A Large Margin Approach to Anaphora Resolution for Neuroscience Knowledge Discovery

I. Burak Ozyurt
Department of Psychiatry, UCSD
9500 Gilman Drive MC 0855
La Jolla, CA 92093-0855
email: iozyurt@ucsd.edu

Abstract

A discriminative large margin classifier based approach to anaphora resolution for neuroscience abstracts is presented. The system employs both syntactic and semantic features. A support vector machine based word sense disambiguation method combining evidence from three methods, that use WordNet and Wikipedia, is also introduced and used for semantic features. The support vector machine anaphora resolution classifier with probabilistic outputs achieved almost four-fold improvement in accuracy over the baseline method.

Introduction

Knowledge-base construction of a semantic search engine involves NLP tasks ranging from semantic parsing for information extraction to anaphora and coreference resolution. Anaphora resolution is defined as the process of determining the antecedent of an anaphor. The antecedent can be located within the same sentence or in any of the previous sentences in the context. It can even be not instantialized or implicit in the meaning of a portion or whole of a sentence prior the anaphor. In this paper, only the antecedents with explicit instantiations as noun phrases (NP) are tackled. Anaphora resolution in information extraction is an important part of the coreference resolution task used to merge partial data objects about the same entities, entity relationships and events described at different discourse positions (Mitkov 2003). Even without full-blown coreference resolution, anaphora resolution by itself is very useful in transforming semantic roles extracted from sentences into self-contained information pieces by expanding any pronouns occurring with their antecedents.

Scientific abstracts being summaries of research papers can be characterized as usually having more dense, long and complex sentences than the full text of the corresponding papers. The interdisciplinary nature of neuroscience is reflected in the contents of neuroscience abstracts which combine vocabulary from multiple diverse fields ranging from clinical psychology to imaging, from genetics to medical informatics. They also contain a heavy dose of highly specialized named entities such as gene/protein names, organic chemical compounds, disease/symptom, drug, body/brain anatomy and clinical assessment names. With their uncommon punctuation and form these entities pose a challenge for syntactic parsers and common NLP tools.

In this paper, as a crucial component of the ongoing knowledge base construction for a semantic search engine to aid research in neuro-degenerative diseases, a large margin classifier based machine learning approach to anaphora resolution for neuroscience abstract information extraction is introduced. Both syntactic and semantic features are employed. In the following section, the corpus and its preprocessing is described. Recognizing the importance of word sense disambiguation (WSD) for semantic features, a novel evidence combining Support vector machine (SVM) based WSD system is introduced next. The features for classification and anaphora resolution classifier are described afterwards followed by test results, discussion and conclusion.

Datasets and Preprocessing

The first 2,000 abstracts returned from a PubMED (the National Library of Medicine's search service) search returned for the keyword 'schizophrenia' are selected as the corpus for anaphora resolution task.

About 1400 anaphor-antecedent pairs are annotated by the author as found in the body text of these abstracts[1]. Each abstract body is first separated into individual sentences by a sentence boundary detector. The detected sentences are parsed using Charniak's syntactic parser (Charniak 2000), which also provides part-of-speech (POS) tags for the parsed sentences. Both personal/possessive and demonstrative pronouns are considered. In these biomedical abstracts, there were no occurrences of second person pronouns, and third person pronouns like 'he', 'she' and their variants were very rare and hence not annotated. Most of the third person pronouns were variants of 'it' and 'they'. There were some first person pronouns, all of them being 'we' and they all implicitly corresponded to the authors of the abstract. In total, seven pronouns, namely 'it' , 'its' , 'this' , 'their' , 'them' , 'they' and 'these' with substantial corpus occurrence are used in annotation. The frequency distribution of these pronouns in the annotated corpus is shown in Table 1.

Scientific abstracts, especially in biomedical fields, are laden with acronyms. To facilitate constructing of features

[1] The dataset is available from the author upon request.

Table 1: Pronoun frequency distribution in the annotated corpus

it	they	them	its	this	these	their
9.7%	9%	2.7%	12%	18.8%	23.8%	24%

for antecedent detection, a robust finite state automata (FSA) based acronym detector is devised. This acronym detector works on the combination of the text of a sentence and its syntactic parse tree, first finding the first declaration of the acronym by regular expressions and working towards left in the parse tree to find the expansion of the acronym. The algorithm collects all leaf nodes which are eligible to be in an acronym expansion including nouns, gerunds, certain prepositions and conjunctions and stops at the first ineligible parse tree node. After reversing the order of the collected leaf nodes, it tries to reconstruct the acronym from these nodes. To accomplish this, first it removes all the occurrences of prepositions and conjunctions from the collection, since they are almost never used in constructing acronyms, then splits nodes containing internal dashes. Then, it tries to find if the first letters of a subset of the leaf nodes can be used to construct the acronym. If this fails, it tries to pivot the first letter of the first leaf node with the acronym start. After that, using backtracking, combination of multiple letters from the remaining collection of leaf nodes are tried to reconstruct the acronym. Using this algorithm, expansions for 658 unique acronyms, ranging from organic compound names to anatomical regions, are recognized in this 2000 abstract corpus.

Word Sense Disambiguation

For semantic similarity determination used in features for anaphora resolution classifier, detection of the correct sense of the head noun is essential. To facilitate this, a word sense disambiguator that combines evidence from multiple word sense disambiguation (WSD) approaches is introduced. Three WSD approaches are considered all relying on WordNet (Fellbaum 1998). These approaches are a) dictionary based WSD method using WordNet glosses as dictionary entries; b) Adaptation of Yarowsky's adaptive thesaurus based WSD method (Yarowsky 1992) for WordNet; c) Simple fallback method assigning the most frequent sense for a noun as the selected sense. The evidence from these three WSD approaches are combined by a support vector machine classifier to yield the final estimated sense of the noun. To increase the coverage of the WordNet in biology and medicine specific areas, Wikipedia is used as final resource for WSD. In the following paragraphs, each of these approaches are explained and experimental results are provided.

Dictionary based WSD approach uses the body of the abstract as the context $\{a_i|i = 1, \ldots, N_a\}$ and the short WordNet gloss of each senses for the noun to be disambiguated as the dictionary entry. The score for each WordNet sense is calculated by sense frequency weighted intersection of the noun set from sense glosses S_g with the noun set from the corresponding context S_{a_i}.

$$\text{score} = \sum_{n \in S_{a_i} \cap S_g} \frac{1}{ns_n}$$

where ns_n refers to the number of WordNet senses for noun n. This score function favors matching nouns with a small number of senses over nouns with large number of senses. The more senses a matched noun has, the more chance there is for the matched noun to refer to a different meaning than the meaning conveyed in the context. WordNet adaptation of the adaptive thesaurus based WSD method of Yarowsky (Yarowsky 1992), uses the hypernymy relations in WordNet. First, for each noun lemma of the 2000 abstracts, which has an immediate hypernym synonym set, the first synonym is selected as a possible topic for the abstract. The set of topics is denoted by $\{t_j|j = 1, \ldots, N_t\}$, where N_t is the total number of topics seen. The most prominent topics for each abstract is then selected first scoring each topic t_j in each abstract a_i by

$$\begin{aligned}\text{sc}(a_i, t_j) &= \log\left[\frac{p(a_i|t_j)}{p(a_i)} p(t_j)\right] \\ &= \sum_{n \in a_i} \log p(n|t_j) - \sum_{n \in a_i} \log p(n) + p(t_j)\end{aligned}$$

and then building a topic set for abstract a_i, $ts(a_i)$ by only keeping the topics between 100 to 70 percentile of the score values. The abstract topic set can be defined more formally as $ts(a_i) = \{t_j|sc(a_i, t_j) \geq 0.7 \times \max_{j=1,\ldots,N_t} sc(a_i, t_j)\}$. The noun given topic and topic probabilities are then adjusted using the computed topic sets $ts(a_i)$ by the following maximum likelihood estimates

$$\begin{aligned}p(n_k|t_j) &= \frac{|\text{NS}_k \cap \text{TS}_j|}{\sum_k |\text{NS}_k \cap \text{TS}_j|} \\ p(t_j) &= \frac{\sum_k |\text{NS}_k \cap \text{TS}_j|}{\sum_k \sum_j |\text{NS}_k \cap \text{TS}_j|}\end{aligned}$$

where $\text{NS}_k = \{i|n_k \in a_i\}$ and $\text{TS}_j = \{i|t_j \in ts(a_i)\}$. The probabilities are smoothed by Lidstone's law with $\lambda = 0.5$. The disambiguation for a given noun n in context a_i is done by calculating the scores for each sense s_l and selecting the sense with the largest score.

$$\begin{aligned}\text{sc}(s_l) &= \log p(t_{s_l}) + \sum_{n_k \in a_i} \log p(n_k|t_{s_l}) \\ s^* &= \arg\max_{s_l} \text{sc}(s_l)\end{aligned}$$

Here t_{s_l} denotes the main topic of the sense s_l which is defined as the first item of the hypernym synset for the WordNet sense s_l.

A mechanism is devised to combine evidence from multiple WSD approaches that is expected to perform better than each of the above described methods. Instead of explicitly specifying evidence combination structure and estimating its parameters using cross-validation, both the structure and its parameters are learned from data by casting it as a classification problem. Four features are used to train a SVM classifier with a Gaussian kernel. These features can be enumerated as the method used for the disambiguation, the number

Table 2: WSD Performance Results

Method	Accuracy
Baseline	62%
Dictionary-based	59.5%
Thesaurus-based	36.1%
SVM-based evidence combiner	69.1%

of WordNet senses, the number of ties in the sense score values and the fraction of the total sense score value range covered between the best score and the second best one. From the development set, 205 nouns having more than one WordNet sense are annotated with correct sense and the WSD approaches are tested on it. On average, there was 5.9 senses per word in the annotated set. The SVM classifier based evidence combining approach is tested on a random 40% holdout data set. The results are summarized in Table 2.

Surprisingly, the simplest approach, selecting the most frequent sense, performs better than the other single WSD approaches, dictionary-based method being the close second. Thesaurus based method performed worst since a considerable fraction of the nouns to be disambiguated are not topicalized which is a noted weakness of Yarowsky's method. However, the evidence combination by a SVM classifier performed better than each of the WSD approaches alone by combining the strengths of each method. Hence, SVM evidence combiner based WSD is used in building the features for anaphora resolver introduced in this paper.

While WordNet has grown in coverage over the years, its coverage for biomedical areas is still incomplete. To increase its coverage, if there is no match in WordNet for a noun, the online encyclopedia Wikipedia is used. Different senses for a Wikipedia topic is detected by parsing Wikipedia pages and following disambiguation page and section entries. In Wikipedia, categories, similar to thesaurus topics, are assigned to topics by the authors of the entries. However, the granularity and philosophy behind Wikipedia categories are different than WordNet synset hierarchy, therefore a full-blown alignment of WordNet topic hierarchy with Wikipedia category hierarchy is not attempted for WSD purposes. However, most biomedical terms usually have a single sense and as a last resort, single sense Wikipedia topic categories are used to align a WordNet unknown noun within WordNet hypernym structure. The number of senses for a Wikipedia topic is determined by using Wikipedia web services to retrieve topic page content in MediaWiki format and parsing them to detect disambiguation page indicators plus links and topic categories. The list of categories are then matched against WordNet and the first category matched against a WordNet entry and its most frequent sense, if there is more than one WordNet sense, is used as the sense category. What is aligned here is not the actual sense of the word but the semantic category under which the context sense of the word would be assigned. Once aligned with WordNet, the sense category then can be used in semantic similarity feature for anaphora resolution.

Anaphora Resolution

Anaphora resolution problem is cast as a discriminative classification problem estimating the conditional probability $p(y|\vec{X})$ directly from labeled training data set $\{\vec{X}, y_l\}$. Each antecedent candidate (AC) is represented by a set of its features. The corresponding binary output y represents the classifier's prediction whether the AC is the antecedent or not. The discriminative classifier employed is the most prominent member of the family of large margin classifiers, namely, support vector machine (SVM) which uses empirical risk minimization in selecting the parameters for the classifier that minimizes generalization error on unseen test cases.

The candidates presented to anaphora resolving SVM classifier are selected using Hobbs' pure syntax-based pronoun resolving algorithm (Hobbs 1976). This algorithm is also used to generate candidates and distance-from-anaphor information by Ge et al. (Ge, Hale, and Charniak 1998). Hobbs algorithm searches syntax tree(s) starting from anaphor node in a left-to-right, breadth-first fashion for AC NP nodes taking into account well-established reflexive pronoun constraints. For inter-sentence antecedent-anaphor pairs, Hobbs algorithm searches the parse trees of previous sentences in recency order similarly in a left-to-right, breadth-first manner, preferring subject antecedents. The algorithm relies on a certain parse tree structure with special node type $\bar{N}$, especially useful for reflexive pronouns. Since parse trees from Charniak parser used here does not have this type of tree nodes, Hobbs algorithm implemented differs from the original. In the corpus, reflexive pronouns were almost non-existent, therefore the effect of the implementation difference should be minimal.

Features Used

Eleven features are devised for SVM based anaphora resolution. Each of them is described in detail in the following paragraphs.

While there is no upper bound how far away from its anaphor an antecedent can be (Hobbs 1976), usually, antecedents and their anaphors are found in relative close proximity to each other. The distance (Dist) of an AC from its anaphor is the first feature considered. Two variants of distance feature is taken account, namely *Hobbs distance*, the order of ACs returned by Hobbs algorithm taken as relative distance and *Surface Distance*, relative surface distance of Hobbs algorithm generated candidates from the anaphor.

The second group of syntactic features considered are the head word lemma (Head) for the immediately dominating phrase of the anaphor, and head word lemmas of the antecedent candidates. Using WordNet synonyms, noun and verb head words are clustered. Both individual head words and semantic cluster prototype words are tested as features. In addition, their part-of-speech (POS) tags are also considered as additional features (Head POS).

The third feature group relies on number (Number) agreement grammatical restriction of a pronoun with its antecedent. A three valued feature with values 'singular', 'plural' and 'unknown', is used to encode the number of AC and the anaphor. To determine the number of a candidate, first the head of it is checked if it is morphologically plural by the type of POS tag assigned by the Charniak parser and XTAG morphology utilities. If it is not plural, a semantic plurality check is done using WordNet hypernymy structures to determine if the word can be considered a collective noun which can act as plural. For possessive nouns, the position of the apostrophe is used to determine the number of the phrase. Also, coordinated conjunctions as in in the phrase 'olanzapine, risperidone, and quetiapine' are recognized as plural. If the number of a candidate cannot be determined by this process it is set to 'unknown'.

The fourth feature group considered is another semantic feature namely the animacity (Anim) agreement of the antecedent with the anaphor. This feature is considered especially for demonstrative and possessive pronouns. For these pronouns, the immediately dominating phrase of the anaphor is a NP. The head lemma for this phrase and the one of the AC needs to agree in animacity if they refer to the same entity. The determination of animacity is done using WordNet hypernym structure. In case of multiple word senses, a sense-frequency weighted voting mechanism is used to determine if the head noun is animate or not. If the animacity can not be determined, the feature value is set to 'unknown'.

The fifth feature type, tried is the mention count (Count) for the AC in the context as in (Ge, Hale, and Charniak 1998). Acronyms are also taken into account in mention count determination.

Semantic similarity (Sem) between the dominating parent phrase head for the anaphor and the AC head is another feature introduced. The phrase head word sense for each context is estimated using the evidence combining SVM based WSD system discussed in the previous section. First possible synonymy of the head lemmas are checked using WordNet. If no match has been found at this stage, up to three levels up in the hypernym tree, hypernym synonyms for the AC are matched against the synonym set of the anaphor parent head lemma and vice versa. An upper cap on how high in hypernym tree structure matches will be attempted is put in-place in order to avoid similarity matches which are too generic. If there is no match at this stage also and if either AC or anaphor original sense disambiguation knowledge source is Wikipedia, its Wikipedia-to-WordNet mapping topic gloss nouns are also matched against the WordNet synonyms for other half of the head lemma pair that is semantically compared. If there is no match after this stage, semantic similarity feature is set to 'no'. If for any of the lemma pairs, the word sense cannot disambiguated because of neither WordNet nor Wikipedia match, similarity feature is set to 'unknown'.

Syntactic parallelism (Mitkov 1997) is taken into account by introducing weighted edit distance between syntactic frames of anaphor (sub)sentence and AC (sub)sentence. Syntactic frames are introduced as feature for semantic role labeling (Xue and Palmer 2004). For example, from the flattened partial sub-sentence parse trees for antecedent (anhedonia) and anaphor (it) below, the syntactic frames are defined as 'CUR→VP→NP' and 'CUR→VP→NP', respectively and their edit distance is 0.

> We do know that $[_S$ $[_{NP}$ anhedonia$]$ $[_{AUX}$is $]$ common in $[_{NP}$ schizophrenia$]$ $]$, that $[_S$ $[_{NP}$ it $]$ $[_{AUX}$ has $]$ $[_{NP}$ significant negative consequences$]$ $]$, and that current treatments are insufficient.

The edit distance (ED) is a metric for measuring the amount of distance between two sequences by finding the minimum number of operations necessary to transform one sequence into the other. The transformation operations available are deletion, insertion or substitution of a single element. The edit distance measure is modified by assigning weights to each of these three operations and introducing an offset (penalty) if the anaphor and AC are on on opposite sides of their respective predicates. The weights are determined by a Nelder-Mead simplex direct search optimization method. The optimization tried to maximize the number of correct antecedents selected on the development training set using the weighted edit distance alone at an estimated threshold value. At each optimization step, the threshold above which candidates are not considered is selected by finding the edit distance value with best correct to incorrect antecedent ratio at fixed set of weights. The final set of weights used are 0.167 for deletion, 0.307 for insertion, 0.526 for substitution and 0.203 as the offset. Since substitution of a phrase is more likely to disturb a parallel syntax structure than the deletion of one, the estimated weights make sense.

A string representation of the syntactic parse tree path (Path) from the anaphor to the AC is another feature considered for anaphora resolution. This feature is originally suggested for semantic role labeling (Gildea and Jurafsky 2002). Sequences of same phrase type in the path feature are collapsed to a single one as an attempt to cluster similar paths together.

Other features considered include a boolean feature (BEFORE) indicating that antecedent is before anaphor or after (for cataphors). Another simple boolean feature ON_SAME_SENTENCE (OSS) is used to indicate to the classifier if the candidate is on the same sentence as the anaphor or not. This feature is intended to facilitate the classifier's possible usage of different strategies for intra-sentence and inter-sentence anaphora resolution.

Results and Discussion

From the 1400 annotated anaphor-antecedent pairs 1214 can be aligned with their parse trees. Due to the syntactic parser errors, a mapping between a labeled phrase and syntax tree is not always possible. For inter-sentence anaphor-antecedent pairs the number of preceding sentences of the anaphora considered is fixed to two or the beginning of the abstract whichever comes first. There were also some annotated pairs falling beyond this threshold. The pair count 1214 also reflects this. A random 75%/25% training/testing split is performed. For SVM learning and classification, SVM^{light} (Joachims 1998) package is used. There were 16,158 ACs for 911 anaphor-antecedent pairs in the training set (on average 18 ACs/pair). Both polynomial and radial basis function (RBF) kernels were tried. Since RBF kernel

SVMs performed significantly better than polynomial kernel ones, RBF kernels are used for all the results reported. SVM classifier parameters, trade-off between training error and margin C, training error cost factor CF and RBF kernel smoothing factor γ, were optimized within the grid $[0.01, 0.1, 0.5, 1.0, 10] \times [1, 5, 7] \times [0.01, 0.05, 0.1, 0.5, 1, 5]$, using 25% of the training data as holdout test set. The parameter combination with best F_1 performance value on the holdout test set, namely $C = 0.5, CF = 7, \gamma = 0.1$, was used for the reported results.

To establish a baseline, as in (Ge, Hale, and Charniak 1998), a simple algorithm that always takes the last mentioned NP before the anaphor as the antecedent, yielded an accuracy of 15.7%. Compared to the 43% accuracy on the Wall Street Journal corpus baseline method, the baseline performance is almost three times lower, indicating the difficulty of the anaphora resolution in neuroscience abstracts domain. Table 3 summarizes the experimental results. The base feature set included the pronoun itself, BEFORE, OSS and Head. The contributions of each remaining individual feature on the anaphora resolution performance are listed in Table 3. The results indicate that Head Pos, Number and Sem features have improved performance over base feature set both on F_1 and Accuracy better than the others. Hobbs distance performed better than the surface distance in this corpus. The best combination of features yielded F_1 of 72.4% and overall accuracy of 56.8%. The 'best set' consists of the features; the pronoun itself, BEFORE, OSS, semantically clustered Head, Head POS, Hobbs Dist, Number, Anim, Sem and Count.

The SVM classifier is also configured to generate probabilistic outputs using Platt's method (Platt 2000) by fitting a sigmoid to its output. This approach can be seen as recalibration of the SVM classifier outputs. The parameters of the sigmoid used for SVM output to probability mapping is trained by using three-fold cross-validation on the 25% of the training data as holdout set. The AC with largest probability is selected as the antecedent from the AC set of each test case. This approach yielded 62.7% accuracy on the test set.

Besides best-first AC selection strategy employed here, one can also use closest-first strategy as in (Soon, Ng, and Lim 2001), where from the most likely candidates based on the classifier output, the one closest to the pronoun is selected. Two versions of this strategy is applied to the 'best set' trained SVM anaphora resolver to investigate the effectiveness of the closest-first strategy. The first one did not take into account the relative similarity of the scores/probabilities for the most likely candidates and resulted in large test accuracy decrease (40.3% without probabilistic outputs and 49.8% with probabilistic outputs). The second version only applied the closest-first strategy if the decreasing order sorted scores/probabilities for the most likely candidates are within 20% of each other, which also resulted in a slight decrease in the test accuracy (56.1%/61.7% without/with probabilistic outputs).

Relative to the Ge et al. (Ge, Hale, and Charniak 1998), approach developed and tested on Wall Street Journal articles with perfect Penn Treebank parse trees, achieving 84.2% accuracy, the achieved 62.7% accuracy seems low, however taking into account the baseline method performance of 43% on Ge et al. corpus vs. 15.7% on our corpus, use of perfect human generated parse trees versus syntactic parser generated parse trees and much larger training set (2230 vs. 911), the performance improvement over the baseline method is very encouraging.

Table 3: Anaphora Resolution Results[a]

Feature Set	**P**	**R**	**F_1**	**Accuracy**
Baseline	-	-	-	15.7
Base	53.7	52.9	53.3	36.3
Base + Hobbs Dist	53	57.2	55.0	38.0
Base + Surface Dist	48.1	53.4	50.8	34.0
Base + Head Pos	53.2	63.9	58.1	40.9
Base + Number	53.7	64.1	58.4	41.3
Base + Anim	54.3	50.2	52.2	35.3
Base + Sem	52.7	76.1	62.3	45.2
Base + ED	57	52.5	54.7	37.6
Base + Path	57	50.0	53.3	36.3
Base + Count	53.7	52.8	53.3	36.3
Best Set	69.1	76.1	72.4	56.8
Using probabilistic outputs				
Best Set	-	-	-	62.7

[a] P: Precision, R: Recall

Related Work

Corpus based approaches for anaphora resolution fall mostly into two categories; Rule based approaches (Lappin and Leass 1994; Mitkov 1997) and statistical/machine learning approaches (Ge, Hale, and Charniak 1998; Soon, Ng, and Lim 2001). Some of the features used in classification are taken from (Ge, Hale, and Charniak 1998). Our approach differs from the generative approach of Ge et al. (Ge, Hale, and Charniak 1998), where the joint probability $p(\vec{X}, y)$ is modeled by making independence assumptions, in using a discriminative approach modeling the conditional probability $p(y|\vec{X})$ without any assumptions on the relationship between inputs $\vec{X}$ and the outputs y that also minimizes the risk of generalization error. Demonstrative pronouns are also not considered in their case. The approach of Soon et al. (Soon, Ng, and Lim 2001) is based on decision trees, a greedy information theoretic discriminative machine learning approach, which does not take generalization error risk minimization as constraint like our SVM approach. Recently, a generative approach (Gasperin and Briscoe 2008) based on an extended Naive Bayes classifier is introduced and applied to animal biology texts.

Conclusion and Future Directions

In this paper, a large margin classifier based approach to anaphora resolution is presented, with about four-fold improvement over baseline method accuracy, of 62.7%. For robust semantic features a SVM classifier based WSD system combining evidence from three separate subordinate WSD approaches using WordNet and Wikipedia as thesaurus and dictionary is also introduced.

From looking at the errors made by the anaphora resolver introduced, it is clear that the consideration of part-of (meronymy) semantic relationships will be beneficial. The meronymy relationships encoded in the WordNet are too coarse to be useful, however. In the future, domain-specific ontologies with rich-set of semantic entity relationships will be tried. Also there was a significant amount of implicit antecedents not materialized as a noun phrase in the corpus. An approach to transform the implicit antecedent into an explicit antecedent needs to be investigated.

Acknowledgments

This research was supported by 1 U24 RR021992 to the Function Biomedical Informatics Research Network (BIRN, http://www.nbirn.net), that is funded by the National Center for Research Resources (NCRR) at the National Institutes of Health (NIH).

References

Charniak, E. 2000. A maximum-entropy-inspired parser. In *Proceedings of NAACL*, 132–139.

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

Gasperin, C., and Briscoe, T. 2008. Statistical anaphora resolution in biomedical texts. In *Proceedings of COLING 2008*.

Ge, N.; Hale, J.; and Charniak, E. 1998. A statistical approach to anaphora resolution. In *In Proceedings of the Sixth Workshop on Very Large Corpora*, 161–170.

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

Hobbs, J. R. 1976. Pronoun resolution. Technical Report 76-1, City College, New York.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning* 137–142.

Lappin, S., and Leass, H. J. 1994. An algorithm for pronomial anaphora resolution. *Computational Linguistics* 535–561.

Mitkov, R. 1997. Factors in anaphora resolution: they are not the only things that matter. a case study based on two different approaches. In *Proceedings of the ACL'97/EACL '97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*.

Mitkov, R. 2003. *The Oxford Handbook of Computational Linguistics*. Oxford University Press, NY, USA. chapter Anaphora Resolution, 266–283.

Platt, J. 2000. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In Smola, A.; Bartlett, P.; Schoelkopf, B.; and Schuurmans, D., eds., *Advances in Large Margin Classifiers*, 61–74.

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

Xue, N., and Palmer, M. 2004. Calibrating features for semantic role labeling. In *Proceedings of EMNLP-2004*.

Yarowsky, D. 1992. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *COLING 14*, 454–460.