Challenges in Pronoun Resolution System for Biomedical Text

Ngan Nguyen[1], Jin-Dong Kim[1], Junichi Tsujii[1,2,3]

[1] University of Tokyo,7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan
[2] University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[3] National Centre for Text Mining, 131 Princess Street, Manchester, M1 7DN, UK
{nltngan, jdkim, tsujii}@is.s.u-tokyo.ac.jp

Abstract

This paper presents our findings on the feasibility of doing pronoun resolution for biomedical texts, in comparison with conducting pronoun resolution for the newswire domain. In our experiments, we built a simple machine learning-based pronoun resolution system, and evaluated the system on three different corpora: MUC, ACE, and GENIA. Comparative statistics not only reveal the noticeable issues in constructing an effective pronoun resolution system for a new domain, but also provides a comprehensive view of those corpora often used for this task.

1. Introduction

A pronoun is considered anaphoric when it refers to some entity mention appearing previously in text, and pronoun resolution determines such referenced mentions or antecedents. This is a key subtask in anaphora resolution and co-reference resolution, the two significant tasks required to be solved when approaching the goal of natural language understanding.

The shift from heuristics and knowledge-based (Mitkov, 1998) to machine learning and corpus-based methods (Ng, 2005), (Soon et al., 2001) has made the annotated corpora a vital resource in both training and evaluating resolution models. Because of similar reference characteristics, the same co-reference annotated corpora are often employed for both co-reference resolution, and for anaphora resolution tasks. Among such, MUC and ACE data sets are very popular for the newswire domain. Many experiments on these corpora produced good results (Yang et al., 2006), (Haghighi and Klein, 2007). Recently, for the biomedical domain, the GENIA corpus has been annotated for co-references. In this work, we aim to compare these three corpora with respect to corpus-based pronoun resolution tasks. A better understanding of these corpora can make it possible to inherit achievements selectively from previous works for the newswire domain, thus building an effective pronoun resolution system for the bio domain. For this purpose, in our experiments, we employed a simple corpus-based pronoun resolution system composed of three common components: markable detection, anaphoricity determination, and pronoun resolution engine. All comparative analysis statistics are based on the MUC7 (both dryrun and formal data sets), BNEW, NPAPER, and NWIRE data sets (both train and devtest) in ACE, and GENIA.

In Section 2, we briefly describe our pronoun resolution system. Section 3 focuses on the pronoun resolution engine, the main component of the system. The evaluation results of this component on each dataset shows the different contributions of features in the process of anaphora resolution for different domains. In section 4, we analyze the differences of anaphoric pronouns in the three corpora, one of the main causes of variation in system performance. Then,
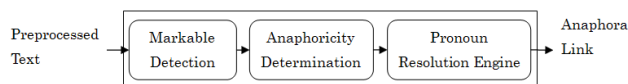

Figure 1: Pronoun resolution system

the experimental results on markable detection component in Section 5 also presents some challenges that deal with various types of entity mentions in each domain. Finally, we conclude our paper in section 6 with future directions.

2. Pronoun resolution system

We built a simple pronoun resolution system containing three main components: markable detection, anaphoricity determination, and pronoun resolution engine (Figure 1). The markable detection detects all mentions called markables, which may join in pronominal anaphora relations, including pronouns. These mentions are basically noun phrases extracted from a base noun phrase chunker, and pronouns recognized by a part-of-speech tagger. Markables are then input to the anaphoricity determination component, which is in charge of determining whether a pronoun is anaphoric or not.

However, in order to estimate the complexity of anaphoricity determination on each dataset, we relax the system with the assumption that all pronouns detected by markable detection are anaphoric. Anaphoric pronouns will then be fed into the last component: the pronoun resolution engine, which will then pick out one antecedent for each anaphor from a set of its candidate markables, thus producing an anaphora link.

3. Pronoun resolution engine

3.1. Pronoun resolution model

We built a machine learning based pronoun resolution engine using a Maximum Entropy ranker model 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 create a positive instance, and the 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 at most three sentences from their anaphors. In the resolution phase, the same style for collecting instances was also applied.

3.2. Features

Table 1 shows the *primitive features* used in our system, which are grouped into *feature groups* according to their common information. Note that the actual features used by the ranker are distance features (*sdist*, and *tdist*), and are not only the primitive features themselves, but also the combinations of these primitive features.

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."*

The feature set includes the combination features of the primitive features.

3.3. Baseline

In this experiment, we use a compact feature set containing two distance features, and the combinations of these features, and the primitive features presented in the section 3.2. One of the reasons why we chose this feature set for the baseline system, is that they are very basic features that have been used by almost all of the previous reference resolution systems. Moreover, we would like to see how these features contribute to the resolution process for different corpora, presented in the next section.

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

In these experiments, we used the gold mentions, which are manually annotated in the corpora, because we would like to compare our duplicate resolver directly with Denis and Baldridge's resolver. It should be noted that many of the previous works were unclear about this point.

Our baseline system achieved 71.41% accuracy on the BNEWS data set (Table 2), which are comparable results to their system (72.9%). Moreover, we can see that the differences caused by the two criteria are not the same for every data set. For the newswire domain data sets, the differences

Table 2: Baseline system evaluation (C1: Criteria 1, C2: Criteria 2, D: Difference)

	GENIA	BNEWS	NPAPER	NWIRE	MUC
C1	70.31	64.61	62.64	63.35	57.08
C2	71.43	71.41	72.10	72.67	61.25
D	1.12	6.8	9.46	9.32	4.17

vary from 4.17% (MUC-7) to 9.46% (NPAPER), 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 newswire domain texts are used more repeatedly than pronouns in bio-medical texts. Because bio-entities are neutral-gender mentions, and are referenced by the same gender and third person pronouns, the repeated use of pronouns may increase the ambiguity of the text, confusing the readers.

3.4. Cross-corpus evaluation

In Table 3, the cell at position [X, Y] shows the evaluation statistic of the system trained on the X corpus, and evaluated on the Y corpus. All of the evaluation results are given in *success rate* 2, a common evaluation scoring in anaphora resolution proposed by Mitkov (Mitkov, 2001). It is calculated as the total successfully resolved anaphoric pronouns divided by the total number of manually annotated anaphoric pronouns.

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

We can see in this table that the system trained on the biomedical corpus shows a more significant degradation on either ACE or MUC, in comparison with the scores achieved when we train on ACE and apply on MUC and vice versa. In the remainder of this paper, we focus on finding the probable causes of this performance degradation by comparing the corpora statistically, as well as observing the contributions of features into the resolution process through the experiment in the following section.

Table 3: Evaluation results of the pronoun resolution system

	GENIA	ACE	MUC
GENIA	**71.43**	60.03	56.67
ACE	68.63	**71.41**	57.92
MUC	65.23	67.46	**61.25**

3.5. 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:

Table 1: Features used in the pronoun resolver

Group	Primitive Feature	Explanation	Example
mention type	P_type	pronoun type	possessive pronoun
	C_type	candidate mention type	proper name
sdist	CP_sdis	distance in sentence	1
tdist	CP_tdis	normalized distance in token	17
numb	P_numb	number of p	singular
	C_numb	number of c	unknown
pers	P_pers	person of p	third person
	C_pers	person of c	third person
gend	P_gend	gender of p	neutral
	C_gend	gender of c	neutral
pfam	P_pfam	family of p	it
	C_pfam	family of c	null
string	P_word	pronoun string	*its*
	C_head	candidate head string	*PMA*

GENIA, BNEWS, and MUC-7. Pronoun type and mention type are the most significant features, and thus, are not omitted in this experiment.

Table 4 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, while the news wire 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 particilar, the MUC corpus contains many coordinated-structured mentions, 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 coreference 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, so the person features do not have any profits.

- **Distance features (*sdist* and *tdist*) :** Our baseline resolver again confirmed that sentence distance is an

Table 4: Feature contributions in the baseline system (evaluation criteria 1)

Without	GENIA	ACE	MUC
Baseline	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)

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 misrecognized 10 antecedents. We should thus use the distance-based features with care.

4. Distributions of pronouns in the corpora

Since each type of pronoun has its own usage depending on its referring characteristics, the distributions of pronoun types in the corpora provide us with much information about the reference phenomenon in data, based on which we should design proper features to solve the task.

In Figure 2, we compare the distributions of anaphoric pronouns in five data sets (ACE contain three data sets), from different perspectives: pronoun types, pronoun gender, and number, the morphological factors proven to be very important in pronoun resolution for the newswire domain (Bergsma and Lin, 2006). However, it can be observed in the graphs that while the distributions of MUC and ACE data sets are similar to each other, GENIA is very different. In GENIA, or bio-texts, a majority of the anaphoric pronouns are neutral-gender, and third-person pronouns (b)(c), which means gender- and person-agreement features are not so effective in this domain as in the newswire domain. Moreover, we can see that the possessive and demonstrative pronouns occupy around 70 percents of the total num-

Table 5: Sizes of the data sets (number of anaphoric pronoun)

	GENIA	BNEWS	NPAPER	NWIRE	MUC
Train	1442	2427	2058	2177	371
Test	357	633	613	450	240

Table 6: Markable detection with an NP chunker

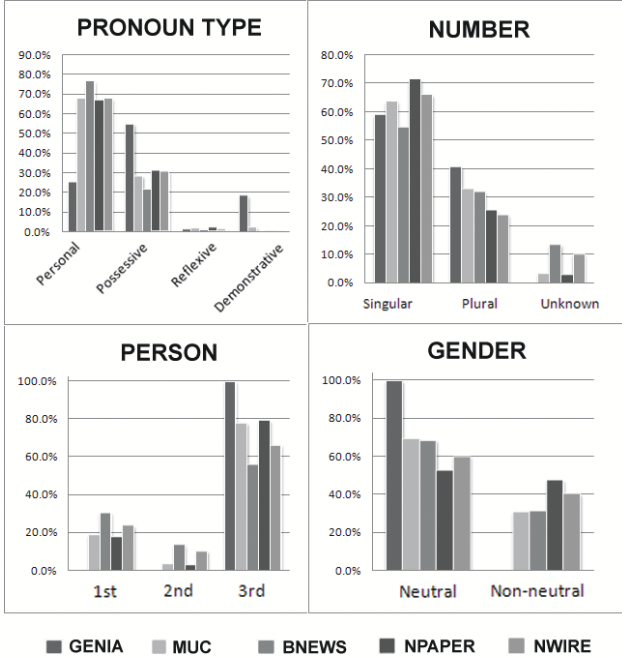|	GENIA	ACE	MUC
Total of mentions	3491	4818	1047
Mention coverage rate	94.59%	95.66%	94.46%
Total of links	1799	3191	617
Link coverage rate	89.55%	92.98%	90.76%

Figure 2: Corpus analysis

ber of anaphoric pronouns (a), while in the newswire text (ACE, MUC), personal pronouns occupy the majority. This indicates that the current set of features may not be good enough to make use of the linguistic characteristics of these types of pronouns, and thus, should be re-designed.

5. Markable detection for biomedical text

The markable detection module is responsible for recognizing all kinds of mentions in texts joinable in the anaphora relationship. In the previous works, a base noun phrase chunker is often used for this task, and all base noun phrases produced are considered as markables. Some researchers additionally merge named entities produced by a NER with base NP chunks to form the markable set (Soon et al., 2001).

In our system, we built a chunker-based markable detector using the GENIA Tagger (Tsuruoka and Tsujii, 2005). For each input NP chunk, the markable detector creates a markable. If a chunk contains a possessive pronoun, then that pronoun will form another markable. Every markable is then characterized by gender, number, person, head, etc., mainly using the part-of-speech information of its content tokens.

In this experiment, we tested our markable detector on three data sets: GENIA, ACE (BNEWS), and MUC-7 (including both the training and test sets). The evaluation results are reported in the coverage rate, as shown in the Table 6. *Men-*

tion coverage rate represents the number of correct markables divided by the total mention of gold mentions in the data set. *Link coverage rate* is the percentage of anaphora links whose anaphor and antecedent are included in the detected markables. This number will become the upper bound for the system recall. It can be seen that the coverage rates are not significantly different from each other, though the link coverage for the GENIA corpus is smaller than the other two corpora.

Furthermore, in order to know the kinds of errors that occurred, we picked out 50 errors, and classified them into several common types for each corpus. The error analysis results given in Table 8 show several interesting issues:

- **Coordination:** The GENIA and MUC corpora have many mentions with coordinated structures. Such complex noun phrases are often split by the noun phrase chunker, and are missed from the the set of detected markables.

- **Named entity:** While in ACE and MUC, many markablesexactly match some named entities, in GENIA this is very rare (only 1 over 50). We can conclude that although bio-texts contain many named entities, the pronouns do not often co-refer with those named entities, except in the case where other concepts have more complex structures. Such is the case with: *"The Epstein-Barr virus early antigen diffuse component (EA-D)," and "the nuclear affinity (Ka) for T3."*

  Therefore, with bio-domain texts, using named entities as markables is not as profitable as in newswire domain.

- **NP contains special characters:** The GENIA tagger tends to break down those noun phrases containing some special characters such as: hyphens, squares, or round brackets, points, as in *"11 alpha-methyl-1 alpha,25-(OH)2D3."* Though such noun phrases do not often appear in the newswire domain, they present a big problem in bio-domain texts. In order to solve this problem, the chunker should be retrained, or we need to build another bio-mention detector.

- **"that":** Another source of errors comes from the demonstrative pronoun, *that*, which is quite popular in technical papers. The word, *that* has several possible part-of-speech tags: **DT** when it plays a role as a determiner, or **IN** when a subordinating conjunction. This ambiguity is error-prone in POS-tagging and chunking. When *that* is not correctly recognized as a noun phrase, it is ignored in markable detection. For example, it may be ignored in the sentence, *"The results showed **that** there was no significant difference*

Table 7: Average lengths in token of markable-detection errors

	GENIA	ACE	MUC
Average length	6.7	3.3	4.5

*between the ER content of lymphocytes from the controls and **that** from patients with SLE .”* The former *that* is tagged with **IN**, and produces an **SBAR**-typed chunk, while the latter *that* should be tagged with **DT**, producing an **NP** chunk.

It may also be a confusing word to the annotators, and is the cause of some annotation mistakesn. An example of such a case is in the sentence, *“Cotransfection studies with **this cDNA** indicate <u>that</u> it can repress basal promoter activity,”* where “that” is tagged as an **IN** but should be tagged as a **DT**.

- **“here”/“there” in ACE:** These two pronouns in the ACE corpus referring to a location are tagged as adverbial phrases, and are thus ignored by our markable detection.

Our error analysis also shows that the system failed to recognize long markables appearing very frequently in biomedical text (Table 7). This reflects the fact that linguistic structures of the antecedents in biomedical texts are much more complicated than those in the newswire domain, in which referred mentions are usually proper names. The antecedents in biomedical texts are usually complex noun phrases with modifying phrases or clauses; or base noun phrases with many modifiers for the head nouns. For examples, the antecedent of the pronoun *their* in “ …and macrophages express closely related immunoglobulin G (IgG) Fc receptors (Fc gamma RII) that differ only in the structures of their cytoplasmic domains,” is “ closely related immunoglobulin G (IgG) Fc receptors (Fc gamma RII) ” which is very complicated and difficult for analysis tools to correctly recognize.

In summary, the error analysis of the simple markable detection reveals that using a base noun phase chunker and a named entity recognizer is insufficient for markable detection in the bio-domain, because of the complex mentions. Since the markables detected in this phase affects the proceeding process in the next steps, it is necessary to build a bio-chunker that has the capability to detect complex bio-mentions for the bio-domain, in order to achieve a high performance for the whole pronoun resolution system for bio-texts.

6. Conclusion and future work

In this paper, we present our study on the significant differences among some popular co-reference annotated corpora for two text genres: newswire and biomedical technical papers. The analysis statistics provide us with a valuable indication of the complexity of the pronoun resolution problem. Based on this study, we plan on improving the pronoun resolution system for biomedical texts by adding new features

Table 8: Markable detection error analysis

Type of error	GENIA	ACE	MUC
coordinated NP	8	2	18
named entity	1	15	12
chunker error	3	13	7
embedded in NP	3	0	5
preprocessing error	0	1	3
NP contains special characters	24	0	2
annotation error	7	0	1
that (GENIA specific)	4	0	0
here, there (ACE specific)	0	17	0
other	0	2	2
Total	50	50	50

designed for specific pronoun types, and making use of the rich domain information.

7. References

Shane Bergsma 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.

Pascal Denis and J. Baldridge. 2007. A ranking approach to pronoun resolution. In *Proceedings of IJCAI-2007*.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *In proceedings of ACL 2007*.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *In Proceedings of ACL '98*, pages 869–875.

Ruslan Mitkov. 2001. Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. *Applied Artificial Intelligence*, 15(3):253–276(24).

Vincent Ng. 2005. Supervised ranking for pronoun resolution: Some recent improvements. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*.

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

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *In Proceedings of HLT/EMNLP 2005*, pages 467–474.

Xiaofeng Yang, Jian Su, and Chew-Lim Tan. 2006. Kernel-based pronoun resolution with structured syntactic knowledge. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pages 41–48.