Automatic Acquisition of Gender Information for Anaphora Resolution

Shane Bergsma

Department of Computing Science,
University of Alberta,
Edmonton, Alberta, Canada T6G 2E8
bergsma@cs.ualberta.ca

Abstract. We present a novel approach to learning gender and number information for anaphora resolution. Noun-pronoun pair counts are collected from gender-indicating lexico-syntactic patterns in parsed corpora, and occurrences of noun-pronoun pairs are mined online from the web. Gender probabilities gathered from these templates provide features for machine learning. Both parsed corpus and web-based features allow for accurate prediction of the gender of a given noun phrase. Together they constructively combine for 96% accuracy when estimating gender on a list of noun tokens, better than any of our human participants achieved. We show that using this gender information in simple or knowledge-rich pronoun resolution systems significantly improves performance over traditional gender constraints. Our novel gender strategy would benefit any of the current top-performing coreference resolution systems.

1 Introduction

Anaphora resolution determines which previous entity (the antecedent) a given noun phrase (the anaphor) refers to. We focus on resolving third-person pronominal anaphora, including reflexives. Performing these resolutions has long been considered a challenging yet vital task for a number of Natural Language Processing applications. We present a new approach for determining the antecedents of pronouns using enhanced statistical gender and number information. Gender and number agreement provide one of the most important, intuitive, and widely-accepted constraints for resolving anaphora. In the following example, gender information allows us to select the correct antecedent:

1. John never saw the car. He arrived late. (resolve "he" to"John").
2. John never saw the car. It arrived late. (resolve "it" to "the car").

How does one encode gender and number agreement when implementing an anaphora resolution system? Often, number information is available from the parsers used to pre-process the text. Number agreement can improve the performance of these technologies, while gender information is neglected [1]. We use Dekang Lin's parser Minipar to induce dependency trees and extract information on plurals [8]. There are also a number of so-called "surface clues" that give

a textual indication to the gender[1] of a noun phrase. In English, pronouns are some of the few remaining overtly gender-indicating words. Gendered designators (such as Mr., Mrs., etc.) also provide gender information for a noun phrase, but gender-indicating suffixes such as those used in "actress" or "chairman" have fallen out of favour, and are hence unreliable [1].

In Section 2 we summarize related anaphora resolution systems and previous gender strategies. Most previous approaches treat gender as a hard constraint, a filter on candidate antecedents. If the gender is not known exactly, it is not used. Our approach treats gender as a probability – another factor that can be used in anaphora resolution. We seek to determine the probability that a given noun is masculine, feminine, neutral or plural. By counting occurrences of noun-pronoun patterns in parsed corpora and on the web, we automatically learn the probabilities that a name like "Alex" is used in a masculine (very common), feminine (less common), neutral or plural (zero probability) context. We automatically learn the neutral preference of company names and organizations, and which words, like "child" or "parent," are likely to be masculine or feminine, but not neutral or plural. Section 3 describes the gender-gathering templates and how gender probability is modelled from the resulting frequency counts.

The accuracy of our acquired data is tested in a novel gender classification task. A classifier uses the gender probabilities to guess the gender of noun tokens extracted from text. Section 4 describes how the within-context gender of these nouns is obtained. Section 5 shows that our classifier's results surpass those of human guessers, and that classifiers using the web-based probabilities, despite their noise, actually outperform those using the corpus information.

Finally, in Section 6, we demonstrate the superiority of using additional gender information by testing machine-learned anaphora resolution classifiers based on the gender statistics. When deciding whether a candidate antecedent co-refers with a pronoun, it makes sense to incorporate the probability this antecedent is of the pronoun's gender. When gender is used as a hard constraint, restricted to completely certain cases, significantly more antecedents are missed.

2 Related Work

Anaphora resolution systems typically employ some combination of constraints and preferences to select the correct antecedent. Constraints eliminate possible candidates by virtue of gender and number disagreement, binding theory violations, etc., while preferences encourage selection of antecedents which are more recent, more frequent, etc. These approaches are generally not based on machine learning from a corpus [7], [6]. Our approach follows a more recent trend toward using an annotated corpus to learn an anaphora resolution classifier [2], or coreference resolution classifier [12]. The first machine learning approach to anaphora

[1] There are four possibilities for gender and number of third-person pronouns: masculine, feminine, neutral and plural (e.g., *he, she, it, they*). Since our approach gathers information for all four, whenever we subsequently refer to the *gender* of a noun phrase, we implicitly include plural as one of the options.

resolution that uses the web is reportedly due to Modjeska, Markert and Nissim, who look at page counts of various patterns for other-anaphora resolution [10].

Current, top-performing coreference resolution systems selectively use surface clues and information derived from WordNet [12]. If a noun's most frequent sense in the WordNet synset is a subclass of a predefined gendered class (e.g., object, which is neutral), gender can be assigned. Despite the availability of surface clues and gender information in the lexicon, gender mismatch has been reported to account for over a third of all pronoun resolution errors [6].

Ge *et al* were the first to learn the gender of noun phrases from unlabelled text [2]. They applied simple, gender-unaware pronoun resolution algorithms (such as selecting the noun phrase at Hobbs distance one), and collected pronoun-antecedent pairs. Nouns were assigned the gender of whichever pronoun they most often pair with. Gender was correctly attributed in about 70% of the cases, using a group of proper names occurring with designators as the test set.

The majority of anaphora resolution approaches involve some form of manual involvement [9]. Ge *et al* use a manually-parsed corpus [2], while Lappin and Leass manually correct parser output [7]. We parse the text and perform noun-phrase identification fully automatically. Lappin and Leass have a module for automatic identification of pleonastic pronouns [7], such as the non-anaphoric "it" in "it is raining," while Kennedy and Boguraev manually identify and exclude these pronouns, as well as those that refer to verb phrases or propositions [6]. We also manually identify and exclude pleonastic pronouns, and those that do not refer to preceding noun phrases, including cataphoric pronouns (pronouns occurring *before* their antecedents). Of the 2779 total pronouns labelled, 144 are pleonastic, 59 do not refer to an explicit noun phrase, and 16 are cataphora. Results stated below do not include these excluded cases.

3 Pattern Matching for Noun-Pronoun Gender Pairs

Principle A of Government and Binding Theory states that a reflexive pronoun must be bound by an antecedent in its governing category [3]. For example, after seeing "John explained himself," we know that *John* binds *himself*, and hence is masculine. To determine the probability that a given word is masculine, feminine, neutral or plural, we might examine a large amount of text and count the number of times it binds with masculine, feminine, neutral or plural reflexives.

3.1 Parsed Corpus Frequencies

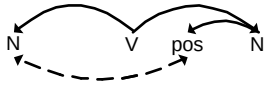
We need not restrict ourselves to the reflexive indicators. We can find other lexico-syntactic patterns in text that typically represent a bound noun and pronoun. On average, "John explained his..." would be more common than "John explained her," "John explained its" or "John explained their." In fact, if we let any verb occupy the place of "explained," then we have a generic pattern with which we can count the number of noun-pronoun occurrences for any noun in text. We collect gender information in corpora via five separate lexico-syntactic

noun-pronoun patterns depicted below. In the following representation of dependency trees, the solid arrows indicate dependency relationships, while the broken arrows connect the noun-pronoun pair extracted by our system:

1. Reflexives (*himself, herself, itself, themselves*):

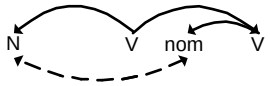
E.g. *John* explained *himself...*

2. Possessives (*his, her, its, their*):



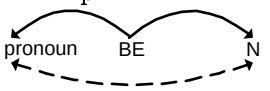
E.g. *John* bought *his* car...

3. Nominatives in *finite* sub-clauses (*he, she, it, they*):



E.g. *John* thought *he* should...

4. Predicates: pronouns are subjects and nouns are in the predicate position:



E.g. *He* is a *father.*

5. Designators: The noun is accompanied by a gendered designator:

E.g. *Mr. Johnson.*

Some noise will be present in these pairs – parser errors, ungrammatical text and false bindings will pollute the results substantially. Thus a large amount of text is required for the correct gender probabilities to prevail.

3.2 Web Frequencies

No matter how large a corpus one has mined, there is always a strong probability one will encounter new words when looking at test data. A growing number of researchers address this problem by using the world wide web as a vast source of example data [5]. We use web page counts of various noun-pronoun patterns to get wide coverage for our gender learning strategy. The Google API and wildcard operator, "*", equal to a single word, are employed. Gender determination is based on the same ideas as were used in the corpus pair searching. Now, we count the number of *pages* returned for the following query patterns:

1. Reflexives: *himself, herself, itself,* and *themselves* in "*noun * reflexive*"
2. Possessives: *his, her, its,* and *their* in "*noun * possessive*"
3. Nominatives: *he, she, it,* and *they* in "*noun * nominative*"
4. Predicates: *he, she, it,* and *they* in "*nominative* is/are [a] *noun*"
5. Designators: *Mr.* and *Mrs.* in "*designator noun*"

A substantial amount of noise also afflicts the web approach. The Google queries do not restrict the wildcard to be a verb, nor the entire query string to be in the same sentence. Also, because our pairings tend to identify nouns in subject positions, we obtain limited data for nouns preferring object positions.

3.3 Modelling Gender Information

For each word, we seek to determine the probability that this word is masculine, feminine, neutral or plural, using the counts from each of the five parsed corpus templates and five web-mining templates. In this sense, the proportion of times a word is a given gender is a parameter we seek to learn from our frequency data. For a given word and gender, each of the ten templates can yield a probability value. The maximum likelihood formulation is to say that the probability the gender of a word is, for example, masculine, is equal to the number of times that word occurs with masculine pronouns in a given template, divided by the number of times it occurs with pronouns of all genders in that template. For example, we captured the parsed corpus pairs *doctor-himself* 224 times, *doctor-herself* 126 times, *doctor-itself* 0 times, and *doctor-themselves* 14 times. Thus our parsed-corpus reflexives indicate there is a 62% chance of doctor being masculine.

There are two issues with the above approach. First small counts will result in large probability swings. Second, we need a measure of how certain we should be in the resulting probabilities – e.g., a 60% chance of being masculine should be taken more seriously when we have five hundred pairs than when we have five. We address these issues by adopting a Bayesian approach. In Bayesian parameter learning, an hypothesis prior distribution is assumed for the parameter, and this distribution is updated as new information is available [11]. We initially assume any value for the gender probability is equally likely. Hence we begin with a uniform prior distribution for the parameter. Subsequently, we treat this prior distribution as the first prior in a family of *Beta* distributions. A *Beta* distribution models binomial proportions in Bayesian analysis. For a given gender, we treat the pair counts as binomial in that all pairs of that gender are treated as one event, while any pair not of that gender is considered a separate event. The *Beta* distribution depends on two hyperparameters, α and β, where $\alpha - 1$ and $\beta - 1$ are the number of times each type of event is observed.

An example will illustrate the procedure. Suppose we're determining whether the word "gretzky" can be replaced with a masculine pronoun. We build the *Beta* distribution for each of the ten gender sources. In our parsed corpus-possessive pairs, for example, we've seen the pair *gretzky-his* 4650 times, the pair *gretzky-her* 0 times, the pair *gretzky-its* 54 times and the pair *gretzky-their* 40 times. The *Beta* distribution that models masculine probability treats 4650 as the number of times masculine has been observed, and $0 + 54 + 40 = 94$ as the number of times non-masculine was seen. The *Beta* distribution is thus *Beta*[4651,95]. The mean, μ, of the *Beta* distribution is given as:

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{1}$$

Here, the mean represents the probability that the word "gretzky," is masculine; it equals 98.0%. The variance of a *Beta* distribution is:

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{2}$$

The variance quantifies how much the probability mass is converging around a single value in the parameter's distribution. Our "gretzky" example has negligible variance, approximately $4.1 * 10^{-6}$. Gender sources with little or no example counts have variance approaching $1/12$ – the variance of a uniform distribution.[2]

The above approach gives the determination of the gender probability and distribution variance for one gender of one of the ten templates. In practice, we need to simultaneously combine each of these sources of information into a single determination of gender (Section 5), or decide whether to exclude a candidate antecedent based on gender mismatch (Section 6). For these tasks, we assign the probability and variance values to dimensions in a gender *feature vector*, and use Support Vector Machine (SVM) learning to determine the optimum use of these values. Sections 5 and 6 describe the experiments and results.

4 The Data Sets

For the acquisition of the gender information, we collected noun-pronoun patterns from a number of corpora, including the AQUAINT corpus and the Reuters corpus. Together, the full set contains about 6 gigabytes of text. We extracted over 4 million reflexive pairs, 32 million possessives, 28 million nominatives, 5 million predicates, and 17 million words with gendered designators.

For the training data and separate testing data used by both the gender classifier described in Section 5 and the full anaphora resolution system described in Section 6, annotated data is required. We labelled third person pronoun-antecedent pairs in 118 documents from the slate section of the American National Corpus.[3] There are 1398 labelled pronouns in 79 documents in the training set and 1381 labelled pronouns in 41 documents in the test set (including non-anaphoric cases (Section 2)). Most of the slate documents are "gist" articles which provide factual background information for stories currently in the news.

Having a set of labelled pronouns also gives us a set of gender-marked nouns: the antecedents of each pronoun have the same gender, in that particular context, as the pronoun which refers to it. Filtering out pronouns which refer to other pronouns and the ignored cases mentioned above, and labelling each antecedent with the gender of its pronoun, we are left with 903 gendered noun phrases from the training section and 876 gendered noun phrases from the test section. These are extracted from the text; they form the lists of noun/gender pairs used for training and testing in Section 5. Roughly 24% of the nouns in the lists are masculine, 7.6% feminine, 33.8% neutral and 34.6% plural.

[2] Note that the information from the beta distributions is equivalent to using pure count statistics as features (the maximum likelihood estimation), except with add-one smoothing and the confidence measure from the variance. Neglecting smoothing or the confidence measure results in small but consistent decreases in performance.

[3] Instructions for obtaining the American National Corpus and our anaphora resolution labels are available at http://www.cs.ualberta.ca/~bergsma/CorefTags/

5 Testing Gender Classification

To test the accuracy of our gender statistics apart from their use in anaphora resolution, we built separate SVM classifiers for masculine, feminine, neutral and plural nouns using features derived from the parsed corpus and web-based data. Each classifier inspects a noun and decides whether it matches the classifier's gender. We used the SVM implementation SVM^{light} with a linear kernel, without normalization [4]. SVM is used because an efficient implementation is available, it easily incorporates the continuous-valued gender features, and it has been shown to provide good performance on various machine learning tasks [4]. For each of the parsed corpus and web-based *Beta* distributions, we include the mean (corresponding to the probability the word matches a given gender) and standard deviation (the certainty in the probability) of the distribution as features for the SVM. We trained the SVM on the pronoun-derived gendered nouns in the training list and tested it on the gendered nouns in the test list.

Note that the genders of the words in the noun lists do not represent the most likely genders or the "true" genders of the given nouns, but merely the genders of the nouns in this particular context. Contradictory instances are present: lawyer is masculine in one article, feminine in another. Nevertheless, the ability of our system to predict the gender of these nouns is a good indication of how adept our system would be in predicting gender for anaphora resolution.

Table 1. Gender Classification Performance Using All Features (%)

	Precision	Recall	F-Score
masculine	88.2	95.2	91.6
feminine	98.2	70.9	82.4
neutral	93.0	93.7	93.3
plural	98.6	89.8	94.0
micro-avg.	93.9	90.6	92.2

The entire set of words was classified once with each classifier, and resulting precision, recall and f-score were calculated (Table 1). Overall the classifiers performed quite well, correctly deciding gender match in 96% of the instances, with a micro-averaged f-score of 92%. Next we determined whether the web-mined or corpus-based sources contributed more to the overall performance. We trained and tested the classification using only the information from the corpus *Beta* distributions and then with only the web-based *Beta* distributions (Table 2). The web-based approach, with a micro-averaged f-score of 90.4%, outperforms the corpus-features, which score 85.4%. As expected, the corpus approach suffered mostly in recall, as many words have few or no instances in the corpus-derived gender pairs, leading to a number of false negatives. It also performs worse, perhaps surprisingly, in precision. We see that the larger

coverage of web-based information extraction more than compensates for the greater noise in this task. It is also interesting to note how values from the web and parsed corpus templates work together in the combined classifier (Table 1). There is no need to choose between web-based or corpus-based information, but instead one may combine the information from each to achieve superior performance.

Table 2. Micro-averaged Performance for Various Classifiers (%)

	Precision	Recall	F-Score
Parsed-Corpus Features	90.9	80.6	85.4
Web-Mined Features	92.4	88.6	90.4
Average Human Performance	88.8	88.8	88.8

To provide further perspective on our results, we asked three native English-speaking graduate students to classify gender on the same list of words. Like our classifier, the students made their decisions blindly without any context from the articles; they were simply asked to assign whatever gender they thought was most likely for each given noun from the noun list. To mitigate any systematic effects, the lists were randomized for the tests. The students achieved micro-averaged scores of 86.6, 88.5, and 91.3, respectively, with average results calculated and tabulated (Table 2). It is interesting that no human performed as well as our full-featured system. We must conclude that humans achieve their near-perfect performance on general pronoun resolution through contextual clues and other techniques, and not through explicit *a priori* noun gender knowledge.

All approaches had low recall with the female classifier. This is because many nouns are most often masculine, and thus appearances of these tokens in a female context is missed as female, providing a false positive to our masculine classifier (reducing precision) and a false negative to our feminine one (reducing recall).

6 Pronoun Resolution with Enhanced Gender

Although knowing the gender of a given word may have interesting applications on its own (e.g. for lexicon development), we are ultimately interested in whether our gender information improves the performance of an anaphora resolution system. We tested this by designing pronoun resolution classifiers of varying complexity and assessing the gain in performance when using gender statistics. The results are summarized in Table 3 and explained in the following subsections.

The general pronoun resolution approach is as follows: for each pronoun, we search backward in the text from this pronoun to previous noun phrases, rejecting ones judged not to match and stopping when we reach one judged to be the true antecedent. The system has correctly identified the antecedent if it

eventually accepts a coreferent noun phrase. A correctly accepted noun phrase is not necessarily the most recent antecedent; more than one antecedent is possible in cases where multiple preceding noun phrases corefer with the pronoun.

Also, we search backwards for an antecedent only through the current and previous sentence in the text. If, by the beginning of the previous sentence we have not yet accepted an antecedent, we lower the threshold for classification in the SVM (i.e., decrease the signed distance from the hyperplane separating positive (coreferent) and negative (non-coreferent) feature vectors), and begin our search again at the most recent noun phrase. We repeat the searching and threshold-lowering until some noun phrase has been accepted as antecedent. This modification is motivated by the observation that over 97% of anaphoric pronouns in our training set had an antecedent in the current or previous sentence.

Table 3. Pronoun Resolution Performance

Method: Set Antecedent to Most Recent NP...	Correct	Incorr.	Rate(%)
Baseline	336	954	26.0
Without Gen. Mismatch	397	893	30.8
Without Gen. Mismatch, Accepted by SVM Gen. Classifier	766	524	59.4
Accepted by SVM Classifier (minus new gender features)	815	475	63.2
Accepted by Full SVM Classifier (with new gender features)	946	344	73.3

6.1 Baseline Anaphora Resolution

The simplest baseline strategy is to always choose the previous noun phrase. This achieves an accuracy of 26.0% (Table 3). We added to this baseline system in two ways. First, we adopted the standard gender approach using only explicit surface clues, rejecting matches where the gender is known and it does not agree with the pronoun. This improved performance to 30.8% (Table 3). Our acquired gender information was then incorporated. Given the gender of a pronoun, we use the corresponding classifier built for the gender classification task in Section 5. We reject previous noun phrases that either mismatch in known gender (standard approach) or the corresponding gender classifier gender, until a match is obtained. This nearly doubles performance, to 59.4%, strikingly illustrating the immediate benefit of our new gender information.

6.2 Robust Anaphora Resolution

To further improve performance, we developed a machine-learned anaphora resolution system based on a number of syntactic and semantic features, including features based on the *Beta* distributions of our gender sources (Table 4). Features are collected after tokenizing, parsing, and linking nouns in the text. Linking

Table 4. Features for Pronoun Resolution

Type	Feature	Description
Pronoun Features	Masculine	1: pronoun masculine; else 0
	Feminine	1: pronoun feminine; else 0
	Neutral	1: pronoun neutral; else 0
	Plural	1: pronoun plural; else 0
Antecedent Features	Antecedent Frequency	Number of Occurrences / 10.0
	Subject	1: subject of clause; else 0
	Object	1: object of clause; else 0
	Predicate	1: predicate of clause; else 0
	Pronominal	1: pronoun; else 0
	Prepositional	1: prepositional complement; else 0
	Head-Word Emphasis	1: parent not noun; else 0
	Conjunction	1: *not* part of conjunction; else 0
	Prenominal modifier	1: noun is a prenominal modifier; else 0
	Org	1: an organization; else 0
	Person	1: a person; else 0
	Time	1: has time units; else 0
	Date	1: a date; else 0
	Money	1: a monetary denomination; else 0
	Price	1: a price; else 0
	Amount	1: ante has measurement units; else 0
	Number	1: number; else 0
	Definite	1: has definite article; else 0
	His/Her	1: ante first word of his/her pattern; else 0
	He/His	1: ante first word of he/his pattern; else 0
Gender Features	Std. Gender Match	1: gender known and matches; else 0
	Std. Gender Mismatch	0 if gender known and mismatches; else 1
	Pronoun Mismatch	0 if both pronouns and mismatch; else 1
	Web/Corpus Genders	mean/std. dev. of *Beta* distributions (20X)
Pronoun-Antecedent Features	Binding Theory	1: satisfies Principles B,C; else 0
	Reflexive Subj. Match	1: ante subj. of reflexive pron's GC; else 0
	Same Sentence	1: ante/pron in same sentence; else 0
	Intra-Sentence Diff.	Within-sentence difference/50.0
	In Previous Sentence	1: ante in previous sentence; else 0
	Inter-Sentence Diff.	Sentence distance/50.0
	Prepositional Parallel	1: ante/pron objs. of same preposition; else 0
	Relation-Match	1: ante/pron have same gramm. rel.; else 0
	Parent Relation Match	1: parents have same gramm. rel.; else 0
	Parent Cat. Match	1: parents have same gramm. category; else 0
	Parent Word Match	1: parents same word; else 0
	Quotation Situation	1: ante/pron both in/out of quotes; else 0
	Singular Match	1: both singular; else 0
	Plural Match	1: both plural; else 0
	MI Value	Mutual Information between ante and pron
	MI Available	1: MI value available; else 0

nouns with matching strings enables us to count noun occurrences and send any gender information learned in one instance of the word to all other occurrences of that word in the chain. To create the training set, we adopt the procedure of Soon *et al* ([12]). Each pronoun and its closest preceding antecedent in our labelled set of training documents form a pairwise positive instance in the set of training vectors. All intervening noun phrases (between the antecedent and the pronoun) form pairwise negative instances with the pronoun. This forms a training set with 1251 positive examples and 2909 negative examples. We train the SVM on this training set, and apply it backward incrementally from each pronoun in the test set until a pronoun-antecedent match is accepted.

Using the full feature set, with all the gender information included, yields a performance of 73.3% on the test data (Table 3), much higher than any of the baseline approaches. We assessed the contribution of the new gender information to this performance by removing the modelled gender features (but including the features for the standard gender approach) and observed a performance of 63.2%. The 10% performance gain obtained by using the new sources of gender information again indicates the clear and immediate benefit of our work.

A pronoun resolution system performing at 73% on a set of challenging news articles provides a good base for future work. Comparison to other systems in the literature is difficult; different data sets are used and different kinds of manual intervention are performed. Kennedy and Boguraev achieved 75% performance on 306 anaphoric pronouns taken from a variety of texts, including news articles [6], while Mitkov *et al* reach 62% on 2263 anaphoric pronouns (excluding pleonastic pronouns) [9].

7 Conclusion

We have proposed a new approach to anaphora resolution using improved gender information. To the best of our knowledge, our system encompasses the broadest and most accurate gender information yet obtained for anaphora resolution and is the first to obtain gender information from web mining. The enhanced gender information is used in a classifier that outperforms humans at gender guessing tasks, and results in significant performance improvements when used as part of either simple or knowledge-rich anaphora resolution systems.

The quality of the obtained gender information depends on many factors. Parsing errors can lead to erroneous noun-pronoun pairings in the parsed corpus templates. Also, the size and variety of the corpora affect the available gender information. For the web-mined pairs, the more data on the web, the more accurate the resulting values. Thus with improved parsers, more data, and the continued growth of the world wide web, there is potential for improved performance using our method.

We will next focus on improving the anaphora resolution system itself. We will investigate new features for pronoun classification, and new approaches to employing the existing features. Modules to detect pleonastic pronouns and re-

solve cataphora are in development. Ultimately, we will incorporate anaphora resolution into a Question Answering system currently under development.

Acknowledgements

Thanks to Dekang Lin and all members of the Natural Language Processing Group at the University of Alberta. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

References

1. Richard Evans and Constantin Orăsan. Improving anaphora resolution by identifying animate entities in texts. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference*, pages 154–162, 2000.
2. Niyu Ge, John Hale, and Eugene Charniak. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, 1998.
3. Liliane Haegeman. *Introduction to Government & Binding theory: Second Edition.* Basil Blackwell, Cambridge, UK, 1994.
4. Thorsten Joachims. Making large-scale SVM learning practical. In B. Schölkopf and C. Burges, editors, *Advances in Kernel Methods.* MIT-Press, 1999.
5. Frank Keller, Maria Lapata, and Olga Ourioupina. Using the web to overcome data sparseness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 230–237, 2002.
6. Christopher Kennedy and Branimir Boguraev. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 113–118, 1996.
7. Shalom Lappin and Herbert J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.
8. Dekang Lin. Dependency-based evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*, 1998.
9. Ruslan Mitkov, Richard Evans, and Constantin Orasan. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 168–186, 2002.
10. Natalia Modjeska, Katja Markert, and Malvina Nissim. Using the web in machine learning for other-anaphora resolution. In *Proceedings of the EACL Workshop on the Computational Treatment of anaphora*, pages 39–46, 2003.
11. Stuart J. Russell and Peter Norvig. *Artificial Intelligence: a modern approach*, chapter 20: Statistical Learning Methods, page 720. Prentice Hall, Upper Saddle River, N.J., 2nd edition edition, 2003.
12. Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.