freq(context,entity) = Max_{mention_i \in entity} freq(P_{context,mention_i})$$

4.4 Role Pair Scores

Unlike the contextual compatibility feature, we only take the role pair of the successive mentions in the candidate entity and the anaphor, because they are more reliably coreferential than arbitrary pairs of mentions within an entity:

$$PairFreq(p_{ana}, p_{cand}) = \frac{freq(p_{ana}, p_{cand})}{freq(p_{cand})}$$

where p_{ana} and p_{cand} are the contextual patterns of the anaphor and the last mention in the candidate entity.

For a set of possible candidates, we compute a relative score:

$$PairScore(p_{ana},p_{cand}) = \frac{PairFreq(p_{ana},p_{cand})}{Max_{pi \in Set(cands)} PairFreq(p_{ana},pi)}$$

Both scores are quantized (binned) in intervals of 0.1 for use as MaxEnt features.

5 Experiment

Our coreference solution system uses ACE annotated data and follows the ACE 2005 English entity guidelines.[2] The baseline coreference system to compare with is the same one used for extracting semantic features from the large corpus. It employs an entity-mention (rather than a mention-pair) model.

Besides entity and mention information, which (as mentioned above) is system output, the semantic information is also automatically extracted by a semantic labeling system. As a result, we report results in section 5.4 which involve no information from the reference (key) annotation.

5.1 Baseline System Description

The baseline system first applies processes like parsing, semantic labeling, name tagging, and entity mention tagging, producing a set of mentions to which coreference analysis is then applied. The coreference phase deals with coreference among mentions that might be pronouns, names or proper nouns, and generates entities when it is finished. The whole is a one-pass process, resolving coreference in the order in which mentions appear in the document. In the pronoun coreference process, every pronoun mention is assigned to one of the candidate entities.

Features	Description
Hobbs_ Distance	Hobbs distance between the last mention in the entity and the anaphor
Head_Pro	Combined word features of the head of the last mention in the entity and anaphor
Is_Subject	True if the last mention in the entity is a subject of the sentence
Last_Cat	Whether the last mention in the entity is a noun phrase, a pronoun or a name
Co_Prior	Number of prior references to this entity

Table 2. Features used in baseline system

The baseline co-reference system has separate, quite elaborate, primarily rule-based systems to handle names, nominals, headless NP's, and adverbs ("here", "there") which may be anaphoric, as well as first- and second-person pronouns. The MaxEnt model under study in this paper is only responsible for third-person pronouns. Also, gender, number, and human/non-human are handled separately outside of the MaxEnt model, and the model only resolves mentions that satisfy these constraints.[3] In the MaxEnt model, 5 basic features (described in table 2) are used. Thus, while the set of features used in the model is relatively small in comparison to many current statistically based reference resolvers, these are the primary features relevant to the limited task

[2] Automatic Content Extraction evaluation, http://projects.ldc.upenn.edu/ace/

[3] Gender information is obtained from a dictionary of gender-specific nouns and from first-name lists from the US Census. Number information comes from large syntactic dictionaries, corpus annotation of collective nouns (syntactically singular nouns which may take plural anaphors), and name tagger information (some organizations and political entities may take plural anaphors).

of the MaxEnt model, and its performance is still competitive[4].

5.2 Corpus Description

There are two kinds of corpora used in our experiment, a small coreference-annotated corpus used for training and evaluating (in cross-validation) the pronoun coreference model, and a large raw-text corpus for extracting semantic information.

For model training and evaluation, we assembled two small corpora from the available ACE data. One consists of news articles (460 documents) from ACE 2005 (330 documents) and ACE 2003 (130 documents), which together contain 3934 pronouns. The other is the full ACE 2005 training set (592 documents), which includes newswire, broadcast news, broadcast conversations (interviews and discussions), web logs, web forums, and Fisher telephone transcripts, and contains 5659 pronouns.

In evaluation, we consider a pronoun to be correctly resolved if its antecedent in the system output (the most recent prior mention in the entity to which the pronoun is assigned) matches the antecedent in the key. We report accuracy (percentage of pronouns which are correctly resolved).

We used a large corpus to extract semantic information, consisting of five years of AFP newswire from the LDC English Gigaword corpus (1996, 2002, 2004, 2005 and 2006), a total of 907,368 documents. We omit news articles written in 1998, 2000 and 2003 to insure there is no overlap between the ACE data and Gigaword data. We pre-processed each document (parsing, name identification, and semantic labeling) and ran the baseline coreference system, which automatically identified all the mentions (including name mentions and nominal mentions) and built a set of entities for each document.

[4]For example, among papers reporting a pronoun accuracy metric, Kehler et al. (2004), testing on a 2002 ACE news corpus, get a pronoun accuracy (without semantic features) of 75.7%; (Yang et al. 2005), testing on the MUC coreference corpora (also news) get for their single-candidate baseline (without semantic features) 75.1% pronoun accuracy. Although the testing conditions in each case are different, these are comparable to our baseline performance.

5.3 Semantic Information Extraction from Large Corpus

In order to remove noise, we only keep contextual compatibility patterns that appear more than 5 times; and only keep role pair patterns which appear more than 15 times, and appear in more than three different years to avoid random pairs extracted from repeated stories. We automatically extracted 626,008 contextual compatibility patterns and 4,736,359 role pairs. Note that we extract fewer patterns than Yang (2005), who extracted in total 2,203,203 contextual compatibility patterns, from a much smaller corpus (173,252 Wall Street Journal articles). This might be for two reasons: first, we pruned low frequency patterns; second, we used a semantic labeling system instead of shallow parsing. Section 5.6 gives a comparison of pattern extraction based on different levels of analysis.

5.4 Results

	News Corpus		2005 Corpus	
	Accu	**SignTest (p <=)**	**Accu**	**SignTest (p <=)**
baseline	75.54		72.04	
context	76.59	0.025	73.35	0.002
role pair	76.28	0.031	73.03	0.003
combine	77.02	0.0005	73.72	0.0015

Table 3. Accuracy of 5-fold cross-validation with statistics-based semantic features

We did a 5-fold cross validation to test the contribution from statistically-based semantic features, and report an average accuracy. All the mentions and their features are obtained from system output; as a result, if the correct antecedent is not correctly discovered and analyzed from the previous phases, we will not be able to co-refer the pronoun correctly. Experiments on the news articles show that each feature provides approximately 1% gain by itself, and contributes to a substantial overall gain of 1.45%. For the 2005 corpus, the baseline is lower because the documents come from different genres, and we get more gain from each semantic feature. We also computed the significance over the baseline using the sign test[5].

[5] In applying the sign test, we treated each pronoun as an independent sample, which is either correctly resolved or incorrectly resolved. Where the individual observations are

5.5 Self-Extracted Bound

To better understand the potential maximum contribution of our semantic features, we constructed an approximation to the most favorable possible semantic features for each test set. We did this by using perfect coreference knowledge and by collecting patterns for each test set *from the test set itself*. For each corpus used for cross-validation, we first collect all the contextual compatibility and role pair patterns corresponding to the correct antecedents (we ignore the patterns corresponding to the wrong antecedents, because we can not get this negative information when we extract them from a large corpus), and score these patterns to produce semantic features for the MaxEnt Model, both training and testing. We then use these features in the model and do a cross-validation as before. Also, as before, we rely on system output to identify and analyze potential antecedents; if the prior phases do not do so correctly, coreference analysis may well fail. This experiment shows that we can get about 3 to 4% gain from each feature type separately; 4.5 to 5.5% gain is achieved from the two features together.

	News Corpus		2005 Corpus	
	Accu	**SignTest (p <=)**	**Accu**	**SignTest (p <=)**
baseline	75.54		72.04	
context	79.23	7e-14	76.04	9e-27
role pair	78.85	6e-13	75.95	1e-26
combine	79.97	4e-16	77.50	2e-38

Table 4. Accuracy of 5-fold cross-validation with self-extracted semantic features

5.6 Comparison between Semantic and Syntax Patterns

To better understand the difference between semantic role labeling and syntactic relations, we did a comparison between patterns extracted from the syntax level and those extracted from semantic role labeling:

Experiments show that using semantic roles (such as Arg0 and Arg1) works better. This may be because the "deeper" representation provides more generalization of relations. For example, the phrases "weapon's use" and "use weapon" share the same semantic relation <Arg1 use weapon>, while they yield different grammatical relations: <T-pos use weapon> and <Obj use weapon>.

	News Corpus		2005 Corpus	
	semantic	**syntax**	**semantic**	**syntax**
baseline	75.54		72.04	
context	79.23	77.73	76.04	75.83
role pair	78.85	76.87	75.95	74.17
combine	79.97	78.42	77.50	76.76

Table 5. Accuracy of 5-fold cross-validation with self-extracted semantic features based on different levels of syntactic/semantic relations

5.7 Error Analysis

We analyzed the errors in the self-extracted results, to see why such corpus-specific semantic features do not produce an even greater reduction in errors. For the contextual compatibility feature, we find cases where an incorrect candidate is equally compatible with the context of the anaphor; for example, if all the candidates are person names, they will share the same context feature because they generate the same ACE type. In other cases, the context does not provide enough information. For example, in a context tuple <Arg0 get *x*>, *x* can be almost any noun, because "get" is too vague to predicate the compatible subjects. There are similar limitations with the role pair feature; for example, <Arg0 get they> can be associated with a lot of other actions.

To quantify this problem, we counted the patterns that appear in both positive examples (correct antecedents) and negative examples (incorrect antecedents). For contextual compatibility patterns, 39.5% of the patterns which appear with positive examples also appear in the negative sample, while for role pair patterns, 19% of the patterns which appear with positive examples also appear in the negative sample. So we see that, even with a pattern set highly tuned to the test set, many patterns do not by themselves serve to distinguish correct from incorrect coreference.

We analyzed some of the cases where the semantic information does not help, or even harms the analysis. In some cases all the antecedent

(changes in) binary outcomes, the sign test provides a suitably sensitive significance test. (In particular, it is comparable to performing a paired t-test over counts of correct resolutions, aggregated over documents.)

scores are very low, either because the patterns are very rare or the antecedent is a common word that appears in a lot of patterns. In other cases, several antecedents have a high compatibility score but the correct one does not have the top score. In these cases, the contextual compatibility is not reliable, as was pointed out by Kehler et al. (2004):

(7) *The model for a republic, adopted over bitter objections from those advocating direct election of a president, is for presidential nominations to be made with public input and the winning candidate decided by a two-thirds majority of Parliament. Former prime minister Paul Keating, who put the republic issue in the spotlight in* ***his*** *unsuccessful 1996 campaign for re-election, welcomed the result.*

Here adding semantic features leads "*his*" to be incorrectly resolved to "*president*" rather than the entity with mentions "*prime minister*" and "*Paul Keating*"; all the relevant patterns are common, but the score for <Arg0 campaign president> is higher (around 0.0012) than for <Arg0 campaign minister> (0.0004) or <Arg0 campaign Person> (0.0006).

Another problem is that the patterns do not capture enough context information, for example:

(8) *The U.S. administration has been pressing the Security Council to adopt a statement condemning Pyongyang for failing to meet* ***its*** *obligations.*

If we can get the semantic context of "*fail to meet its obligations*" instead of "*its obligations*", we might get better solutions for (8).

The role pair information raises similar problems. Some verbs are very vague, like "*get*", "*take*", "*have*", and role pairs with these verbs might not be very useful. Here is an example:

(9) *The retired Greek officer tried to get Ocalan to the Netherlands, home to a large Kurdish community.* ***He*** *claimed he had been manipulated by the government.*

In this sentence, the role pair information is very vague and it is hard to select a proper antecedent by connecting the subject of "*try*" or "*get*" or the object of "*get*" to the subject of "*claim*".

5.8 Limitations of Semantic Features

The availability of very large corpora coupled with improved pre-processing (e.g., faster parsers, accurate semantic labelers) is making it easier to extract large sets of semantic patterns. However, results on "perfect" semantic information show that even if we can get very good semantic features, there are at least two concerns to address:

- How to best capture the context information: larger context patterns may suffer from data sparseness; simple patterns may be insufficiently selective, appearing in both positive and negative samples.
- In some cases, the baseline features are sufficient to select the antecedent and the semantic features only do harm. If we are able to better gauge our confidence in the decisions based on the baseline features and on the semantic features, we may be able to combine these two sources more effectively.

6 Conclusions and Future Work

We have presented two ways to incorporate semantic features into a MaxEnt model-based pronoun coreference system, where these features have been extracted from a large corpus using a baseline IE (Information Extraction) system and a semantic labeling system, with no specific domain information.

We also estimated the maximal benefit of these features and did some error analysis to identify cases where this semantic knowledge did not suffice. Our experiments show the value of these semantic features for pronoun coreference, but also the limitations of our current context representation and reference resolution models.

Last, we compared the features extracted from different levels of analysis, and showed that 'deeper' representations worked better.

References

Shuya Abe, Kentaro Inui, and Yuji Matsumoto. 2008. *Two-phased event relation acquisition: coupling the relation-oriented and argument-oriented approaches.* Proc. 22nd Int'l Conf. on Computational Linguistics (COLING 2008).

David Bean and Ellen Riloff. 2004. *Unsupervised Learning of Contextual Role Knowledge for*

Coreference Resolution. Proc. HLT-NAACL 2004.

Nathanael Chambers and Dan Jurafsky. 2008. *Unsupervised Learning of Narrative Event Chains*. Proc. ACL-08: HLT.

I. Dagan and A. Itai. 1990. *Automatic processing of large corpora for the resolution of anaphora references*. Proc. 13th International Conference on Computational Linguistics (COLING 1990).

J. Hobbs. 1978. *Resolving pronoun references*. Lingua, 44:339–352.

A. Kehler, D. Appelt, L. Taylor, and A. Simma. 2004. *The (non)utility of predicate-argument frequencies for pronoun interpretation*. Proc. HLT-NAACL 2004.

A. Meyers, M. Kosaka, N. Xue, H. Ji, A. Sun, S. Liao and W. Xu. 2009. Automatic Recognition of Logical Relations for English, Chinese and Japanese in the GLARF Framework. *In SEW-2009 (Semantic Evaluations Workshop) at NAACL HLT-2009*

Viktor Pekar. 2006. *Acquisition of verb entailment from text*. Proc. HLT-NAACL 2006.

Simone Paolo Ponzetto and Michael Strube. 2006 *Exploiting semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution*. Proc. HLT-NAACL 2006.

X. Yang, J. Su, G. Zhou, and C. Tan. 2004. *An NP-cluster approach to coreference resolution*. Proc. 20th International Conference on Computational Linguistics (COLING 2004).

Xiaofeng Yang, Jian Su, Chew Lim Tan.2005. *Improving Pronoun Resolution Using Statistics-Based Semantic Compatibility Information*. Proc. 43rd Annual Meeting of the Assn. for Computational Linguistics.

Xiaofeng Yang and Jian Su. 2007. *Coreference Resolution Using Semantic Relatedness Information from Automatically Discovered Patterns*. Proc. 45th Annual Meeting of the Assn. for Computational Linguistics.