Semantic Role Labeling 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

Extending a machine learning based coreference resolution system with a feature capturing automatically generated information about semantic roles improves its performance.

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; Kehler et al., 2004, inter alia). Similarly, many researchers have explored techniques for robust, broad coverage semantic parsing in terms of semantic role labeling (Gildea & Jurafsky, 2002; Carreras & Màrquez, 2005, SRL henceforth).

This paper explores whether coreference resolution can benefit from SRL, more specifically, *which phenomena* are affected by such information. The motivation comes from the fact that current coreference resolution systems are mostly relying on rather shallow features, such as the distance between the coreferent expressions, string matching, and linguistic form. On the other hand, the literature emphasizes since the very beginning the relevance of world knowledge and inference (Charniak, 1973). As an example, consider a sentence from the Automatic Content Extraction (ACE) 2003 data.

(1) 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.

It seems that 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 document-level *event descriptive information* into the relations holding between referring expressions (REs). This layer of semantic context abstracts from the specific lexical expressions used, and therefore represents a higher level of abstraction than predicate argument statistics (Kehler et al., 2004) and Latent Semantic Analysis used as a model of world knowledge (Klebanov & Wiemer-Hastings, 2002). In this respect, the present work is closer in spirit to Ji et al. (2005), who explore the employment of the ACE 2004 relation ontology as a semantic filter.

2 Coreference Resolution Using SRL

2.1 Corpora Used

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 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.

2.2 Learning Algorithm

For learning coreference decisions, we used a Maximum Entropy (Berger et al., 1996) model. 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. First, a set of pre-processing components includ-

[1] We used the training data corpus only, as the availability of the test data was restricted to ACE participants.

	BNEWS				NWIRE			
	#coref ch.	#pron.	#comm. nouns	#prop. names	#coref ch.	#pron.	#comm. nouns	#prop. names
TRAIN.	587	876	572	980	904	1037	1210	2023
DEVEL	201	315	163	465	399	358	485	923
TEST	228	291	238	420	354	329	484	712

Table 1: Partitions of the ACE 2003 training data corpus

ing a chunker and a named entity recognizer is applied to the text in order to identify the noun phrases, which are further taken as REs to be used for instance generation. Instances are created following Soon et al. (2001). During testing the classifier imposes a partitioning on the available REs by clustering each set of expressions labeled as coreferent into the same coreference chain.

2.3 Baseline System Features

Following Ng & Cardie (2002), our baseline system reimplements the Soon et al. (2001) system. The system uses 12 features. Given a pair of candidate referring expressions RE_i and RE_j the features are computed as follows[2].

(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 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 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.

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

(d) Distance features

DISTANCE how many sentences RE_i and RE_j are apart.

2.4 Semantic Role Features

The baseline system employs only a limited amount of semantic knowledge. In particular, semantic information is limited to WordNet semantic class matching. Unfortunately, a simple WordNet semantic class lookup exhibits problems such as coverage and sense disambiguation[3], which make the WN_CLASS feature very noisy. As a consequence, we propose in the following to enrich the semantic knowledge made available to the classifier by using 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 in case 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 ARG_i is always defined with respect to a predicate lemma pred_i. Given such level of semantic information available at the RE level, we introduce two new features[4].

I_SEMROLE the semantic role argument-predicate pairs of RE_i.

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

[4]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

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.

3 Experiments

3.1 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 discard therefore those responses not present in the key, as we are interested here in establishing the upper limit of the improvements given by SRL.

We also 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 in the same key coreference class.

In all experiments, the REs given to the classifier are noun phrases automatically extracted by a pipeline of pre-processing components (i.e. PoS tagger, NP chunker, Named Entity Recognizer).

3.2 Results

Table 2 compares the results between our duplicated Soon baseline and the original system. The systems show a similar performance w.r.t. F-measure. We speculate that the result improvements 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 one incremented with SRL. Performance improvements are highlighted in bold. The tables show that SRL tends to improve system recall, rather than acting as a 'semantic filter' improving precision. Semantic roles therefore seem to trigger a response in cases where more shallow features do not seem to suffice (see example (1)).

	R	P	F_1	A_p	A_{cn}	A_{pn}
baseline	54.5	88.0	67.3	34.7	20.4	53.1
+SRL	**56.4**	**88.2**	**68.8**	**40.3**	**22.0**	52.1

Table 4: Results ACE (merged BNEWS/NWIRE)

Feature	Chi-square
STR_MATCH	1.0
J_SEMROLE	**0.2096**
ALIAS	0.1852
I_SEMROLE	**0.1594**
SEMCLASS	0.1474
DIST	0.1107
GENDER	0.1013
J_PRONOUN	0.0982
NUMBER	0.0578
I_PRONOUN	0.0489
APPOSITIVE	0.0397
PROPER_NAME	0.0141
DEF_NP	0.0016
DEM_NP	0.0

Table 5: χ^2 statistic for each feature

The RE types which are most positively affected by SRL are pronouns and common nouns. On the other hand, SRL information has a limited or even worsening effect on the performance on proper names, where features such as string matching and alias seem to suffice. This suggests that SRL plays a role in pronoun and common noun resolution, where surface features cannot account for complex preferences and semantic knowledge is required.

3.3 Feature Evaluation

We investigated the contribution of the different features in the learning process. Table 5 shows the chi-square statistic (normalized in the $[0, 1]$ interval) for each feature occurring in the training data of the MERGED dataset. SRL features show a high χ^2 value, ranking immediately after string matching and alias, which indicates a high correlation of these features to the decision classes.

The importance of SRL is also indicated by the analysis of the contribution of individual features to the overall performance. Table 6 shows the performance variations obtained by leaving out each feature in turn. Again, it can be seen that removing both I and J_SEMROLE induces a relatively high performance degradation when compared to other features. Their removal ranks 5th out of 12, following only essential features such as string matching, alias, pronoun and number. Similarly to Table 5, the semantic role of the anaphor ranks higher than the one of the antecedent. This re-

	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	37.7	23.1	55.6
+SRL	**50.9**	86.1	**64.0**	**36.8**	**14.3**	**45.7**	**58.3**	86.9	**69.8**	**38.0**	**25.8**	**55.8**

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

Feature(s) removed	Δ F_1
all features	68.8
STR_MATCH	−21.02
ALIAS	−2.96
I/J_PRONOUN	−2.94
NUMBER	−1.63
I/J_SEMROLE	**−1.50**
J_SEMROLE	**−1.26**
APPOSITIVE	−1.20
GENDER	−1.13
I_SEMROLE	**−0.74**
DIST	−0.69
WN_CLASS	−0.56
DEF_NP	−0.57
DEM_NP	−0.50
PROPER_NAME	−0.49

Table 6: Δ F_1 from feature removal

lates to the improved performance on pronouns, as it indicates that SRL helps for linking anaphoric pronouns to preceding REs. Finally, it should be noted that SRL provides much more solid and noise-free semantic features when compared to the WordNet class feature, whose removal induces always a lower performance degradation.

4 Conclusion

In this paper we have investigated the effects of using semantic role information within a machine learning based coreference resolution system. Empirical results show that coreference resolution can benefit from SRL. The analysis of the relevance of features, which had not been previously addressed, indicates that incorporating semantic information as shallow event descriptions improves the performance of the classifier. The generated model is able to learn selection preferences in cases where surface morpho-syntactic features do not suffice, i.e. pronoun resolution.

We speculate that this contrasts with the disappointing findings of Kehler et al. (2004) since SRL provides a more *fine grained* level of information when compared to predicate argument statistics. As it models the semantic relationship that a syntactic constituent has with a predicate, it carries indirectly *syntactic* preference information. In addition, when used as a feature it allows the classifier to infer *semantic* role co-occurrence, thus inducing deep representations of the predicate argument relations for learning in coreferential contexts.

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).

References

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.

Carreras, X. & L. Màrquez (2005). Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proc. of CoNLL-05*, pp. 152–164.

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.

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

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.

Klebanov, B. & P. Wiemer-Hastings (2002). The role of wor(l)d knowledge in pronominal anaphora resolution. In *Proceedings of the International Symposium on Reference Resolution for Natural Language Processing, Alicante, Spain, 3–4 June, 2002*, pp. 1–8.

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.

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.

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.

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.