Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach

Shanheng Zhao and **Hwee Tou Ng**
Department of Computer Science
National University of Singapore
3 Science Drive 2
Singapore 117543
{zhaosh, nght}@comp.nus.edu.sg

Abstract

In this paper, we present a machine learning approach to the identification and resolution of Chinese anaphoric zero pronouns. We perform both identification and resolution automatically, with two sets of easily computable features. Experimental results show that our proposed learning approach achieves anaphoric zero pronoun resolution accuracy comparable to a previous state-of-the-art, heuristic rule-based approach. To our knowledge, our work is the first to perform both identification and resolution of Chinese anaphoric zero pronouns using a machine learning approach.

1 Introduction

Coreference resolution is the task of determining whether two or more noun phrases refer to the same entity in a text. It is an important task in discourse analysis, and successful coreference resolution benefits many natural language processing applications such as information extraction, question answering, etc.

In the literature, much of the work on coreference resolution is for English text (Soon et al., 2001; Ng and Cardie, 2002b; Yang et al., 2003; McCallum and Wellner, 2005). Publicly available corpora for coreference resolution are mostly in English, e.g., the Message Understanding Conference tasks (MUC6 and MUC7)[1]. Relatively less work has been done on coreference resolution for Chinese. Recently, the ACE Entity Detection and Tracking (EDT) task[2] included annotated Chinese corpora for coreference resolution. Florian et al. (2004) and Zhou et al. (2005) reported research on Chinese coreference resolution.

A prominent phenomenon in Chinese coreference resolution is the prevalence of zero pronouns. A zero pronoun (ZP) is a gap in a sentence which refers to an entity that supplies the necessary information for interpreting the gap. An anaphoric zero pronoun (AZP) is a zero pronoun that corefers to one or more overt noun phrases present in the preceding text. Zero pronouns occur much more frequently in Chinese compared to English, and pose a unique challenge in coreference resolution for Chinese. For example, Kim (2000) conducted a study to compare the use of overt subjects in English, Chinese, and other languages. He found that the use of overt subjects in English is over 96%, while this percentage is only 64% for Chinese, indicating that zero pronouns (lack of overt subjects) are much more prevalent in Chinese.

Chinese zero pronouns have been studied in linguistics research (Li and Thompson, 1979; Li, 2004), but only a small body of prior work in computational linguistics deals with Chinese zero pronoun identification and resolution (Yeh and Chen, 2004; Converse, 2006). To our knowledge, all previous research on zero pronoun identification and resolution in Chinese uses hand-engineered rules or heuristics, and our present work is the first to perform both identification and resolution of Chinese

[1] http://www-nlpir.nist.gov/related_projects/muc/

[2] http://www.nist.gov/speech/tests/ace/

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 541–550, Prague, June 2007.

anaphoric zero pronouns using a machine learning approach.

The rest of this paper is organized as follows. In Section 2, we give the task definition, and describe the corpus used in our evaluation and the evaluation metrics. We then give an overview of our approach in Section 3. Anaphoric zero pronoun identification and resolution are presented in Section 4 and 5, respectively. We present the experimental results in Section 6 and related work in Section 7, and conclude in Section 8.

2 Task Definition

2.1 Zero Pronouns

As mentioned in the introduction, a zero pronoun (ZP) is a gap in a sentence which refers to an entity that supplies the necessary information for interpreting the gap. A coreferential zero pronoun is a zero pronoun that corefers to one or more overt noun phrases present in the same text.

Just like a coreferential noun phrase, a coreferential zero pronoun can also corefer to a noun phrase in the preceding or following text, called anaphoric or cataphoric, respectively. Most coreferential zero pronouns in Chinese are anaphoric. In the corpus used in our evaluation, 98% of the coreferential zero pronouns have antecedents. Hence, for simplicity, we only consider anaphoric zero pronouns (AZP) in this work. That is, we only attempt to resolve a coreferential zero pronoun to noun phrases preceding it.

Here is an example of an anaphoric zero pronoun from the Penn Chinese TreeBank (CTB) (Xue et al., 2005) (sentence ID=300):

[中国	机电	产品	进出口	
[China	electronic	products	import and export	
贸易]$_1$	继续	增加	，	ϕ_2
trade]$_1$	continues	increasing	,	ϕ_2
占	总	进出口	的	
represents	total	import and export	’s	
比重	继续	上升	。	
ratio	continues	increasing	.	

The anaphoric zero pronoun ϕ_2 is coreferring to noun phrase 1. The corresponding parse tree is shown in Figure 1. In CTB, IP refers to a simple clause that does not have complementizers. CP, on the other hand, refers to a clause introduced by a complementizer.

Resolving an anaphoric zero pronoun to its correct antecedent in Chinese is a difficult task. Although gender and number information is available for an overt pronoun and has proven to be useful in pronoun resolution in prior research, a zero pronoun in Chinese, unlike an overt pronoun, provides no such gender or number information. At the same time, identifying zero pronouns in Chinese is also a difficult task. There are only a few overt pronouns in English, Chinese, and many other languages, and state-of-the-art part-of-speech taggers can successfully recognize most of these overt pronouns. However, zero pronouns in Chinese, which are not explicitly marked in a text, are hard to be identified. Furthermore, even if a gap is a zero pronoun, it may not be coreferential. All these difficulties make the identification and resolution of anaphoric zero pronouns in Chinese a challenging task.

2.2 Corpus

We use an annotated third-person pronoun and zero pronoun coreference corpus from Converse (2006)[3]. The corpus contains 205 texts from CTB 3.0, with annotations done directly on the parse trees. In the corpus, coreferential zero pronouns, third-person pronouns, and noun phrases are annotated as coreference chains. If a noun phrase is not in any coreference chain, it is not annotated. If a coreference chain does not contain any third-person pronoun or zero pronoun, the whole chain is not annotated.

A zero pronoun is not always coreferential with some noun phrases. In the corpus, if a zero pronoun is not coreferential with any overt noun phrases, it is assigned one of the following six categories: discourse deictic (#DD), existential (#EXT), inferrable (#INFR), ambiguity between possible referents in the text (#AMB), arbitrary reference (#ARB), and unknown (#UNK). For example, in the following sentence, ϕ_3 refers to an event in the preceding text, with no corresponding antecedent noun phrase. So no antecedent is annotated, and ϕ_3 is labeled as #DD.

香港	著名	财团	长江
Hong Kong	famous	syndicate	Cheung Kong

[3] The data set we obtained is a subset of the one used in Converse (2006).

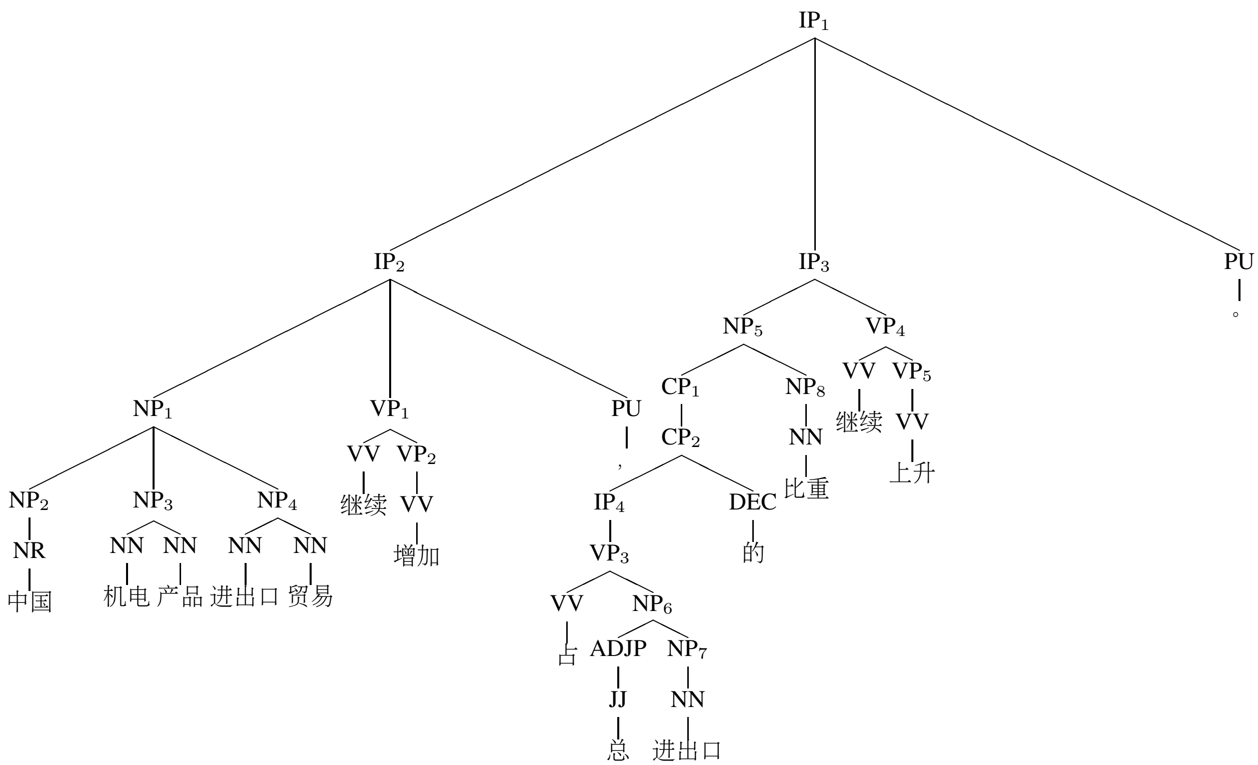

Figure 1: The parse tree which corresponds to the anaphoric zero pronoun example in Section 2.1.

实业 、 百富勤 作为 战略性
Holdings , Peregrine as strategic
投资者 已 购入 了
investors already purchased LE
“ 深业 控股 ” 百分之十二
“ Shenye Holdings ” twenty percent
的 股权 ， ϕ_3 充分 反映 出
's share , ϕ_3 fully reflects out
投资者 的 信心 。
investors 's confidence .

Converse (2006) assumed that all correctly identified AZPs and the gold standard parse trees are given as input to her system. She applied the Hobbs algorithm (Hobbs, 1978) to resolve antecedents for the given AZPs.

In our case, we are only interested in zero pronouns with explicit noun phrase referents. If a coreference chain does not contain AZPs, we discard the chain. We also discard the 6 occurrences of zero pronouns with split antecedents, i.e., a zero pronoun with an antecedent that is split into two separate noun phrases. A total of 383 AZPs remain in the data set used in our experiments.

Among the 205 texts in the data set, texts 1–155 are reserved for training, while the remaining texts (156–205) are used for blind test. The statistics of the data set are shown in Table 1.

	Training	Test
Doc ID	1–155	156–205
# Docs	155	50
# Characters	96,338	15,710
# Words	55,348	9,183
# ZPs	665	87
# AZPs	343	40

Table 1: Statistics of training and test data sets.

2.3 Evaluation Metrics

As in previous work on pronoun resolution, we evaluate the accuracy in terms of recall, precision, and F-measure. The overall recall and precision on the test set are computed by micro-averaging over all test instances. The overall F-measure is then computed.

For AZP identification, recall and precision are defined as:

$$Recall_{AZP} = \frac{\text{\# AZP Hit}}{\text{\# AZP in Key}}$$

$$Precision_{AZP} = \frac{\text{\# AZP Hit}}{\text{\# AZP in Response}}$$

An "AZP Hit" occurs when an AZP as reported in the response (system output) has a counterpart in the same position in the gold standard answer key.

For AZP resolution, recall and precision are defined as:

$$Recall_{Resol} = \frac{\text{\# Resol Hit}}{\text{\# AZP in Key}}$$

$$Precision_{Resol} = \frac{\text{\# Resol Hit}}{\text{\# AZP in Response}}$$

A "Resol Hit" occurs when an AZP is correctly identified, and it is correctly resolved to a noun phrase that is in the same coreference chain as provided in the answer key.

3 Overview of Our Approach

In this section, we give an overview of our approach for Chinese AZP identification and resolution.

Typically, the input raw texts need to be processed by a Chinese word segmenter, a part-of-speech (POS) tagger, and a parser sequentially. Although our approach can apply directly to machine-generated parse trees from raw text, in order to minimize errors introduced by preprocessing, and focus mainly on Chinese zero pronoun resolution, we use the gold standard word segmentation, POS tags, and parse trees provided by CTB. However, we remove all null categories and functional tags from the CTB gold standard parse trees. Figure 1 shows a parse tree after such removal.

A set of zero pronoun candidates and a set of noun phrase candidates are then extracted. If W is the leftmost word in the word sequence that is spanned by some VP node, the gap G that is immediately to the left of W qualifies as a ZP candidate. For example, in Figure 1, gaps immediately to the left of the two occurrences of 继续, and 增加, 占, 上升 are all ZP candidates. All noun phrases[4] that are either maximal NPs or modifier NPs qualify as NP candidates. For example, in Figure 1, NP_1, NP_2, NP_3, NP_5, and NP_6 are all NP candidates. With these ZP and NP candidate extractions, the recalls of ZPs and NPs are 100% and 98.6%, respectively.

[4]A noun phrase can either be NP or QP in CTB. We simply use NP hereafter.

After the ZP and NP candidates are determined, we perform AZP identification and resolution in a sequential manner. We build two classifiers, the AZP identification classifier and the AZP resolution classifier. The AZP identification classifier determines the position of AZPs, while the AZP resolution classifier finds an antecedent noun phrase for each AZP identified by the AZP identification classifier. Both classifiers are built using machine learning techniques. The features of both classifiers are largely syntactic features based on parse trees and are easily computed.

We perform 5-fold cross validation on the training data set to tune parameters and to pick the best model. We then retrain the best model with all data in the training data set, and apply it to the blind test set. In the following sections, all accuracies reported on the training data set are based on 5-fold cross validation.

4 Anaphoric Zero Pronoun Identification

We use machine learning techniques to build the AZP identification classifier. The features are described in Table 2.

In the feature description, Z is the ZP candidate. Let W_l and W_r be the words immediately to the left and to the right of Z, respectively, P the parse tree node that is the lowest common ancestor node of W_l and W_r, P_l and P_r the child nodes of P that are ancestor nodes of W_l and W_r, respectively. If Z is the first gap of the sentence, W_l, P, P_l, and P_r are all NA. Furthermore, let V be the highest VP node in the parse tree that is immediately to the right of Z, i.e., the leftmost word in the word sequence that is spanned by V is W_r. If Z is not the first gap in the sentence, define the ceiling node C to be P, otherwise to be the root node of the parse tree. In the example shown in Figure 1, for the ZP candidate ϕ_2 (which is immediately to the left of 占), W_l, W_r, P, P_l, P_r, V, and C are "，", 占, IP_1, IP_2, IP_3, VP_3, and IP_1, respectively. Its feature values are also shown in Table 2.

To train an AZP identification classifier, we gen-

Feature	Description	ϕ_2
First_Gap	If Z is the first gap in the sentence, T; else F.	F
P_l_Is_NP	If Z is the first gap in the sentence, NA; otherwise, if P_l is an NP node, T; else F.	F
P_r_Is_VP	If Z is the first gap in the sentence, NA; otherwise, if P_r is a VP node, T; else F.	F
P_l_Is_NP & P_r_Is_VP	If Z is the first gap in the sentence, NA; otherwise, if P_l is an NP node and P_r is a VP node, T; else F.	F
P_Is_VP	If Z is the first gap in the sentence, NA; otherwise, if P is a VP node, T; else F.	F
IP-VP	If in the path from W_r to C, there is a VP node such that its parent node is an IP node, T; else F.	T
Has_Ancestor_NP	If V has an NP node as ancestor, T; else F.	T
Has_Ancestor_VP	If V has a VP node as ancestor, T; else F.	F
Has_Ancestor_CP	If V has a CP node as ancestor, T; else F.	T
Left_Comma	If Z is the first gap, NA; otherwise if W_l is a comma, T; else F.	T
Subject_Role	If the grammatical role of Z is subject, S; else X.	X
Clause	If V is in a matrix clause, an independent clause, a subordinate clause, or none of the above, the value is M, I, S, X, respectively.	I
Is_In_Headline	If Z is in the headline of the text, T; else F.	F

Table 2: Features for anaphoric zero pronoun identification. The feature values of ϕ_2 are shown in the last column.

erate training examples from the training data set. All ZP candidates in the training data set generate training examples. Whether a training example is positive or negative depends on whether the ZP candidate is an AZP.

After generating all training examples, we train an AZP identification classifier using the J48 decision tree learning algorithm in Weka[5]. During testing, each ZP candidate is presented to the learned classifier to determine whether it is an AZP. We conduct experiments to measure the performance of the model learned. The results of 5-fold cross validation on the training data set are shown in Table 3.

Model	R	P	F
Heuristic	99.7	15.0	26.1
AZP Ident	19.8	51.1	28.6
AZP Ident ($r = 8$)	59.8	44.3	50.9

Table 3: Accuracies of AZP identification on the training data set under 5-fold cross validation.

We use heuristic rules as a baseline for comparison. The rules used by the heuristic model are as follows. For a node T in the parse tree, if

1. T is a VP node; and
2. T's parent node is not a VP node; and
3. T has no left sibling, or its left sibling is not an NP node,

then the gap that is immediately to the left of the word sequence spanned by T is an AZP. This simple AZP identification heuristic achieves an F-measure of 26.1%.

Imbalanced Training Data

From Table 3, one can see that the F-measure of the machine-learned AZP identification model is 28.6%, which is only slightly higher than baseline heuristic model. It has a relatively high precision, but much lower recall. The problem lies in the highly imbalanced number of positive and negative training examples. Among all the 155 texts in the training set, there are 343 positive and 10,098 negative training examples. The ratio r of the number

[5] http://www.cs.waikato.ac.nz/ml/weka/

of negative training examples to the number of positive training examples is 29.4. A classifier trained on such highly imbalanced training examples tends to predict more testing examples as negative examples. This explains why the precision is high, but the recall is low.

To overcome this problem, we vary r by varying the weight of the positive training examples, which is equivalent to sampling more positive training examples. The values of r that we have tried are $1, 2, 3, \ldots, 29$. The larger the value of r, the higher the precision, and the lower the recall. By tuning r, we get a balance between precision and recall, and hence an optimal F-measure. Figure 2 shows the effect of tuning r on AZP identification. When $r = 8$, the optimal F-measure is 50.9%, which is much higher than the F-measure without tuning r.

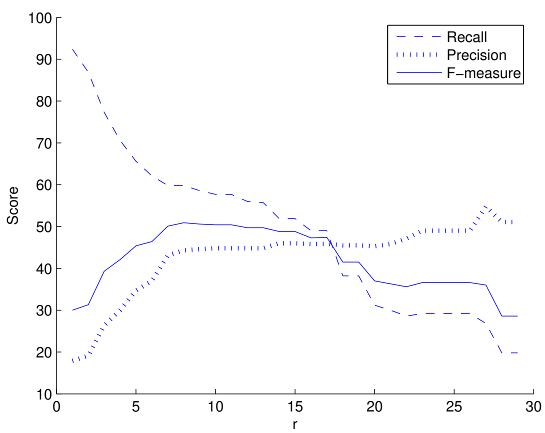

Figure 2: Effect of tuning r on AZP identification

Ng and Cardie (2002a) reported that the accuracies of their noun phrase anaphoricity determination classifier were 86.1% and 84.0% for the MUC6 and MUC7 data sets, respectively. Noun phrases provide much fruitful information for anaphoricity identification. However, useful information such as gender, number, lexical string, etc, is not available in the case of zero pronouns. This makes AZP identification a much more difficult task, and hence it has a relatively low accuracy.

5 Anaphoric Zero Pronoun Resolution

In anaphoric zero pronoun resolution, we also use machine learning techniques to build a classifier. The features are described in Table 4.

In the feature description, Z is the anaphoric zero pronoun that is under consideration, and A is the potential NP antecedent for Z. V is the same as in AZP identification. The feature values of the pair NP_1 and ϕ_2 (the gap immediately to the left of 占) in Figure 1 are shown in Table 4.

To train the AZP resolution classifier, we generate training examples in the following way. An AZP Z and its immediately preceding coreferential NP antecedent A in the gold standard coreference chain form a positive training example. Between A and Z, there are other NP candidates. Each one of these NP candidates, together with Z, form a negative training example. This is similar to the approach adopted in Soon et al. (2001). We also train the AZP resolution classifier using the J48 decision tree learning algorithm.

After building both AZP identification and resolution classifiers, we perform AZP identification and resolution in a sequential manner. For a ZP candidate Z, the AZP identification classifier determines whether Z is an AZP. If it is an AZP, all NP candidates that are to the left of Z in textual order are considered as potential antecedents. These potential antecedents are tested from right to left. We start from the NP candidate A_1 that is immediately to the left of Z. A_1 and Z form a pair. If the pair is classified as positive by the resolution classifier, A_1 is the antecedent for Z. If it is classified as negative, we proceed to the NP candidate A_2 that is immediately to the left of A_1, and test again. The process continues until we find an antecedent for Z, or there is no more NP candidate to test.

This right-to-left search attempts to find the closest correct antecedent for an AZP. We do not choose the best-first search strategy proposed by Ng and Cardie (2002b). This is because we generate training examples and build the resolution classifier by pairing each zero pronoun with its closest preceding antecedent. In addition, a zero pronoun is typically not too far away from its antecedent. In our data set, 92.6% of the AZPs have antecedents that are at most 2 sentences apart. Our experiment shows that this closest-first strategy performs better than the best-first strategy for Chinese AZP resolution.

Table 5 shows the experimental results of 5-fold cross validation on the training data set. For com-

Feature	Description	NP$_1$-ϕ_2
	Features between Z and A	
Dist_Sentence	If Z and A are in the same sentence, 0; if they are one sentence apart, 1; and so on.	0
Dist_Segment	If Z and A are in the same segment (where a segment is a sequence of words separated by punctuation marks including "，", "；", "。", "！", and "？"), 0; if they are one segment apart, 1; and so on.	1
Sibling_NP_VP	If Z and A are in different sentences, F; Otherwise, if both A and Z are child nodes of the root node, and they are siblings (or at most separated by one comma), T; else F.	F
Closest_NP	If A is the closest preceding NP candidate to Z, T; else F.	T
	Features on A	
A_Has_Anc_NP	If A has an ancestor NP node, T; else F.	F
A_Has_Anc_NP_In_IP	If A has an ancestor NP node which is a descendant of A's lowest ancestor IP node, T; else F.	F
A_Has_Anc_VP	If A has an ancestor VP node, T; else F.	F
A_Has_Anc_VP_In_IP	If A has an ancestor VP node which is a descendant of A's lowest ancestor IP node, T; else F.	F
A_Has_Anc_CP	If A has an ancestor CP node, T; else F.	F
A_Grammatical_Role	If the grammatical role of A is subject, object, or others, the value is S, O, or X, respectively.	S
A_Clause	If A is in a matrix clause, an independent clause, a subordinate clause, or none of the above, the value is M, I, S, X, respectively.	M
A_Is_ADV	If A is an adverbial NP, T; else F.	F
A_Is_TMP	If A is a temporal NP, T; else F.	F
A_Is_Pronoun	If A is a pronoun, T; else F.	F
A_Is_NE	If A is a named entity, T; else F.	F
A_In_Headline	If A is in the headline of the text, T; else F.	F
	Features on Z	
Z_Has_Anc_NP	If V has an ancestor NP node, T; else F.	T
Z_Has_Anc_NP_In_IP	If V has an ancestor NP node which is a descendant of V's lowest ancestor IP node, T; else F.	F
Z_Has_Anc_VP	If V has an ancestor VP node, T; else F.	F
Z_Has_Anc_VP_In_IP	If V has an ancestor VP node which is a descendant of V's lowest ancestor IP node, T; else F.	F
Z_Has_Anc_CP	If V has an ancestor CP node, T; else F.	T
Z_Grammatical_Role	If the grammatical role of Z is subject, S; else X.	X
Z_Clause	If V is in a matrix clause, an independent clause, a subordinate clause, or none of the above, the value is M, I, S, X, respectively.	I
Z_Is_First_ZP	If Z is the first ZP candidate in the sentence, T; else F.	F
Z_Is_Last_ZP	If Z is the last ZP candidate in the sentence, T; else F.	F
Z_In_Headline	If Z is in the headline of the text, T; else F.	F

Table 4: Features for anaphoric zero pronoun resolution. The feature values of the pair NP$_1$ and ϕ_2 are shown in the last column.

parison, we show three baseline systems. In all three baseline systems, we do not perform AZP identification, but directly apply the AZP resolution classifier. In the first baseline, we apply the AZP resolution classifier on all ZP candidates. In the second baseline, we apply the classifier only on ZPs annotated in the gold standard, instead of all ZP candidates. In the third baseline, we further restrict it to resolve only AZPs. The F-measures of the three baselines are 2.5%, 27.6%, and 40.6% respectively.

Model	R	P	F
All ZP Candidates	40.5	1.3	2.5
Gold ZP	40.5	20.9	27.6
Gold AZP	40.5	40.6	40.6
AZP Ident (r=8 t=0.5)	23.6	17.5	20.1
AZP Ident (r=11 t=0.6)	22.4	20.3	21.3

Table 5: Accuracies of AZP resolution on the training data set under 5-fold cross validation.

Tuning of Parameters

Ng (2004) showed that an NP anaphoricity identification classifier with a cut-off threshold t = 0.5 pruned away many correct anaphoric NPs and harmed the overall recall. By varying t, the overall resolution F-measure was improved. We adopt the same tuning strategy and accept a ZP candidate ZP_i as an AZP and proceed to find its antecedent only if $P(ZP_i) \geq t$. The possible values for t that we have tried are $0, 0.05, 0.1, \ldots, 0.95$.

In Section 4, we show that $r = 8$ yields the best AZP identification F-measure. When we fix $r = 8$ and vary t, the overall F-measure for AZP resolution is the best at $t = 0.65$, as shown in Figure 3. We then try tuning r and t at the same time. An overall optimal F-measure of 21.3% is obtained when $r = 11$ and $t = 0.6$. We compare this tuned F-measure with the F-measure of 20.1% at $r = 8$ and $t = 0.5$, obtained without tuning t. Although the improvement is modest, it is statistically significant ($p < 0.05$).

6 Experimental Results

In the previous section, we show that when $r = 11$ and $t = 0.6$, our sequential AZP identification and resolution achieves the best F-measure under 5-fold cross validation on the 155 training texts. In order to utilize all available training data, we generate

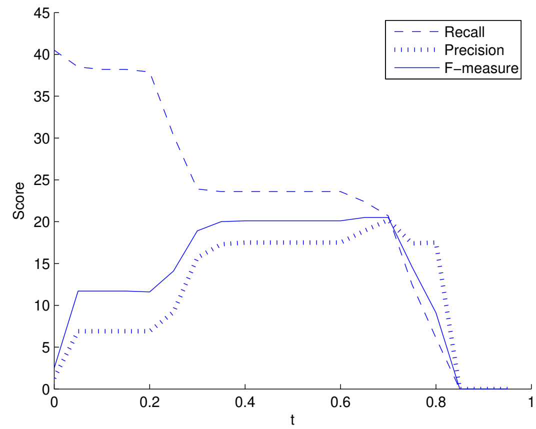

Figure 3: Effect of tuning t on AZP resolution

training examples for the AZP identification classifier with $r = 11$, and generate training examples for the AZP resolution classifier, on all 155 training texts. Both classifiers are trained again with the newly generated training examples. We then apply both classifiers with anaphoricity identification cut-off threshold $t = 0.6$ to the blind test data. The results are shown in Table 6.

R	P	F
27.5	24.4	25.9

Table 6: Accuracies of AZP resolution on blind test data.

By utilizing all available information on the gold standard parse trees, Converse (2006) finds an antecedent for each AZP given that all AZPs are correctly input to her system. The accuracy of her rule-based approach is 43.0%. For comparison, we determine the antecedents for AZPs in the gold standard annotation, under 5-fold cross validation on all 205 texts in the corpus. The recall, precision, and F-measure are 42.3%, 42.7%, and 42.5%, respectively. This shows that our proposed machine learning approach for Chinese zero pronoun resolution is comparable to her state-of-the-art rule-based approach.

7 Related Work

Converse (2006) assumed that the gold standard Chinese anaphoric zero pronouns and the gold standard parse trees of the texts in Penn Chinese Tree-

Bank (CTB) were given as input to her system, which performed resolution of the anaphoric zero pronouns using the Hobbs algorithm (Hobbs, 1978). Her system did not identify the anaphoric zero pronouns automatically.

Yeh and Chen (2004) proposed an approach for Chinese zero pronoun resolution based on the Centering Theory (Grosz et al., 1995). Their system used a set of hand-engineered rules to perform zero pronoun identification, and resolved zero pronouns with a set of hand-engineered resolution rules.

In Iida et al. (2006), they proposed a machine learning approach to resolve zero pronouns in Japanese using syntactic patterns. Their system also did not perform zero pronoun identification, and assumed that correctly identified zero pronouns were given as input to their system.

The probabilistic model of Seki et al. (2002) both identified and resolved Japanese zero pronouns, with the help of a verb dictionary. Their model needed large-scale corpora to estimate the probabilities and to prevent data sparseness.

Ferrández and Peral (2000) proposed a hand-engineered rule-based approach to identify and resolve zero pronouns that are in the subject grammatical position in Spanish.

8 Conclusion

In this paper, we present a machine learning approach to the identification and resolution of Chinese anaphoric zero pronouns. We perform both identification and resolution automatically, with two sets of easily computable features. Experimental results show that our proposed learning approach achieves anaphoric zero pronoun resolution accuracy comparable to a previous state-of-the-art, heuristic rule-based approach. To our knowledge, our work is the first to perform both identification and resolution of Chinese anaphoric zero pronouns using a machine learning approach.

Obviously, there is much room for improvement. In future, we plan to apply our model directly on machine-generated parse trees. We also plan to classify non-coreferential zero pronouns into the six categories.

Acknowledgements

We thank Susan Converse and Martha Palmer for sharing their Chinese third-person pronoun and zero pronoun coreference corpus.

References

Susan Converse. 2006. *Pronominal Anaphora Resolution in Chinese.* Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania.

Antonio Ferrández and Jesús Peral. 2000. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000)*, pages 166–172.

Radu Florian, Hany Hassan, Abraham Ittycheriah, Hongyan Jing, Nanda Kambhatla, Xiaoqiang Luo, Nicolas Nicolov, and Salim Roukos. 2004. A statistical model for multilingual entity detection and tracking. In *Proceedings of the Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics Annual Meeting 2004 (HLT-NAACL2004)*, pages 1–8.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2006. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL2006)*, pages 625–632.

Young-Joo Kim. 2000. Subject/object drop in the acquisition of Korean: A cross-linguistic comparison. *Journal of East Asian Linguistics*, 9(4):325–351.

Charles N. Li and Sandra A. Thompson. 1979. Third-person pronouns and zero-anaphora in Chinese discourse. *Syntax and Semantics*, 12:311–335.

Wendan Li. 2004. Topic chains in Chinese discourse. *Discourse Processes*, 37(1):25–45.

Andrew McCallum and Ben Wellner. 2005. Conditional models of identity uncertainty with application to noun coreference. In *Advances in Neural Information Processing Systems 17 (NIPS)*, pages 905–912.

Vincent Ng and Claire Cardie. 2002a. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING2002)*, pages 1–7.

Vincent Ng and Claire Cardie. 2002b. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL2002)*, pages 104–111.

Vincent Ng. 2004. Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL2004)*, pages 152–159.

Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. 2002. A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING2002)*, pages 911–917.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2003. Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL2003)*, pages 176–183.

Ching-Long Yeh and Yi-Chun Chen. 2004. Zero anaphora resolution in Chinese with shallow parsing. *Journal of Chinese Language and Computing*.

Yaqian Zhou, Changning Huang, Jianfeng Gao, and Lide Wu. 2005. Transformation based Chinese entity detection and tracking. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP 2005)*, pages 232–237.