Using the Web in Machine Learning for Other-Anaphora Resolution

Natalia N. Modjeska
School of Informatics
University of Edinburgh and
Department of Computer Science
University of Toronto
natalia@cs.utoronto.ca

Katja Markert
School of Computing
University of Leeds and
School of Informatics
University of Edinburgh
markert@inf.ed.ac.uk

Malvina Nissim
School of Informatics
University of Edinburgh
mnissim@inf.ed.ac.uk

Abstract

We present a machine learning framework for resolving other-anaphora. Besides morpho-syntactic, recency, and semantic features based on existing lexical knowledge resources, our algorithm obtains additional semantic knowledge from the Web. We search the Web via lexico-syntactic patterns that are specific to other-anaphors. Incorporating this innovative feature leads to an 11.4 percentage point improvement in the classifier's F-measure (25% improvement relative to results without this feature).

1 Introduction

Other-anaphors are referential NPs with the modifiers "other" or "another" and non-structural antecedents:[1]

(1) An exhibition of American design and architecture opened in September in *Moscow* and will travel to **eight other Soviet cities**.

(2) [...] the alumni director of *a Big Ten university* "I'd love to see sports cut back and so would a lot of my counterparts at **other schools**, [...]"

(3) You either believe Seymour can do it again or you don't. Beside *the designer's age*, **other risk factors for Mr. Cray's company** include the Cray-3's [...] chip technology.

In (1), "eight other Soviet cities" refers to a set of Soviet cities *excluding* Moscow, and can be rephrased as "eight Soviet cities other than Moscow". In (2), "other schools" refers to a set of schools *excluding* the mentioned Big Ten university. In (3), "other risk factors for Mr. Cray's company" refers to a set of risk factors *excluding* the designer's age.

In contrast, in list-contexts such as (4), the antecedent is available both anaphorically and *structurally*, as the left conjunct of the anaphor.[2]

(4) Research shows AZT can relieve *dementia* and **other symptoms** in children [...]

We focus on cases such as (1–3).

Section 2 describes a corpus of other-anaphors. We present a machine learning approach to other-anaphora, using a Naive Bayes (NB) classifier (Section 3) with two different feature sets. In Section 4 we present the first feature set (*F1*) that includes standard morpho-syntactic, recency, and string comparison features. However, there is evidence that, e.g., syntactic features play a smaller role in resolving anaphors with full lexical heads than in pronominal anaphora (Strube, 2002; Modjeska, 2002). Instead, a large and diverse amount of lexical or world knowledge is necessary to understand examples such as (1–3), e.g., that Moscow is a (Soviet) city, that universities are informally called schools in American English and that age can be viewed as a risk factor. Therefore we add lexical knowledge, which is extracted from WordNet (Fellbaum, 1998) and from a Named Entity (NE) Recognition algorithm, to *F1*.

[1] All examples are from the *Wall Street Journal*; the correct antecedents are in italics and the anaphors are in bold font.

[2] Antecedents are also available structurally in constructions "other than", e.g., "few clients other than the state". For a computational treatment of "other" with structural antecedents see (Bierner, 2001).

The algorithm's performance with this feature set is encouraging. However, the semantic knowledge the algorithm relies on is not sufficient for many cases of other-anaphors (Section 4.2). Many expressions, word senses and lexical relations are missing from WordNet. Whereas it includes Moscow as a hyponym of city, so that the relation between anaphor and antecedent in (1) can be retrieved, it does not include the sense of school as university, nor does it allow to infer that age is a risk factor.

There have been efforts to extract missing lexical relations from corpora in order to build new knowledge sources and enrich existing ones (Hearst, 1992; Berland and Charniak, 1999; Poesio et al., 2002).[3] However, the size of the used corpora still leads to data sparseness (Berland and Charniak, 1999) and the extraction procedure can therefore require extensive smoothing. Moreover, some relations should probably not be encoded in fixed context-independent ontologies at all. Should, e.g., underspecified and point-of-view dependent hyponymy relations (Hearst, 1992) be included? Should age, for example, be classified as a hyponym of risk factor independent of context?

Building on our previous work in (Markert et al., 2003), we instead claim that the Web can be used as a huge additional source of domain- and context-independent, rich and up-to-date knowledge, without having to build a fixed lexical knowledge base (Section 5). We describe the benefit of integrating Web frequency counts obtained for lexico-syntactic patterns specific to *other*-anaphora as an additional feature into our NB algorithm. This feature raises the algorithm's F-measure from 45.5% to 56.9%.

2 Data Collection and Preparation

We collected 500 other-anaphors with NP antecedents from the *Wall Street Journal* corpus (Penn Treebank, release 2). This data sample excludes several types of expressions containing "other": (a) list-contexts (Ex. 4) and other-than contexts (footnote 2), in which the antecedents are available structurally and thus a relatively unsophisticated procedure would suffice to find them; (b) idiomatic and discourse connective "other", e.g., "on the other hand", which are not anaphoric; and (c) reciprocal "each other" and "one another", elliptic phrases e.g. "one X ... the other(s)" and *one*-anaphora, e.g., "the other/another one", which behave like pronouns and thus would require a different search method. Also excluded from the data set are samples of other-anaphors with non-NP antecedents (e.g., adjectival and nominal pre- and postmodifiers and clauses).

Each anaphor was extracted in a 5-sentence context. The correct antecedents were manually annotated to create a training/test corpus. For each anaphor, we automatically extracted a set of potential NP antecedents as follows. First, we extracted all base NPs, i.e., NPs that contain no further NPs within them. NPs containing a possessive NP modifier, e.g., "Spain's economy", were split into a possessor phrase, "Spain", and a possessed entity, "economy". We then filtered out null elements and lemmatised all antecedents and anaphors.

3 The Algorithm

We use a Naive Bayes classifier, specifically the implementation in the Weka ML library.[4]

The training data was generated following the procedure employed by Soon et al. (2001) for coreference resolution. Every pair of an anaphor and its closest preceding antecedent created a positive training instance. To generate negative training instances, we paired anaphors with each of the NPs that intervene between the anaphor and its antecedent. This procedure produced a set of 3,084 antecedent-anaphor pairs, of which 500 (16%) were positive training instances.

The classifier was trained and tested using 10-fold cross-validation. We follow the general practice of ML algorithms for coreference resolution and compute *precision (P)*, *recall (R)*, and *F-measure* (F) on all possible anaphor-antecedent pairs.

As a first approximation of the difficulty of our task, we developed a simple rule-based baseline algorithm which takes into account the fact that the lemmatised head of an other-anaphor is sometimes the same as that of its antecedent, as in (5).

[3] In parallel, efforts have been made to enrich WordNet by adding information in glosses (Harabagiu et al., 1999).

[4] `http://www.cs.waikato.ac.nz/~ml/weka/`. We also experimented with a decision tree classifier, with Neural Networks and Support Vector Machines with Sequential Minimal Optimization (SMO), all available from Weka. These classifiers achieved worse results than NB on our data set.

Table 1: Feature set *F1*

Type	Feature	Description	Values
Gramm	NP_FORM	Surface form (for all NPs)	definite, indefinite, demonstrative, pronoun, proper name, unknown
Match	RESTR_SUBSTR	Does lemmatized antecedent string contain lemmatized anaphor string?	yes, no
Syntactic	GRAM_FUNC	Grammatical role (for all NPs)	subject, predicative NP, dative object, direct object, oblique, unknown
Syntactic	SYN_PAR	Anaphor-antecedent agreement with respect to grammatical function	yes, no
Positional	SDIST	Distance between antecedent and anaphor in sentences	1, 2, 3, 4, 5
Semantic	SEMCLASS	Semantic class (for all NPs)	person, organization, location, date, money, number, thing, abstract, unknown
Semantic	SEMCLASS_AGR	Anaphor-antecedent agreement with respect to semantic class	yes, no, unknown
Semantic	GENDER_AGR	Anaphor-antecedent agreement with respect to gender	same, compatible, incompatible, unknown
Semantic	RELATION	Type of relation between anaphor and antecedent	same-predicate, hypernymy, meronymy, compatible, incompatible, unknown

(5) *These three countries* aren't completely off the hook, though. They will remain on a lower-priority list that includes **other countries** [. . .]

For each anaphor, the baseline string-compares its last (lemmatised) word with the last (lemmatised) word of each of its possible antecedents. If the words match, the corresponding antecedent is chosen as the correct one. If several antecedents produce a match, the baseline chooses the most recent one among them. If string-comparison returns no antecedent, the baseline chooses the antecedent closest to the anaphor among all antecedents. The baseline assigns "yes" to exactly one antecedent per anaphor. Its P, R and F-measure are 27.8%.

4 Naive Bayes without the Web

First, we trained and tested the NB classifier with a set of 9 features motivated by our own work on other-anaphora (Modjeska, 2002) and previous ML research on coreference resolution (Aone and Bennett, 1995; McCarthy and Lehnert, 1995; Soon et al., 2001; Ng and Cardie, 2002; Strube et al., 2002).

4.1 Features

A set of 9 features, *F1*, was automatically acquired from the corpus and from additional external resources (see summary in Table 1).

Non-semantic features. NP_FORM is based on the POS tags in the *Wall Street Journal* corpus and heuristics. RESTR_SUBSTR matches lemmatised strings and checks whether the antecedent string contains the anaphor string. This allows to resolve examples such as "one woman ringer . . . another woman". The values for GRAM_FUNC were approximated from the parse trees and Penn Treebank annotation. The feature SYN_PAR captures syntactic parallelism between anaphor and antecedent. The feature SDIST measures the distance between anaphor and antecedent in terms of sentences.[5]

Semantic features. GENDER_AGR captures agreement in gender between anaphor and antecedent, gender having been determined using gazetteers, kinship and occupational terms, titles, and WordNet. Four values are possible: "same", if both NPs have same gender; "compatible", if antecedent and anaphor have compatible gender, e.g., "lawyer . . . other women"; "incompatible", e.g., "Mr. Johnson . . . other women"; and "unknown", if one of the NPs is undifferentiated, i.e., the gender value is "unknown". SEMCLASS: Proper names were classified using ANNIE, part of the GATE2 software package (`http://gate.ac.uk`). Common nouns were looked up in WordNet, considering only the most frequent sense of each noun (the first sense in WordNet). In each case, the output was mapped onto one of the values in Table 1. The SEMCLASS_AGR fea-

[5] We also experimented with a feature MDIST that measures intervening NP units. This feature worsened the overall performance of the classifier.

ture compares the semantic class of the antecedent with that of the anaphor NP and returns "yes" if they belong to the same class; "no", if they belong to different classes; and "unknown" if the semantic class of either the anaphor or antecedent has not been determined. The RELATION between other-anaphors and their antecedents can partially be determined by string comparison ("same-predicate")[6] or WordNet ("hypernymy" and "meronymy"). As other relations, e.g. "redescription" (Ex. (3), cannot be readily determined on the basis of the information in WordNet, the following values were used: "compatible", for NPs with compatible semantic classes, e.g., "woman … other leaders"; and "incompatible", e.g., "woman … other economic indicators". Compatibility can be defined along a variety of parameters. The notion we used roughly corresponds to the root level of the WordNet hierarchy. Two nouns are compatible if they have the same SEMCLASS value, e.g., "person". "Unknown" was used if the type of relation could not be determined.

4.2 Results

Table 2 shows the results for the Naive Bayes classifier using *F1* in comparison to the baseline.

Table 2: Results with *F1*

Features	P	R	*F*
baseline	27.8	27.8	27.8
F1	51.7	40.6	45.5

Our algorithm performs significantly better than the baseline.[7] While these results are encouraging, there were several classification errors.

Word sense ambiguity is one of the reasons for misclassifications. Antecedents were looked up in WordNet for their most frequent sense for a context-independent assignment of the values of semantic class and relations. However, in many cases either the anaphor or antecedent or both are used in a sense that is ranked as less frequent in Wordnet. This might even be a quite frequent sense for a specific corpus, e.g., the word "issue" in the sense of "shares, stocks" in the WSJ. Therefore there is a strong interaction between word sense disambiguation and reference resolution (see also (Preiss, 2002)).

Named Entity resolution is another weak link. Several correct NE antecedents were classified as "antecedent=no" (false negatives) because the NER module assigned the wrong class to them.

The largest class of errors is however due to insufficient semantic knowledge. Problem examples can roughly be classified into five partially overlapping groups: (a) examples that suffer from gaps in WordNet, e.g., (2); (b) examples that require domain-, situation-specific, or general world knowledge, e.g., (3); (c) examples involving bridging phenomena (sometimes triggered by a metonymic or metaphoric antecedent or anaphor), e.g., (6); (d) redescriptions and paraphrases, often involving semantically vague anaphors and/or antecedents, e.g., (7) and (3); and (e) examples with ellipsis, e.g., (8).

(6) *The Justice Department*'s view is shared by **other lawyers** [. . .]

(7) While Mr. Dallara and Japanese officials say *the question of investors' access to the U.S. and Japanese markets* may get a disproportionate share of the public's attention, a number of **other important economic issues** will be on the table at next week's talks.

(8) He sees *flashy sports* as the only way the last-place network can cut through the clutter of cable and VCRs, grab millions of new viewers and tell them about **other shows** premiering a few weeks later.

In (6), the antecedent is an *organization-for-people* metonymy. In (7), the question of investors' access to the U.S. and Japanese markets is characterized as an important economic issue. Also, the head "issues" is lexically uninformative to sufficiently constrain the search space for the antecedent. In (8), the antecedent is not the flashy sports, but rather flashy sport shows, and thus an important piece of information is omitted. Alternatively, the antecedent is a *content-for-container* metonymy.

Overall, our approach misclassifies antecedents whose relation to the other-anaphor is based on similarity, property-sharing, causality, or is constrained to a specific domain. These relation types are not — and perhaps *should not* be — encoded in WordNet.

[6]Same-predicate is not really a relation. We use it when the head noun of the anaphor and antecedent are the same.

[7]We used a t-test with confidence level 0.05 for all significance tests.

5 Naive Bayes with the Web

With its approximately 3033M pages[8] the Web is the largest corpus available to the NLP community. Building on our approach in (Markert et al., 2003), we suggest using the Web as a knowledge source for anaphora resolution. In this paper, we show how to integrate Web counts for lexico-syntactic patterns specific to other-anaphora into our ML approach.

5.1 Basic Idea

In the examples we consider, the relation between anaphor and antecedent is implicitly expressed, i.e., anaphor and antecedent do not stand in a structural relationship. However, they are linked by a strong semantic relation that is likely to be *structurally explicitly expressed* in other texts. We exploit this insight by adopting the following procedure:

1. In other-anaphora, a hyponymy/similarity relation between the lexical heads of anaphor and antecedent is exploited or stipulated by the context,[9] e.g. that "schools" is an alternative term for universities in Ex. (2) or that age is viewed as a risk factor in Ex. (3).

2. We select patterns that structurally explicitly express the same lexical relations. E.g., the list-context `NP`$_1$ `and other NP`$_2$ (as Ex. (4)) usually expresses hyponymy/similarity relations between the hyponym NP_1 and its hypernym NP_2 (Hearst, 1992).

3. If the implicit lexical relationship between anaphor and antecedent is strong, it is likely that anaphor and antecedent also frequently cooccur in the selected explicit patterns. We instantiate the explicit pattern for all anaphor-antecedent pairs. In (2) the pattern `NP`$_1$ `and other NP`$_2$ is instantiated with e.g., `counterparts and other schools`, `sports and other schools` and `universities and other schools`.[10] These instantiations can be searched in any corpus to determine their frequencies. The rationale is that the most frequent of these instantiated patterns is a good clue for the correct antecedent.

4. As the patterns can be quite elaborate, most corpora will be too small to determine the corresponding frequencies reliably. The instantiation `universities and other schools`, e.g., does not occur at all in the British National Corpus (BNC), a 100M words corpus of British English.[11] Therefore we use the largest corpus available, the Web. We submit all instantiated patterns as queries making use of the Google API technology. Here, `universities and other schools` yields over 700 hits, whereas the other two instantiations yield under 10 hits each. High frequencies do not only occur for synonyms; the corresponding instantiation for the correct antecedent in Ex. (3) `age and other risk factors` yields over 400 hits on the Web and again none in the BNC.

5.2 Antecedent Preparation

In addition to the antecedent preparation described in Section 2, further processing is necessary. First, pronouns can be antecedents of other-anaphors but they were not used as Web query input as they are lexically empty. Second, all modification was eliminated and only the rightmost noun of compounds was kept, to avoid data sparseness. Third, using patterns containing NEs such as "Will Quinlan" in (9) also leads to data sparseness (see also the use of NE recognition for feature SEMCLASS).

(9) [...] *Will Quinlan* had not inherited a damaged retinoblastoma supressor gene and, therefore, faced no more risk than **other children** [...]

We resolved NEs in two steps. In addition to GATE's classification into ENAMEX and NUMEX categories, we used heuristics to automatically obtain more fine-grained distinctions for the categories LOCATION, ORGANIZATION, DATE and MONEY, whenever possible. No further distinctions were made for the category PERSON. We classified LOCATIONS into COUNTRY, (US) STATE, CITY, RIVER, LAKE and OCEAN, using mainly

[8] `http://www.searchengineshowdown.com/stats/sizeest.shtml`, data from March 2003.

[9] In the Web feature context, we will often use "anaphor/antecedent" instead of the more cumbersome "lexical heads of the anaphor/antecedent".

[10] These simplified instantiations serve as an example and are neither exhaustive nor the final instantiations we use; see Section 5.3.

[11] `http://info.ox.ac.uk/bnc`

Table 3: Patterns and Instantiations for other-anaphora

ANTECEDENT	PATTERN	INSTANTIATIONS
common noun	(O1): `(`N_1\{*sg*\} `OR` N_1\{*pl*\}`) and other` N_2\{*pl*\}	I_1^c: "(university OR universities) and other schools"
proper name	(O1): `(`N_1\{*sg*\} `OR` N_1\{*pl*\}`) and other` N_2\{*pl*\}	I_1^p: "(person OR persons) and other children"
		I_2^p: "(child OR children) and other persons"
	(O2): N_1 `and other` N_2\{*pl*\}	I_3^p: "Will Quinlan and other children"

gazetteers.[12] If an entity classified by GATE as ORGANIZATION contained an indication of the organization type, we used this as a subclassification; therefore "Bank of America" is classified as BANK. For DATE and MONEY entities we used simple heuristics to classify them further into DAY, MONTH, YEAR as well as DOLLAR.

From now on we call $\mathcal{A}$ the list of possible antecedents and ana the anaphor. For (2), this list is $\mathcal{A}_2$=\{*counterpart, sport, university*\} (the pronoun "I" has been discarded) and ana_2=*school*. For (9), they are $\mathcal{A}_9$=\{*risk, gene, person [=Will Quinlan]*\} and ana_9 =*child*.

5.3 Queries and Scoring Method

We use the list-context pattern:[13]

(O1) `(`N_1\{*sg*\} `OR` N_1\{*pl*\}`) and other` N_2\{*pl*\}

For common noun antecedents, we instantiate the pattern by substituting N_1 with each possible antecedent from set $\mathcal{A}$, and N_2 with *ana*, as normally N_1 is a *hyponym* of N_2 in (O1), and the antecedent is a hyponym of the anaphor. An instantiated pattern for Ex. (2) is `(university OR universities) and other schools` (I_1^c in Table 3).[14]

For NE antecedents we instantiate (O1) by substituting N_1 with the NE category of the antecedent, and N_2 with ana. An instantiated pattern for Example (9) is `(person OR persons) and other children` (I_1^p in Table 3). In this instantiation, N_1 ("person") is not a hyponym of N_2 ("child"), instead N_2 is a hyponym of N_1. This is a consequence of the substitution of the antecedent ("Will Quinlan") with its NE category ("person"); such an instantiation is not frequent, since it violates standard relations within (O1). Therefore, we also instantiate (O1) by substituting N_1 with ana, and N_2 with the NE type of the antecedent (I_2^p in Table 3). Finally, for NE antecedents, we use an additional pattern:

(O2) N_1 `and other` N_2\{*pl*\}

which we instantiate by substituting N_1 with the original NE antecedent and N_2 with ana (I_3^p in Table 3).

Patterns and instantiations are summarised in Table 3. We submit these instantiations as queries to the Google search engine.

For each antecedent ant in $\mathcal{A}$ we obtain the raw frequencies of all instantiations it occurs in (I_1^c for common nouns, or I_1^p, I_2^p, I_3^p for proper names) from the Web, yielding $freq(I_1^c)$, or $freq(I_1^p)$, $freq(I_2^p)$ and $freq(I_3^p)$. We compute the maximum M_{ant} over these frequencies for proper names. For common nouns M_{ant} corresponds to $freq(I_1^c)$. The instantiation yielding M_{ant} is then called $Imax_{ant}$.

Our scoring method takes into account the individual frequencies of ant and ana by adapting *mutual information*. We call the first part of $Imax_{ant}$ (e.g. "university OR universities", or "child OR children") X_{ant}, and the second part (e.g. "schools" or "persons") Y_{ant}. We compute the probability of $Imax_{ant}$, X_{ant} and Y_{ant}, using Google to determine $freq(X_{ant})$ and $freq(Y_{ant})$.

$$Pr(Imax_{ant}) = \frac{M_{ant}}{\text{number of GOOGLE pages}}$$

$$Pr(X_{ant}) = \frac{freq(X_{ant})}{\text{number of GOOGLE pages}}$$

$$Pr(Y_{ant}) = \frac{freq(Y_{ant})}{\text{number of GOOGLE pages}}$$

[12] They were extracted from the Web. Small gazetteers, containing in all about 500 entries, are sufficient. This is the only external knowledge collected for the Web feature.

[13] In all patterns in this paper, "OR" is the boolean operator, "N_1" and "N_2" are variables, all other words are constants.

[14] Common noun instantiations are marked by a superscript "c" and proper name instantiations by a superscript "p".

We then compute the final score MI_{ant}.

$$MI_{ant} = log \frac{Pr(Imax_{ant})}{Pr(X_{ant})Pr(Y_{ant})}$$

5.4 Integration into ML Framework and Results

For each anaphor, the antecedent in $\mathcal{A}$ with the highest MI_{ant} gets feature value "webfirst".[15] All other antecedents (including pronouns) get the feature value "webrest". We chose this method instead of e.g., giving score intervals for two reasons. First, since score intervals are unique for each anaphor, it is not straightforward to incorporate them into a ML framework in a consistent manner. Second, this method introduces an element of *competition* between several antecedents (see also (Connolly et al., 1997)), which the individual scores do not reflect.

We trained and tested the NB classifier with the feature set *F1*, plus the Web feature. The last row in Table 4 shows the results. We obtained a 9.1 percentage point improvement in precision (an 18% improvement relative to the *F1* feature set) and a 12.8 percentage point improvement in recall (32% improvement relative to *F1*), which amounts to an 11.4 percentage point improvement in F-measure (25% improvement relative to *F1* feature set). In particular, all the examples in this paper were resolved.

Our algorithm still misclassified several antecedents. Sometimes even the Web is not large enough to contain the instantiated pattern, especially when this is situation or speaker specific. Another problem is the high number of NE antecedents (39.6%) in our corpus. While our NER module is quite good, any errors in NE classification lead to incorrect instantiations and thus to incorrect classifications. In addition, the Web feature does not yet take into account pronouns (7.43% of all correct and potential antecedents in our corpus).

6 Related Work and Discussion

Modjeska (2002) presented two hand-crafted algorithms, SAL and LEX, which resolve the anaphoric references of other-NPs on the basis of grammatical salience and lexical information from WordNet, respectively. In our own previous work (Markert et al., 2003) we presented a preliminary symbolic approach that uses Web counts and a recency-based tie-breaker for resolution of other-anaphora and bridging descriptions. (For another Web-based symbolic approach to bridging see (Bunescu, 2003).) The approach described in this paper is the first machine learning approach to other-anaphora. It is not directly comparable to the symbolic approaches above for two reasons. First, the approaches differ in the data and the evaluation metrics they used. Second, our algorithm does not yet constitute a full resolution procedure. As the classifier operates on the whole set of antecedent-anaphor pairs, more than one potential antecedent for each anaphor can be classified as "antecedent=yes". This can be amended by e.g. incremental processing. Also, the classifier does not know that each other-NP is anaphoric and therefore has an antecedent. (This contrasts with e.g. definite NPs.) Thus, it can classify all antecedents as "antecedent=no". This can be remedied by using a back-off procedure, or a competition learning approach (Connolly et al., 1997). Finally, the full resolution procedure will have to take into account other factors, e.g., syntactic constraints on antecedent realization.

Our approach is the first ML approach to any kind of anaphora that integrates the Web. Using the Web as a knowledge source has considerable advantages. First, the size of the Web almost eliminates the problem of data sparseness for our task. For this reason, using the Web has proved successful in several other fields of NLP, e.g., machine translation (Grefenstette, 1999) and bigram frequency estimation (Keller et al., 2002). In particular, (Keller et al., 2002) have shown that using the Web handles data sparseness better than smoothing. Second, we do not process the returned Web pages in any way (tagging, parsing, e.g.), unlike e.g. (Hearst, 1992; Poesio et al., 2002). Third, the linguistically motivated patterns we use reduce long-distance dependencies

Table 4: Results with *F1* and *F1+Web*

Features	P	R	F
baseline	27.8	27.8	27.8
F1	51.7	40.6	45.5
F1+Web	60.8	53.4	56.9

[15] If several antecedents have the highest MI_{ant} they all get value "webfirst".

between anaphor and antecedent to local dependencies. By looking up these patterns on the Web we obtain semantic information that is not and perhaps should not be encoded in an ontology (redescriptions, vague relations, etc.). Finally, these local dependencies also reduce the need for prior word sense disambiguation, as the anaphor and the antecedent constrain each other's sense within the context of the pattern.

7 Conclusions

We presented a machine learning approach to other-anaphora, which uses a NB classifier and two sets of features. The first set consists of standard morpho-syntactic, recency, and semantic features based on WordNet. The second set also incorporates semantic knowledge obtained from the Web via lexico-semantic patterns specific to other-anaphora. Adding this knowledge resulted in a dramatic improvement of 11.4% points in the classifier's F-measure, yielding a final F-measure of 56.9%.

To our knowledge, we are the first to integrate a Web feature into a ML framework for anaphora resolution. Adding this feature is inexpensive, solves the data sparseness problem, and allows to handle examples with non-standard relations between anaphor and antecedent. The approach is easily applicable to other anaphoric phenomena by developing appropriate lexico-syntactic patterns (Markert et al., 2003).

Acknowledgments

Natalia N.Modjeska is supported by EPSRC grant GR/M75129; Katja Markert by an Emmy Noether Fellowship of the Deutsche Forschungsgemenschaft. We thank three anonymous reviewers for helpful comments and suggestions.

References

C. Aone and S. W. Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proc. of ACL'95*, pages 122–129.

M. Berland and E. Charniak. 1999. Finding parts in very large corpora. In *Proc. of ACL'99*, pages 57–64.

G. Bierner. 2001. Alternative phrases and natural language information retrieval. In *Proc. of ACL'01*.

R. Bunescu. 2003. Associative anaphora resolution: A Web-based approach. In R. Dale, K. van Deemter, and R. Mitkov, editors, *Proc. of the EACL Workshop on the Computational Treatment of Anaphora*.

D. Connolly, J. D. Burger, and D. S. Day. 1997. A machine learning approach to anaphoric reference. In Daniel Jones and Harold Somers, editors, *New Methods in Language Processing*, pages 133–144. UCL Press, London.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

G. Grefenstette. 1999. The WWW as a resource for example-based MT tasks. In *Proc. of ASLIB'99 Translating and the Computer 21*, London.

S. Harabagiu, G. Miller, and D. Moldovan. 1999. Wordnet 2 - a morphologically and semantically enhanced resource. In *Proc. of SIGLEX-99*, pages 1–8.

M. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of COLING-92*.

F. Keller, M. Lapata, and O. Ourioupina. 2002. Using the Web to overcome data sparseness. In *Proc. of EMNLP 2002*, pages 230–237.

K. Markert, M. Nissim, and N. N. Modjeska. 2003. Using the Web for nominal anaphora resolution. In R. Dale, K. van Deemter, and R. Mitkov, editors, *Proc. of the EACL Workshop on the Computational Treatment of Anaphora*, pages 39–46.

J. F. McCarthy and W. G. Lehnert. 1995. Using decision trees for coreference resolution. In *Proc. of IJCAI-95*, pages 1050–1055.

N. N. Modjeska. 2002. Lexical and grammatical role constraints in resolving *other*-anaphora. In *Proc. of DAARC 2002*, pages 129–134.

V. Ng and C. Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proc. of ACL'02*, pages 104–111.

M. Poesio, T. Ishikawa, S. Schulte im Walde, and R. Viera. 2002. Acquiring lexical knowledge for anaphora resolution. In *Proc. of LREC 2002*, pages 1220–1224.

J. Preiss. 2002. Anaphora resolution with word sense disambiguation. In *Proc. of SENSEVAL-2*, pages 143–146.

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

M. Strube, S. Rapp, and C. Müller. 2002. The influence of minimum edit distance on reference resolution. In *Proc. of EMNLP 2002*, pages 312–319.

M. Strube. 2002. NLP approaches to reference resolution. Tutorial notes, *ACL'02*.