Machine Learning Based Pronoun Resolution for Biomedical Text

Ngan L.T. Nguyen[1]　Yusuke Miyao[1]　Jin-Dong Kim[1]　Junichi Tsujii[1,2,3]
[1] Department of Computer Science, University of Tokyo
[2] School of Computer Science, University of Manchester, UK
[3] NaCTeM (National Center for Text Mining), Manchester, UK
{nltngan, yusuke, jdkim, tsujii}@is.s.u-tokyo.ac.jp

1 Introduction

In this research, we investigated the influence of domain differences on the pronoun resolution problem, laying the foundations for solving the same task for the biomedical domain.

Pronoun resolution is known as the task of determining the *antecedent* of an *anaphoric* pronoun, a pronoun which points back to some previously mentioned item in a text. For example, in this sentence, *"The dog bites Mary on her leg,"* the possessive pronoun *"her"* should be resolved to point back to the person *"Mary,"*, but not to *"the dog."*

For other domains, especially the news wire domain, many works on pronoun resolution have been carried out by the researchers in the NLP field [5][6]. Nonetheless, there are still not many works for the bio-domain [2][4]. In order to recognize the important factors in building an efficient pronoun resolution system, 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 wire domain, we used the MUC-7 and ACE corpora, and for the biomedical domain, we employed the GENIA co-reference corpus, containing 1999 MEDLINE abstracts annotated with co-reference information.

Section 2 briefly introduces the ranker-based pronoun resolution model, and the primitive features used in our resolver. In section 3, we present our experiment settings, the evaluation scheme, and our experimental results. Finally, we conclude our paper in section 4.

2 Approach

2.1 Pronoun resolution model

We built a machine learning based pronoun resolution engine using a Maximum Entropy ranker model in a similar way to Denis and Baldridge [3]. For every anaphoric pronoun π, the ranker selects the most likely antecedent candidate α from a set of k candidate markables.

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

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 which appear 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 3 sentences from their anaphors. In the resolution phase, the same style of collecting instances was also applied.

2.2 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 the combinations of these primitive features, not only the primitive features themselves.

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."*

Each primitive feature is from a layer of text analysis (see *Layer*), which can be morphological (*mor.*), syntactic (*syn.*), semantic (*sem.*), or discourse (*disc.*). The second column represents the feature sets that are used in our experiments.

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

Table 1: Features used in the pronoun resolver

Layer	Feature set	Group	Primitive Feature	Explanation	Example
mor.	fundamental	mention type	P_type	pronoun type	possessive p.
			C_type	candidate mention type	proper name
	baseline	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*
syn.			C_head	candidate head string	*PMA*
	additional	pos	P_lpos	POS of the left word of p	TO
			P_rpos	POS of the right word of p	NN
			C_lpos	POS of the left word of c	COMMA
			C_rpos	POS of the right word of c	VBZ (*is*)
		parg	P_parg	argument role of p	null
			C_parg	argument role of c	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 3.4*	*effect*
disc.			C_1stnp	first NP or not	false

3 Experiments

3.1 Experiment settings and evaluation method

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 on the corresponding *devtest* data set. For the GENIA corpus, we randomly split 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.

All of the experimental results in this paper are reported in *success rate* [5], calculated using the following formula.

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

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 response 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 response antecedent just needs to match one of the antecedents of a pronoun in its coreference chain. This criterion has been used by most of the previous works, including Denis and Baldridge's system [3].

3.2 Baseline resolver

In this experiment, we use the feature set presented in the section 2.2. One of the reasons why we chose this feature set for the baseline system, is that they are basic features that have been used by almost all 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 2), which is comparable to the results of Denis and Baldridge's 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 news wire domain data sets, the differences vary from 4.17% (MUC) to 6.8% (ACE), which is high in comparison to that of GENIA, which is less than 2 percent. This can be explained by the fact that pronouns in news wire domain 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, confusing the readers.

Table 2: Baseline system evaluation

	GENIA	ACE	MUC
Criterion 1 (C1)	70.31	64.61	57.08
Criterion 2 (C2)	71.43	71.41	61.25
Difference	+1.12	+6.8	+4.17

3.3 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 3 data sets: GENIA, BNEWS, and MUC-7. Pronoun type and mention type are the most significant features, and thus, are not omitted in this experiment.

Table 3 shows the experimental results: the first column is the feature group name, and the following 3 columns show the resolution accuracy of the 3 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 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 particular, 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 pronominal antecedent than to determine a non-pronominal one. The same phenomenon 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 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. We should thus use the distance-based features with care.

Table 3: Feature contributions in the baseline system (evaluation criteria 1)

	GENIA	ACE	MUC
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)
string	68.07(−2.24)	61.93(−2.68)	55.83(−1.25)
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)

3.4 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 4 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 wire domain data sets. The following is our analysis of how 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 [8] and Bergsma [1]. 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*

Table 4: Additional features and their contributions (evaluation criteria 1)

	GENIA	ACE	MUC
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)

*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, *it-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 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 are examples that show the correct anaphora links resulting from using these 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 that 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 [7], 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.

We found that the ***pos*** group also has the effects similar to parse features.

Table 5: Feature integration

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

Feature integration

Finally, we integrated all of the positive feature groups for each data set in the above experiments, and tested this new feature set. The results are shown in Table 5. As we can see in the table, the performance of the resolver on GENIA increased up to 80.95 %, and 66.25 % on the MUC-7 data set.

4 Conclusion and future work

Our experiments showed that the contributions of the popular basic features to each domain were not the same. Additionally, the combinations of semantic features and syntactic features proved to be useful in the disambiguation of antecedent candidates for the bio-domain. This study thus contributes to the design of a better solution to the pronoun resolution problem, as well as to the co-reference task for a new domain, like the bio-domain.

References

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

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

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

[4] Jung jae Kim and Jong C.Park. Bioar: Anaphora resolution for relating protein names to proteome database entries. In *Proc. of the ACL 2004: Workshop on Reference Resolution and its Applications*, pp. 79–86, 2004.

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

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

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

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