Exploring Domain Differences for the Design of Pronoun Resolution Systems for Biomedical Text

Ngan L.T. Nguyen **Jin-Dong Kim**
Department of Computer Science, University of Tokyo, Hongo 7-3-1, Tokyo, Japan
{nltngan, jdkim}@is.s.u-tokyo.ac.jp

Abstract

Much effort in the research community has been spent on solving the anaphora resolution or pronoun resolution problem, and in particular for news texts. In order to selectively inherit the previous works and solve the same problem for a new domain, we carried out a comparative study with three different corpora: MUC, ACE for the news texts, and GENIA for bio-medical papers. Our corpus analysis and experimental results show the significant differences in the use of pronouns in the two domains, thus by properly considering the characteristics of a domain, we can improve the performance of pronoun resolution for that domain.

1 Introduction

Pronoun resolution is the task of determining the **antecedent** of an **anaphoric** pronoun, or a pronoun pointing back to some previously mentioned item in a text. For example, in the sentence, *"**The IL-2 gene** displays both T cell specific and inducible expression: **it** is only expressed in CD4+ T cells after antigenic or mitogenic stimulation,"* the pronoun *"it"* should be resolved to refer to *"the IL-2 gene,"* and thus, we have an **anaphora link**.

Pronoun resolution is an important task in the family of reference resolution tasks, including anaphora resolution and co-reference resolution, which are known as significant parts of text understanding systems. Recently the need to have more powerful information extraction systems for biomedical technical papers has motivated researchers to solve the same task for the biomedical domain. Castano (Castano et al., 2002) resolved the sortal and pronominal anaphora, by using a salience measure, which is the sum of all feature scores. Kim and Park (Kim and C.Park, 2004) introduced BioAR, a biomedical anaphora resolution system that relates entity mentions in text with their corresponding Swiss-Prot entries. This system resolves anaphoric pronouns by using heuristic rules and seven patterns for parallelism. However, the sizes of the data sets used in their experiments were small. In the former system, 46 and 54 MEDLINE abstracts were used for the development set, and the test set respectively, and the test set in the latter work contained only sixteen anaphoric pronouns. Contrary to their work, in this work we made use of GENIA, a large co-reference annotated corpus for the bio domain, containing 1999 MEDLINE abstracts.

While there are quite a few works on this task for the bio-medical domain, for other domains, and especially for the news domain, a myriad of works on pronoun resolution has been carried out by the NLP researchers (Mitkov, 2002). Since Soon (Soon et al., 2001) started the trend of using the machine learning approach by using a binary classifier in a pairwise manner for solving co-reference resolution problem, many machine learning-based systems have been built, using both supervised and, unsupervised learning methods (Haghighi and Klein, 2007). Such methods were claimed to be comparable with traditional methods. However, the problems caused by domain differences, which strongly affect a deep-semantics related task like pronoun resolution, have not yet been studied well enough.

In order to recognize the important factors in

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 625–632
Manchester, August 2008

building an effective machine learning-based pronoun resolution system, and in particular for the bio-domain, we have built a machine learning-based pronoun resolver and observed the contributions of different features in the pronoun resolution process. In our experiments for the news domain, we used the MUC-7 and ACE corpora, and for the biomedical domain, we employed the GENIA co-reference corpus.

Section 2 describes the noticeable issues related to the corpora, and their preprocessing. Section 3 describes the implementation of our pronoun resolution system, including the resolution model and the features used. Our experiment settings, evaluation scheme, and experimental results are presented in Section 4. Finally, we conclude our paper in Section 5.

2 Corpora

In this section, we briefly introduce three corpora used in our experiments: MUC-7, ACE, and GENIA, and discuss the differences in their annotation schemes. Afterwards, we analyzed the major differences in the distributions of anaphoric pronouns in these data sets, which provide important information for the design of features used in the pronoun resolution process.

The MUC-7 co-reference corpus is a collection of news wire articles from the source for North American News Text Corpora. It contains the *training*, *dry run* test, and *formal run* test sets. The dry run and formal run have different domains; the dry run (and training) consists of aircrash scenarios, while the formal run consists of missile launch scenarios. The ACE (phase 2) corpus for named entity detection contains three data sets: news wire (NWIRE), broadcast news (BNEWS), and newspaper (NPAPER). Each data set is divided into 2 parts for training (*train*), and for development testing (*devtest*). For the bio-domain, we use the GENIA co-reference corpus, containing 1999 abstracts selected from MEDLINE: a huge source of bio-domain scientific papers.

These three corpora are all manually annotated with co-reference information; i.e., the information where mentions refer to the same entities. However, since the annotation schemes used are not the same, these corpora contain some significant differences, which may affect our reference resolution systems.

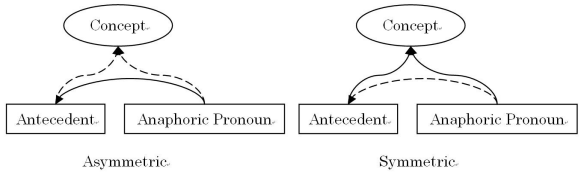

Figure 1: The symmetric and asymmetric annotation schemes. The dotted lines represent implicit links between the elements.

2.1 Variations in co-reference annotation schemes

We started by introducing some important terminologies together with some noticeable issues related to the common co-reference annotation scheme. Later, we mention the differences among the annotation schemes of the three corpora used in our experiments.

There are three main elements in the co-reference corpus annotation: the anaphoric expressions, which are anaphoric pronouns in the case of the pronominal anaphora, their antecedents, and the referred concepts. Depending on either the *asymmetric scheme* employed in MUC (Lynette, 1997) and GENIA (Hong, 2004) or the *symmetric scheme* in ACE (NIST, 2003), the annotation task is defined as either an anaphor-antecedent linking, or mention-concept linking task, correspondingly (See Figure 1). Moreover, each annotation scheme provides its own guidelines for recognizing and annotating these three elements, causing the variations across different co-reference annotated corpora.

In the annotation schemes, mentions which may join in the co-reference relationship are called *markable*. All of the three annotation schemes record both a maximal and a minimal boundary of markables, in concerning the evaluation schemes. However, the types of markables to be annotated, and the ways to decide their maximal boundary, are not the same in every annotation scheme.

Table 1 shows the concepts annotated for each corpora according to the annotation schemes. While the number of concepts in the ACE corpus is limited to only 5 entity types, the GENIA and MUC annotation schemes do not clearly specify the concept types. This means that every possible concept in the text domains can join the anaphora relations; i.e., can be annotated as markables. This in turn makes the resolution task become more difficult.

Table 1: Possible concepts according to the annotation schemes

GENIA	ACE	MUC
(Not specified explicitly) -Bio entities	5 types of entities -Person -Organization -Facility -Location -GPE(Geo-political Entity)	(Not specified explicitly) -Person -Organization -Location -Date -Time -Money -Percent

Table 2: Possible types of anaphor according to the annotation schemes (O: allowed, X: not allowed, U: unspecified)

TYPE	GENIA	ACE	MUC
Personal pronoun	O	O	O
Demonstrative pronoun	O	O	O
Possessive pronoun	O	O	O
Reflexive pronoun	O	O	U
Indefinite pronoun (e.g., *both*)	O	U	U
Pleonastic pronoun *it*	X	U	U
Bound anaphor	X	U	O
Mention with empty head (e.g., *five* of)	X	U	U
here, *there*	U	O	U

The possible types of annotated anaphoric pronouns are given in Table 2. **O** denotes the type of pronoun, which may be annotated as markable, in contrast to **X**, which denotes the type of pronoun, which is not allowed to be annotated as markable. The notation **U** represents the annotation scheme that does not state how a type should be treated because that type is not popular in the domain, or the scheme does not allow the annotation of such a type implicitly.

Using the similar notations as in Table 2, Table 3 shows the possible syntactic structures of antecedents according to the annotation schemes, which are also the structures of markables in real annotations. In practice, such structural variations may cause troubles for automatically markable recognition, so in the experiments with pronoun resolution, gold markables are often used to eliminate error-prone problems.

2.2 Corpus preprocessing

Our objective anaphoric pronouns are limited to the following types: personal pronouns (all cases), possessive pronouns, and demonstrative pronouns, which have a nominal antecedent. In addition

Table 3: Possible types of antecedent according to the annotation schemes (O: allowed, X: not allowed, U: unspecified)

TYPE	GENIA	ACE	MUC
Pronominal	X	O	O
Noun used as a modifier (embedded in NP)	X	O	O
Name, named entity (embedded in NP)	X	O	O
Gerund	U	U	X
NP with a head noun (definite and indefinite)	O	O	O
Conjoint NP (with more than one head)	O	X	O
Coordinated NP	O	O	O
Predicate nominal	X	O	O
NP with a restrictive appositive phrase	X	O	O
NP with a non-restrictive appositive phrase	X	O	O
NP with a restrictive prepositional phrase	O	O	O
NP with a non-restrictive prepositional phrase	X	O	O
NP with a restrictive relative clause	O	O	O
NP with a non-restrictive relative clause	O	O	O
Infinitive clause	O	U	U
Date, Currency expressions, and percentages	U	U	O
Proper adjective (e.g.,*French*)	U	O	U
here, *there*	X	O	U

to these types of pronouns, the annotated corpora contain other types of pronominal anaphora, including "both," "one," numeric mentions (GENIA), and bound anaphora (ACE). However, analysis statistics show that such pronouns occupy less than 5% of the total pronouns in the GENIA corpus, thus we have ignored them.

In the preprocessing step, for each corpus, we extract the gold pronominal anaphora links, which link the anaphoric pronouns with their antecedents. Although MUC and GENIA used the same asymmetric annotation schemes, picking one gold antecedent in a set of co-referenced mentions is not straightforward, since pronouns in GENIA are not allowed to be linked with a pronomial antecedent, while in the MUC corpus, this kind of link is allowed. In order to achieve the fairest comparative experimental results, we uniformly choose the nearest item in the co-reference chain of a pronoun, and make a gold anaphora link. This policy is best suited for ACE, thanks to the symmetric scheme used.

Table 4: Sizes of the data sets (number of anaphoric pronoun)

	GENIA	ACE	MUC
Training set	1442	2427	371
Test set	357	633	240

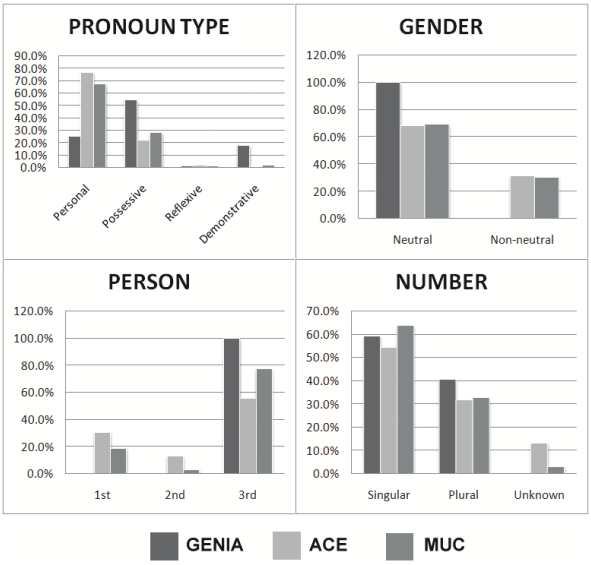

Figure 2: Analysis of anaphoric pronoun in different data sets

2.3 Statistics

In the following step, we analyze the extracted anaphora links for the three corpora. The analysis statistics in Figure 2 show the differences of the distributions of pronoun types and pronoun properties in three data sets: MUC-7, GENIA, and BNEWS from ACE. Note that only four major types out of the nine types of anaphoric pronouns mentioned in the previous section are counted. In particular, the chosen types correspond to those rows in Table 2 that contain at least two **O**.

We can see that all of the anaphoric pronouns in GENIA are neutral-gender and third-person pronouns. Another difference is that the number of demonstrative pronouns in GENIA comes to about 20%, which is much more than in other data sets.

As each type of pronoun has its own referential characteristics, such differences in the distributions of pronouns can significantly affect the pronoun resolution. This will be shown in our experiments, and analysis of the experimental results will be given in the following section.

3 Implementation

3.1 Pronoun resolution model

We built a machine learning based pronoun resolution engine using a Maximum Entropy ranker model (Berger et al., 1996), similar with Denis and Baldridge's model (Denis and Baldridge, 2007). For every anaphoric pronoun π, the ranker selects the most likely antecedent candidate α, from a set of k candidate markables.

$$P_r(\alpha_j|\pi) = \frac{\exp\left(\sum_{i=1}^{n} \lambda_i f_i(\pi, \alpha_j)\right)}{\sum_k \exp\left(\sum_{i=1}^{n} \lambda_i f_i(\pi, \alpha_k)\right)} \quad (1)$$

We constructed the training examples in the following way: for each gold anaphora link in the training corpus, we created a positive instance, and negative training instances are created by pairing the pronoun with all of the other markables appearing in a window of w preceding sentences. In all the experiments on ACE and MUC, we set w to 10 sentences, while for GENIA, w is set to 5. This setting is based on our corpus analysis showing that many of the gold antecedents in the bio-domain texts are in at most three sentences from their anaphors. In the resolution phase, the same method for collecting instances was also applied.

3.2 Features

Table 5 shows the *primitive features* used in our system, which are grouped into *feature groups* according to the type of information that they carry. Note that the actual features used by the ranker are distance features (*sdist*, and *tdist*), and not only the primitive features themselves, but also the combinations of these primitive features. The pronoun resolution model makes use of the discriminative power of these combinatory features. For example, the combination of *P_num* and *C_num* tests the *agreement* in number between the anaphoric pronoun and its candidate. Such agreements in number and gender are one of the constraints in the anaphora phenomenon, and have been exploited in almost all machine learning-based pronoun resolution frameworks (Soon et al., 2001).

Each primitive feature is from a layer of text analysis (see *Layer*), which can be morphological (*mor.*), syntactic (*syn.*), or semantic (*sem.*). The second column represents the feature sets that are used in our experiments. The *explanation* column in the table shows the way we extract feature values from texts, with the exception of the primitive

feature *P_semw*, reflecting the context information of the anaphoric pronoun. This feature value is determined in the following way. If the pronoun is a subject, then *P_semw* is its governing head verb, and if it is a possessive adjective, then *P_semw* is the head noun of the noun phrase containing that pronoun. A default value is used if the pronoun belongs to neither of the above cases.

The last column of this table shows an example of the feature characterization for the anaphora link *PMA-its* in this discourse: *"By comparison, **PMA** is a very inefficient inducer of the jun gene family in Jurkat cells. Similar to **its** effect on the induction of AP1 by okadaic acid, PMA inhibits the induction of c-jun mRNA by okadaic acid."*

We divided the feature groups into 3 feature sets: *fundamental*, *baseline* and *additional*. The *fundamental* feature set contains the indispensable features for solving pronoun resolution. The *baseline* feature set mostly includes morphological features, reflecting the properties of text mentions, and in particular the pronoun properties such as *gender*, *number*, etc. The features in the *additional* feature set are used to exploit higher levels of knowledge through more semantic and syntactic features. We also include in this feature set the features that have been used in some previous work in order to clarify their contributions in our system.

4 Experiments

4.1 Experiment setting and evaluation scoring

For each corpus, we trained our resolver on the training set, and then applied it to the development test set. For the case of the ACE corpus, we only used the *train* part of the BNEWS data set for training, and applied on the corresponding *devtest* data set. We randomly splitted the GENIA corpus it into 2 parts: the *train*, and the *heldout* data sets, which contain 1599 and 400 abstracts, respectively. For the MUC corpus, we used the *dryrun* part for training, and the *formal* part for testing.

Similar to previous works, all of the experimental results in this paper are reported in *success rate* (Mitkov, 2002), calculated using the following formula.

$$\textit{Success rate} = \frac{\textit{Number of successfully resolved anaphors}}{\textit{Number of all anaphors}} \quad (2)$$

The input of our resolver are the gold mentions annotated in the corpora. The output anaphora links of a pronoun resolution system are evaluated following two criteria. In criterion 1, the recognized antecedent of an anaphoric pronoun is considered correct only when it matches the antecedent in the gold anaphora link of that pronoun. Criterion 2 is a bit looser when the recognized antecedent just needs to match one of the antecedents of a pronoun in its co-reference chain. This criterion has been used by most of the previous works, including Denis and Baldridge's system (Denis and Baldridge, 2007).

4.2 Baseline resolver

In this experiment, we use the baseline feature set presented in section 3.2. One of the reasons in choosing these features for the baseline system, is that they are basic features that have been used by most of the previous reference resolution systems. Moreover, we wanted to see how these features contribute to the resolution process for different corpora, presented in the next section.

Our baseline system achieved a 71.41% success rate on the BNEWS data set (Table 6, criterion 2), which is comparable to the result of Denis and Baldridge's system on the same data set (Denis and Baldridge, 2007). Moreover, we can see that the differences caused by the two criteria are not the same for every data set. For the news domain data sets, the differences vary from 4.17% (MUC) to 6.8% (ACE), which is high in comparison with the percentages of GENIA, which were less than two percent. This can be explained by the fact that pronouns in news texts are used more repeatedly than those in bio-medical texts. Because bio-entities are neutral-gender mentions, and are referred by the neutral gender and third person pronouns, the repeated use of pronouns may increase the ambiguity of the text, and confuse the readers.

To prevent the confusion of the readers, we chose just one data set BNEWS to represent the ACE corpus and present our further analysis experiments on these three data sets: GENIA, ACE (BNEWS), and MUC (MUC-7).

4.3 Contributions of the features in the baseline resolver

In order to observe the effects of the features in the baseline pronoun resolver, we omitted each feature group from the whole feature set, retrained our resolution models with the new feature set, and applied them to the three data sets: GENIA,

Table 5: Features used in the pronoun resolver

Layer	Feature set	Group	Primitive Feature	Explanation	Example
mor.	fundamental	mention type	P_type C_type	pronoun type candidate mention type	possessive pronoun proper name
	baseline	sdist	CP_sdis	distance in sentence	1
		tdist	CP_tdis	normalized distance in token	17
		numb	P_numb C_numb	number of p number of c	singular unknown
		pers	P_pers C_pers	person of p person of c	third person third person
		gend	P_gend C_gend	gender of p gender of c	neutral neutral
		pfam	P_pfam C_pfam	family of p family of c	it null
		string	P_word C_head	pronoun string candidate head string	*its* *PMA*
syn.	additional	pos	P_lpos P_rpos C_lpos C_rpos	POS of the left word of p POS of the right word of p POS of the left word of c POS of the right word of c	TO NN COMMA VBZ (*is*)
		parg	P_parg C_parg	argument role of p argument role of c	null arg1
sem.		netype	C_netype	entity type of c	null
mor.		last3c	C_last3c	the last 3 characters of c	*pma*
syn.		comb	P_semw	*see Section 3.2*	*effect*
other			C_1stnp	first NP in a sentence or not	false

Table 6: Baseline system evalutation (C1: criterion 1, C2: criterion 2, D: difference between criterion 1 and 2)

	GENIA	ACE	MUC
C1	70.31	64.61	57.08
C2	71.43	71.41	61.25
Diff	1.12	6.80	4.17

BNEWS, and MUC. Pronoun type and mention type are the most significant features, and thus, are not omitted in this experiment.

Table 7 shows the experimental results: the first column is the feature group name, and the following three columns show the resolution accuracy of the three corpora. The figures in the parentheses show the degradation, when we exclude the corresponding group from the baseline feature set. Our data analysis show some noticeable issues:

Number features (*numb*) :

The number-combination features are the most significant features in bio-texts while they are not so effective on ACE, and even perform negatively on MUC. One of the reasons behind this, is that in the bio-texts, all of the anaphoric pronouns have a deterministic number; i.e., either singular or plural (Section 2.3), while the news texts contain first- and second-person pronouns whose numbers are unspecified. Another reason emerges from the non-pronominal types of mentions, which play a role as antecedents. The number property of these mentions is characterized in the markable detection phase based on the part-of-speech tag, the head noun, and the phrase structure of those mentions. In particular, the MUC corpus contains many coordinated-structured mentions (Section 2.1), which are difficult for markable characterization.

Person features and pronoun family (*pers* and *pfam*) :

The absence of the *pers* features caused the biggest loss for the resolution success rate on the ACE corpus, because the co-reference chains in this corpus contain a lot of pronouns, and it is easier for the pronoun resolver to determine a pronominal antecedent than to determine a non-pronominal antecedent. The same phenomena can be observed with *pfam* features. The bio-text only contains third-person anaphoric pronouns (Section 2.3), so the person features do not have any profits.

Distance features (*sdist* and *tdist*) :

Our baseline resolver again confirmed that the sentence distance is an indispensable feature in pronoun resolution. However, the token-based distance did not show any improvements on the MUC corpus. Analyzing the MUC anaphora links, we found that these *tdist* features resulted in 10 correct anaphora links, but also mis-recognized 10 antecedents.

Table 7: Feature contributions in the baseline system (evaluation criterion 1)

Excluded	GENIA	ACE	MUC
none	70.31	64.61	57.08
-sdist	67.23(−3.08)	63.51(−1.10)	51.67(−5.41)
-tdist	70.03(−0.28)	59.56(−5.05)	57.08(+0.00)
-numb	65.83(−4.48)	61.77(−2.84)	58.33(+1.25)
-pers	70.31(+0.00)	57.19(−7.42)	55.42(−1.66)
-gend	69.75(−0.56)	64.45(−0.16)	56.67(−0.41)
-pfam	71.15(+0.84)	63.51(−1.10)	57.92(+0.84)
-string	68.07(−2.24)	61.93(−2.68)	55.83(−1.25)

4.4 Contributions of additional features to the baseline feature set

In addition to the baseline feature set, we enhanced our resolver with more features. Among them, there are two noticeable features: the grammatical role of pronouns or antecedent candidates, and the named entity type of the candidates. The other feature groups are used in Denis and Baldridge's system, which we also want to test in our system.

Table 8 shows the resolution results and the increase when adding the corresponding feature group. With the exception of the *last3c* features, the others significantly improved the resolution success rate on bio-texts, although they did not have clear contributions to the news domain data sets. The following is our further analysis to see the way that these features can contribute to the pronoun resolution process.

Semantic features (*netype*)

The first feature we would like to observe is the combination of *C_netype* and *P_semw* features, which contributed to the increase by 3.64 points. We further conducted a small test by excluding this combination from the *netype* feature group, but the success rate remained unchanged from the baseline result. This signifies that this combination contributed the most to the above increase.

The combination of *C_netype* and *P_semw* features exploits the co-ocurrence of the semantic type of the candidate antecedent and the *context word*, which appears in some relationship with the pronoun. This combination feature uses the information similar to the semantic compatibility features proposed by Yang (Yang et al., 2005) and Bergsma (Bergsma and Lin, 2006). Depending on the pronoun type, the feature extractor decides which relationship is used. For example, the resolver successfully recognizes the antecedent of the pronoun *its* in this discourse: *"**HSF3** is constitutively expressed in the erythroblast cell line HD6 , the lymphoblast cell line MSB , and embryo fibroblasts , and yet **its** DNA-binding activity is induced only upon exposure of HD6 cells to heat shock ,"* because *HSF3* was detected as a Protein entity, which has a strong association with the governing head noun *activity* of the pronoun.

Another example is the correct anaphora link between *"it"* and *"the viral protein"* in the following sentence, which the other features failed to detect. *"Tax , **the viral protein** , is thought to be crucial in the development of the disease , since **it** transforms healthy T cells in vitro and induces tumors in transgenic animals."* The correct antecedent was recognized due to the bias given to the association of the Protein entity type, and the governing verb, *"transform"* of the pronoun. The experimental results show the contribution of the domain knowledge to the pronoun resolution, and the potential combination use of such knowledge with the syntactic features.

Parse features (*parg*)

The combinations of the primitive features of grammatical roles significantly improved the performance of our resolver. The following examples show the correct anaphora links resulting from using the parse features:

- *"By comparison, **PMA** is a very inefficient inducer of the jun gene family in Jurkat cells. Similar to **its** effect on the induction of AP1 by okadaic acid, PMA inhibits the induction of c-jun mRNA by okadaic acid."*

In this example, the possessive pronoun *"its"* in the second sentence corefers to *"PMA"*, the subject of the preceding sentence.

Among the combination features in this group, one noticeable feature is the combination of *C_parg*, *Sdist*, and *P_type* which contains the association of the grammatical role of the candidate, the sentence-based distance, and the pronoun type. The idea of adding this combination is based on the Centering theory (Walker et al., 1998), a theory of discourse successfully used in pronoun resolution. This simple feature shows the potential of encoding centering theory in the machine learning features, based on the parse information.

Feature integration

Finally, we integrated all of the positive feature groups for each data set in the above experiments, and tested this combining feature set. Table

Table 8: Additional features and their contributions (evaluation criterion 1)

Added	GENIA	ACE	MUC
none	70.31	64.61	57.08
+**pos**	75.63(+5.32)	62.88(−1.73)	57.50(+0.42)
+**parg**	73.67(+3.36)	63.82(−0.79)	58.75(+1.67)
+**netype**	73.95(+3.64)	64.30(−0.31)	58.33(+1.25)
+**last3c**	67.51(−2.80)	62.09(−2.52)	56.67(−0.41)
+**comb**	72.83(+2.52)	63.82(−0.79)	56.25(−0.83)

Table 9: Feature integration

	GENIA	ACE	MUC
C1	79.55 (+9.24)	64.61 (+0.00)	60.42 (+3.34)
C2	**80.95** (+9.52)	**71.41** (+0.00)	**66.25** (+5.00)

9 shows a significant increase in the performance of the resolver on GENIA and MUC.

5 Conclusion and future work

Through the differences in the corpus annotation schemes, in the corpora themselves, and in contributions of resolution factors to the pronoun resolution process, we can see that adapting pronoun resolution for a different domain is not an easy task. A good study on the types of anaphoric pronouns and entity mention structures beforehand can help design a better feature set for our machine learning-based pronoun resolution system and thus, can save much time and labor.

As shown in this paper, for the news domain, the properties of anaphoric pronouns contain rich information about their antecedents, which is very useful in the resolution process. While in biomedical text, it is more important to capture the information to connect a pronoun and its antecedent from their surrounding context, because the anaphoric pronouns themselves contain almost no information of their antecedents with the exception of the *numbers*.

As a future work, it would be interesting to see how the system performs in other domains. More experiments should be designed to make the influences of annotation schemes on the pronoun resolution process clearer.

References

Berger, Adam L., Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Bergsma, Shane and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 33–40.

Castano, Jose, Jason Zhang, and James Pusterjovsky. 2002. Anaphora resolution in biomedical literature. In *Int'l Symposium Reference Resolution in NLP*.

Denis, Pascal and J. Baldridge. 2007. A ranking approach to pronoun resolution. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI07)*.

Haghighi, Aria and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855.

Hong, Huaqing. 2004. Coreference annotation scheme for medco corpus.

Kim, Jung-Jae and Jong C.Park. 2004. Bioar: Anaphora resolution for relating protein names to proteome database entries. In *Proceedings of the ACL 2004: Workshop on Reference Resolution and its Applications*, pages 79–86.

Lynette, Hirschman. 1997. Muc-7 coreference task definition.

Mitkov, Ruslan. 2002. *Anaphora resolution*. Pearson Education, London, Great Britain.

NIST. 2003. Entity detection and tracking - phrase 1 edt and metonymy annotation guidelines version 2.5.1 20030502.

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

Walker, Marilyn A., Aravind K. Joshi, and Ellen F. Prince. 1998. *Centering Theory in Discourse*. Clarendon Press, Oxford.

Yang, Xiaofeng, Jian Su, and Chew-Lim Tan. 2005. Improving pronoun resolution using statistics-based semantic compatibility information. In *Proceedings of the 43rd Annual meeting of the Association for Computational Linguistics (ACL05)*, pages 427–434.