Chinese Zero Pronoun Resolution: An Unsupervised Approach Combining Ranking and Integer Linear Programming

Chen Chen and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{yzcchen,vince}@hlt.utdallas.edu

Abstract

State-of-the-art approaches to Chinese zero pronoun resolution are supervised, requiring training documents with manually resolved zero pronouns. To eliminate the reliance on annotated data, we propose an unsupervised approach to this task. Underlying our approach is the novel idea of employing a model trained on manually resolved *overt* pronouns to resolve *zero* pronouns. Experimental results on the OntoNotes 5.0 corpus are encouraging: our unsupervised model surpasses its supervised counterparts in performance.

Introduction

A zero pronoun (ZP) is a gap in a sentence that is found when a phonetically null form is used to refer to a real-world entity. An anaphoric zero pronoun (AZP) is a ZP that corefers with one or more preceding noun phrases (NPs) in the associated text. Below is an example taken from the Chinese TreeBank (CTB), where *pro* is used to denote a ZP.

俄罗斯作为米洛舍夫维奇一贯的支持者，*pro* 曾经提出调停这场政治危机。

Russia is a consistent supporter of Milošević, *pro* has proposed to mediate this political crisis.

In this example, the antecedent of *pro* is 俄罗斯 (Russia). The ability to correctly interpret ZPs is crucial to the automatic processing of pro-drop languages. Note that ZPs lack grammatical attributes essential for overt pronoun resolution, such as Number and Gender. This makes ZP resolution more challenging than overt pronoun resolution.

ZP resolution is composed of two steps. The first step, AZP identification, involves extracting ZPs that are anaphoric. The second step, AZP resolution, aims to identify an antecedent for an AZP. State-of-the-art ZP resolvers have tackled both steps in a supervised manner, training a classifier for AZP identification and another one for AZP resolution (e.g., Zhao and Ng (2007), Chen and Ng (2013)).

In this paper, we focus on the second task, AZP resolution. In other words, we assume that there is a separate process that identifies AZPs, and our goal is to resolve the given AZPs in a document. Note that the task of AZP resolution alone is by no means easy: for instance, given gold-standard AZPs, state-of-the-art supervised resolvers only achieve an F-score of 47.7% for *resolving* Chinese AZPs (Chen and Ng 2013).

Our contribution in this paper lies in the proposal of an *unsupervised language-independent* approach to AZP resolution. In other words, our approach does not require any data with manually resolved AZPs and is applicable to any language where such annotated data is not readily available. Underlying our approach is a novel, unexplored hypothesis: to resolve an AZP, we could apply a model trained on manually resolved *overt* pronouns to *rank* its candidate antecedents. In other words, we recast unsupervised AZP resolution as a supervised ranking problem, where we employ training data composed of manually resolved overt pronouns only. In addition, we show how our ranking model can be enhanced by incorporating grammatical compatibility information via integer linear programming (ILP). Results on resolving the Chinese AZPs in the OntoNotes 5.0 corpus are encouraging: our unsupervised approach achieves results that surpass those achieved by state-of-the-art supervised AZP resolvers.

The rest of the paper is organized as follows. After discussing related work and the grammatical properties of Chinese overt pronouns, we describe how we train our ranking model on overt pronouns and apply it to resolve AZPs. We then show how the model can be enhanced by incorporating grammatical compatibility information via ILP. Finally, we present evaluation results and an analysis of the errors.

Related Work

Chinese ZP resolution. Early approaches to Chinese ZP resolution are rule-based. Converse (2006) applied Hobbs' algorithm (Hobbs 1978) to resolve the ZPs in the CTB documents. Yeh and Chen (2007) hand-engineered a set of rules for ZP resolution based on Centering Theory (Grosz, Joshi, and Weinstein 1995).

Recent approaches to this task are based on *supervised* learning. Zhao and Ng (2007) are the first to employ a supervised machine learning approach to Chinese ZP resolution. They trained an AZP resolver by employing a set of syntactic and positional features in combination with a decision tree learner. Unlike Zhao and Ng, Kong and

Zhao (2010) employed context-sensitive convolution tree kernels (Zhou, Kong, and Zhu 2008) to model syntactic information. Extending Zhao and Ng's (2007) feature set, Chen and Ng (2013) improved supervised Chinese ZP resolution by proposing novel features to capture the contextual information between candidate antecedents and ZPs. They also exploited the coreference links between ZPs as bridges to find far-away antecedents for ZPs.

ZP resolution for other languages. There have been rule-based and supervised machine learning approaches for resolving ZPs in other languages. For example, to resolve ZPs in Spanish texts, Ferrández and Peral (2000) proposed a set of hand-crafted rules that encode preferences for candidate antecedents. In addition, supervised approaches have been extensively employed to resolve ZPs in Korean (e.g., Han (2006)) and Japanese (e.g., Seki, Fujii, and Ishikawa (2002), Isozaki and Hirao (2003), Iida, Inui, and Matsumoto (2006; 2007)). More recently, Iida and Poesio (2011) have applied ILP to resolve Japanese and Italian ZPs, but their goal when applying ILP is completely different from ours. Specifically, they used ILP to coordinate the decisions made by two classifiers trained in a supervised manner on ZP coreference annotations, one for ZP *detection* and the other for ZP *resolution*. In contrast, we will see in a later section that we use ILP to coordinate the decisions made by two different components regarding which overt pronoun should be used to fill a ZP gap.

Ranking for coreference resolution. While we are the first to recast unsupervised AZP resolution as a supervised ranking task, we are not the first to employ supervised ranking for coreference resolution. Connolly, Burger, and Day (1994), Iida et al. (2003), and Yang et al. (2003) trained a decision tree-based pairwise ranker that ranks two candidate antecedents for an anaphoric NP. Advances in machine learning have enabled Denis and Baldridge (2007; 2008) to train a maximum entropy ranker that simultaneously ranks all the candidate antecedents for an anaphoric NP. Extending this idea, Rahman and Ng (2009) have trained a model for ranking the partial coreference clusters preceding an anaphoric NP.

Chinese Overt Pronouns

As mentioned before, our approach relies heavily on Chinese *overt* pronouns. Specifically, we exploit ten personal pronouns, including 你 (singular you), 我 (I), 他 (he), 她 (she), 它 (it), 你们 (plural you), 我们 (we), 他们 (masculine they), 她们 (feminine they), and 它们 (impersonal they). These ten pronouns are chosen because they have well-defined grammatical attribute values.

Each of the overt pronouns can be uniquely identified by four grammatical attributes, Number, Gender, Person, and Animacy. Number has two values, *singular* and *plural*. Gender has three values, *neuter*, *masculine* and *feminine*. Person has three values, *first*, *second* and *third*. Finally, Animacy has two values, *animate* and *inanimate*. The attribute values associated with each pronoun are shown in Table 1. These four attributes are crucial because they are used to represent an overt pronoun in our approach.

Pronouns	Number	Gender	Person	Animacy
我 (I)	singular	neuter	first	animate
你 (you)	singular	neuter	second	animate
他 (he)	singular	masculine	third	animate
她 (she)	singular	feminine	third	animate
它 (it)	singular	neuter	third	inanimate
你们 (you)	plural	neuter	second	animate
我们 (we)	plural	neuter	first	animate
他们 (they)	plural	masculine	third	animate
她们 (they)	plural	feminine	third	animate
它们 (they)	plural	neuter	third	inanimate

Table 1: Chinese overt pronoun attributes.

Ranking Model

In this section, we describe how we (1) train a ranking model to rank candidate antecedents for an anaphoric overt pronoun, and (2) apply the resulting model to resolve AZPs.

Training the ranker. We create training instances as follows. Each training instance corresponds to an anaphoric overt pronoun op and one of its candidate antecedents, c, and is represented using the 36 features shown in Table 2. Some of these features are employed in state-of-the-art supervised AZP resolvers (Zhao and Ng 2007; Chen and Ng 2013), while others are commonly used in overt pronoun resolution (e.g., features that encode semantic compatibility and agreement in Number, Gender, Person, and Animacy).[1]

Since our goal is to learn a ranker for ranking the candidate antecedents of an overt pronoun, each set of training instances created from the same overt pronoun corresponds to a *ranking problem*. The question, then, is: how can we assign a rank value to each candidate antecedent? The answer depends on the algorithm we use to train the ranker. In our experiments, we use YASMET[2], a maximum entropy-based ranking toolkit. Recall that YASMET can be used to train a ranking model that, when given an unseen ranking problem, distributes probability mass over the instances in the ranking problem. In the context of our pronoun resolution problem, we want the ranker to give those instances corresponding to correct antecedents a higher probability mass than those corresponding to incorrect antecedents. As a result, we assign rank values to the training instances as follows. Assume that S_{op} is the set of training instances created from op. If c is a correct antecedent of op, its rank value is $\frac{1}{|coref|}$, where $|coref|$ is the number of candidate antecedents that are coreferent with op. Otherwise, its rank value is 0.

Two points deserve mention. First, to avoid having to consider a potentially large number of candidate antecedents (and thus unnecessarily complicating the ranking task), we consider all and only those NPs that are at most two sentences away from an overt pronoun to be its candidate antecedents.[3] Second, we create training instances from an overt pronoun

[1] Since a ZP is by definition a null pronoun, all the features that are applicable to ZPs are also applicable to overt pronouns.

[2] http://www.fjoch.com/YASMET.html

[3] Only 8% of the overt pronouns do not have any antecedent in the preceding two sentences.

Features between c and op (14)	the sentence distance between c and op; the segment distance between c and op, where segments are separated by punctuations; whether c is the closest NP to op; whether c and op are siblings in the associated parse tree; the four concatenations of Number, Gender, Person and Animacy attributes of c and op; whether c is also a subject and its predicate verb is identical with v; whether c is the nearest candidate antecedent with subject grammatical role and is semantically compatible[4] with v, if not, whether c is the first semantically compatible candidate antecedent encountered; the concatenation of the head of c and v; the concatenation of the head of c, v and the head of object if exists; the concatenation of the head of c and the punctuation at the end of sentence that op is in.
Features on c (12)	whether c has an ancestor NP, and if so, whether this NP is a descendent of c's lowest ancestor IP; whether c has an ancestor VP, and if so, whether this VP is a descendent of c's lowest ancestor IP; whether c has an ancestor CP; the grammatical role of c; the clause type in which c appears; whether c is an adverbial NP, a temporal NP, a pronoun, or a named entity; whether c is in the headline of the text.
Features on op (10)	whether v has an ancestor NP, and if so, whether this NP node is a descendent of v's lowest ancestor IP; whether v has an ancestor VP, and if so, whether this VP is a descendent of v's lowest ancestor IP; whether v has an ancestor CP; the grammatical role of op; the type of the clause in which v appears; whether op is the first or last to-be-resolved pronoun in the sentence; whether op is the beginning of a sentence; whether op is the beginning of an IP cause; whether op is in the headline of the text.

Table 2: Features used to represent an instance. op is the overt pronoun, and c is candidate antecedent. v is the predicate verb after op.

op in a training text if and only if it satisfies the following conditions: (1) op is one of the 10 overt pronouns described in the previous section; (2) the closest antecedent of op is at most two sentences away from it; and (3) op is a surface or deep subject in the corresponding sentence. Condition (2) ensures that there is something to rank in each ranking problem (i.e., not all instances have the same rank). Condition (3) is motivated by our observation that 99.56% of the ZPs in our corpus (i.e., OntoNotes 5.0) are surface or deep subjects. We impose this condition so that the ranker can focus its effort on ranking overt pronouns that are subjects.[5]

Applying the ranker. After training, we can apply the ranker to resolve the AZPs in the test set. The question is: since the ranker was trained on overt pronouns, how can it be applied to resolve AZPs? Specifically, the issue is that some of the features the ranker employs are derived from the four grammatical attributes of overt pronouns, so it will not be applicable to AZPs as they lack such attributes.

To address this issue, we explore a new idea: for each AZP zp to be resolved, we fill the gap left behind by zp with an *overt* pronoun. The question, then, is: which of the 10 overt pronouns should we use to fill the gap? This is not a trivial question: if it were easy to find the right overt pronoun to fill the gap, AZP resolution would not be more difficult than overt pronoun resolution. Hence, rather than attempting to answer this non-trivial question, we fill the gap with each of the 10 overt pronouns. Specifically, for each overt pronoun op, we fill the gap with op and then create test instances in the same way as the training instances. Hence, assuming that zp has $|C|$ candidate antecedents, the total number of test instances created from zp is $10 \times |C|$.

Before applying the ranker to these test instances, we make the ranker's job easier by reducing the number of test instances. Specifically, in some of these test instances, the candidate antecedent and the overt pronoun are not compatible with respect to all four grammatical attributes we defined earlier (e.g., candidate antecedent 电脑 (the computer) is incompatible with overt pronoun 你们 (plural you) with respect to Animacy, Person, and Number).[6] To prevent the ranker from assigning non-zero probability mass to these linguistically implausible cases, we remove them and apply the ranker to rank the remaining ones. We then select the candidate antecedent associated with the most probable test instance created from AZP zp as its antecedent.

Although we have successfully converted unsupervised AZP resolution into overt pronoun resolution above, the resolution procedure can be improved further. The improvement is motivated by a problem we observed previously (Chen and Ng 2013): an AZP and its closest antecedent can sometimes be far away from each other, thus making it difficult to correctly resolve the AZP. To address this problem, we employ the following resolution procedure in our experiments. Given a test document, we process its AZPs in a left-to-right manner. As soon as we resolve an AZP to a preceding NP c, we fill the corresponding AZP's gap with c. Hence, when we process an AZP zp, all of its preceding AZPs in the associated text have been resolved, with their gaps filled by the NPs they are resolved to. To resolve AZP zp, we create test instances between zp and its candidate antecedents in the same way as described before. The only difference is that the set of candidate antecedents of zp may now include those NPs that are used to fill the gaps of the AZPs resolved so far. In other words, this incremental resolution procedure may increase the number of candidate antecedents (and hence the number of test instances) for each AZP zp. Some of these additional candidate antecedents are closer to zp than the original candidate antecedents, thereby facilitating the resolution of zp. If the ranker resolves zp to the additional candidate antecedent that fills the gap left behind by, say, AZP zp', we postprocess the output by resolving zp to the NP that zp' is resolved to.[7]

[4] We employ Bergsma and Lin's approach (2006) to compute semantic compatibility. See their paper for details.

[5] This is by no means a limitation of our approach: if we were given a corpus in which many ZPs occur as grammatical objects, we could similarly train another ranker on overt objects.

[6] We compute the attribute values of a candidate antecedent heuristically, in essentially the same way as these values are computed for an NP in English pronoun resolution.

[7] This postprocessing step is needed because the additional candidate antecedents are only gap fillers.

Enforcing Pronoun-Verb Compatibility

Recall that when applying the ranker, we enforced compatibility between an overt pronoun and the associated candidate antecedent in a test instance with respect to the four grammatical attributes. A natural question is: can we similarly enforce compatibility between the overt pronoun and its governing verb in a test instance with respect to the four grammatical attributes? Note that enforcing this pronoun-verb compatibility amounts to ensuring that the overt pronoun satisfies all the grammatical constraints the governing verb places on its subject.[8]

Somewhat unfortunately, many Chinese verbs place little or even no grammatical constraints on their subject NPs. More specifically, while Chinese verbs have the same Animacy and Gender constraints on their subject NPs as English verbs, the vast majority of them do not have any Number or Person constraints on their subject NPs. The reason is that Chinese has no morphology. This implies that the form of a Chinese verb does not change as its subject's Number and Person change. For example, consider the verb 宣布 (announce). It turns out that in this case, the subject of 宣布 (announce) is unconstrained with respect to not only Gender and Person, but also Animacy and Number.

The fact that Chinese verbs place little or no grammatical constraints on their subject NPs seems to suggest that enforcing pronoun-verb compatibility is unlikely to have a big impact on AZP resolution. Nevertheless, we hypothesize that pronoun-verb compatibility could be enforced in a better way. Specifically, we make the following observation: while many Chinese verbs do not have *hard* grammatical constraints on their subject NPs, many of them have grammatical *preferences* for their subject NPs. For example, 解决 (resolve) prefers an animate subject, and 打扮 (dress up), when used in a Chinese context, prefers a feminine subject. Such grammatical preferences could be employed as soft constraints when enforcing pronoun-verb compatibility.

The question, then, is: how can we obtain the grammatical preferences of a verb? Rather than manually specifying such preferences, we learn them from a large, unannotated corpus. Specifically, for each verb v that governs an AZP in our test corpus, we first collect from the Chinese Gigaword corpus (Graff and Chen 2003) the set of NPs that serve as the subject of v. Then, for each possible value b of each of the four grammatical attributes a, we compute $P_a(b)$, the probability that an NP in this set has b as its value for a. Finally, we use the resulting probabilities to represent a verb's grammatical preferences for its subject NP.

Now that we know how to compute a verb's grammatical preferences for its subject NP, the question is: how can they be used as soft constraints to enforce pronoun-verb compatibility? To answer this question, note that given a verb v governing an AZP zp, its grammatical preferences tell us which overt pronoun it prefers to use to fill the gap left behind by zp, e.g., if v prefers an animate, singular, and neuter subject, it implies that v prefers the gap to be filled by 我 (I) or 你 (singular you). In other words, 我 (I) and 你 (singular you) are more compatible with v than the remaining pronouns.

So far, we have seen that given an AZP's gap to be filled, a verb has a preference for which pronouns should be used to fill the gap. The question, then, is: how can a verb's preference be used to improve AZP resolution? Before answering this question, recall from the previous section that for each AZP zp in the test set, the ranker assigns a probability $P(op, c)$ to each test instance $i(op, c)$ created from zp, where $i(op, c)$ corresponds to an overt pronoun op that fills the AZP's gap and one of its candidate antecedents c. $P(op, c)$ can be interpreted as the ranker's preference for selecting c as the antecedent of zp and using op to fill the gap left behind by zp. Given that both the ranker and the governing verb have their own preferences for the pronoun filling the AZP's gap, we coordinate their preferences using ILP by defining an objective function as the linear combination of the probabilities encoded in these preferences.

Before describing the objective function, let us introduce some notation. The set $A = \{Num, Gen, Per, Ani\}$ has four elements, which correspond to Number, Gender, Person and Animacy respectively. We use a to denote an attribute in A; V_a to denote the set of possible values of a; op_a to denote overt pronoun op's value for attribute a; and PR to denote the set of the 10 Chinese overt pronouns we employ. As described before, for each AZP zp, $P(op, c)$ is the probability the ranker assigns the instance created from overt pronoun op and candidate antecedent c, and $P_a(b)$ is the probability that the value of attribute a is b according to the grammatical preferences of the verb governing zp.

In addition, we have to define binary indicator variables whose values are to be determined by an ILP solver. Specifically, we define $x(op, c)$ to be a variable that takes on the value 1 if and only if the solver selects c to be zp's antecedent and fills the gap left behind by zp with op. We define another variable $y_a(b)$, whose value is 1 if and only if the solver decides that the gap left behind by zp should be filled by an overt pronoun whose attribute a has value b.

For each AZP zp, we create one ILP program whose objective function is a linear combination of $P(op, c)$ and the four probabilities that encode a verb's preference, $P_{Num}(b)$, $P_{Gen}(b)$, $P_{Per}(b)$, and $P_{Ani}(b)$, as shown below:

$$\begin{aligned}&\operatorname*{argmax}_{op,c}[\sum_{op\in PR}\sum_{c\in C} P(op,c)x(op,c)+\\ &\alpha \sum_{b\in V_{Num}} P_{Num}(b)y_{Num}(b)+\beta\sum_{b\in V_{Gen}} P_{Gen}(b)y_{Gen}(b)+\\ &\gamma\sum_{b\in V_{Per}} P_{Per}(b)y_{Per}(b)+\delta\sum_{b\in V_{Ani}} P_{Ani}(b)y_{Ani}(b)]\end{aligned} \tag{1}$$

subject to the following constraints:

$$x(op, c) \in \{0, 1\}, \forall op \in PR, \forall c \in C \tag{2}$$

$$\sum_{op\in PR}\sum_{c\in C} x(op, c) = 1 \tag{3}$$

$$y_a(b) \in \{0, 1\}, \forall a \in A, \forall b \in V_a \tag{4}$$

[8] We focus on enforcing subject-verb compatibility but not verb-object compatibility because 99.56% of the ZPs in our corpus occur as subjects. If we were given a corpus in which many ZPs occur as objects, we could similarly enforce verb-object compatibility.

$$\sum_{b=1}^{|a|} y_a(b) = 1, \forall a \in A \quad (5)$$

$$y_{Ani}(animate) \geq y_{Gen}(masculine) + y_{Gen}(feminine) \quad (6)$$

The four parameters, α, β, γ and δ, denote the relative importance of the terms in the objective function and are jointly tuned to maximize F-score on development data.[9]

Constraints (2) and (4) ensure that $x(op, c)$ and $y_a(b)$ are binary values. Constraint (3) ensures that exactly one overt pronoun is used to fill the AZP's gap and exactly one candidate antecedent is selected as the AZP's antecedent. Constraint (5) ensures that the overt pronoun used to fill the AZP's gap has exactly one value for each attribute. Constraint (6) ensures that if the overt pronoun string is masculine or feminine, then it has to be animate.

Note that we need an additional constraint to ensure the choice of op takes into account both the ranker's preference and the preference of the verb v governing the AZP:

$$\sum_{op_a=b} \sum_{c \in C} x(op, c) = y_a(b), \forall a \in A, \forall b \in V_a \quad (7)$$

Constraint (7) ensures the consistency between the choice of op and the value $y_a(b)$ for each attribute $a \in A$. With this constraint, v affects the choice of op through the last four terms in the objective function and thereafter has an impact on the choice of the AZP's antecedent.

Evaluation

Experimental Setup

Dataset. For evaluation, we employ the Chinese portion of the OntoNotes 5.0 corpus that was used in the official CoNLL-2012 shared task. In the CoNLL-2012 data, only the training set and development set contain ZP coreference annotations, while the test set does not.[10] Therefore, we employ the training set of the CoNLL-2012 data for model training and parameter tuning, and perform evaluation on the CoNLL-2012 development set.[11] Statistics on the datasets are shown in Table 3. The documents in these datasets come from six sources, including Broadcast News (BN), Newswire (NW), Broadcast Conversation (BC), Telephone Conversation (TC), Web Blog (WB) and Magazine (MZ).

Evaluation measures. Following Zhao and Ng (2007) and Chen and Ng (2013), we express the results of AZP resolution in terms of recall (R), precision (P) and F-score (F).

Evaluation setting. Since we focus on AZP resolution, we assume that gold AZPs and gold parse trees are given.

	Training	Test
Documents	1,391	172
Sentences	36,487	6,083
Words	756,063	110,034
Qualified Overt Pronouns	9,239	–
AZPs	–	1,713

Table 3: Statistics on the training and test sets.

Results

Baseline systems. We employ three baseline systems: (1) Zhao and Ng (2007); (2) Kong and Zhou (2010); and (3) Chen and Ng (2013). All three baseline systems are *supervised*, meaning that they are trained on data manually annotated with the antecedents of AZPs.

Table 4 shows the overall scores (row 1) and the per-source scores (rows 2 to 7). The parenthesized number beside a source's name is the number of AZPs in that source. As we can see, the best-performing baseline is the Chen and Ng baseline: it significantly outperforms the Zhao and Ng baseline and the Kong and Zhou baseline by 6.2% and 2.8% in F-score respectively.[12] For per-source results, the Chen and Ng baseline yields the best scores on four sources: it only underperforms Zhao and Ng's system on NW and Kong and Zhou's system on BN.

Our ranker. The performance of our ranker is also shown in Table 4. Although our ranker does not exploit any AZP coreference annotations, it significantly beats Kong and Zhou's system ($p < 0.07$) and Zhao and Ng's system by 1.2% and 4.6% in F-score respectively, and only significantly underperforms the Chen and Ng baseline by 1.6% in F-score.

Our ILP method. Next, we examine the performance of our approach after imposing constraints via ILP. In comparison to the ranking results, we can see that after applying ILP, the overall F-score increases significantly by 2.6%. In addition, our approach beats the best baseline system significantly by 1.0% ($p < 0.06$), achieving the best overall F-score on this dataset reported to date.

While our approach's overall F-score is significantly better than that of the best baseline, it outperforms the best baseline in only three of the six sources (namely WB, BC and TC). These results suggest that further performance gains might be possible if we combine our system with the best baseline via ensemble learning, for instance.

In an attempt to better understand the role of joint inference in our approach, we conduct an experiment where we replace ILP with a pipeline approach. Specifically, we (1) determine the Gender, Number, Animacy, or Person attributes of the overt pronoun for filling an AZP's gap based on pronoun-verb compatibility, and then (2) apply the ranker to rank the test instances created based on the grammatical attributes determined in the first step. Our results show that the pipeline method significantly underperforms our approach: its F-score is only 32.2%. We attribute the poor performance

[9] We attempted values of 0, 0.02, 0.04, 0.06, 0.08, and 0.10 for each of these parameters.

[10] AZP resolution is not part of the CoNLL-2012 shared task.

[11] We tune the parameters (i.e., α, β, γ and δ) as follows. We first train a ranking model on 90% of the training data and use the remaining 10% for parameter tuning. Then we retrain the model on all of the training data before applying it to the test data.

[12] All significance tests are paired t-tests. Unless otherwise stated, $p < 0.05$.

	Baseline Systems									Our Approach					
	Kong and Zhou			Zhao and Ng			Chen and Ng			Ranking Model			ILP Method		
Source	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
Overall (1713)	44.9	44.9	44.9	41.5	41.5	41.5	47.7	47.7	47.7	45.9	46.4	46.1	48.4	48.9	**48.7**
NW (84)	34.5	34.5	34.5	40.5	40.5	**40.5**	38.1	38.1	38.1	31.0	31.0	31.0	38.1	38.1	38.1
MZ (162)	32.7	32.7	32.7	28.4	28.4	28.4	34.6	34.6	**34.6**	29.0	29.2	29.1	30.9	31.1	31.0
WB (284)	45.4	45.4	45.4	40.1	40.1	40.1	46.1	46.1	46.1	45.4	45.4	45.4	50.4	50.4	**50.4**
BN (390)	51.0	51.0	**51.0**	43.1	43.1	43.1	47.2	47.2	47.2	45.1	45.1	45.1	45.9	45.9	45.9
BC (510)	43.5	43.5	43.5	44.7	44.7	44.7	52.7	52.7	52.7	50.2	50.7	50.4	53.5	54.1	**53.8**
TC (283)	48.4	48.4	48.4	42.8	42.8	42.8	51.2	51.2	51.2	53.7	56.1	**54.9**	53.7	56.1	**54.9**

Table 4: Resolution results on the test set. The strongest F-score in each row is boldfaced.

	Chen and Ng			Ranking Model			ILP Method		
	R	P	F	R	P	F	R	P	F
without	46.2	46.2	46.2	44.2	44.9	44.5	47.1	47.8	47.4
with	47.7	47.7	47.7	45.9	46.4	46.1	48.4	48.9	48.7

Table 5: AZP resolution results without and with the improved resolution procedure.

of the pipeline approach to the difficulty of correctly selecting an overt pronoun to fill an AZP's gap: as mentioned before, many Chinese verbs do not have hard grammatical constraints on their subject NPs.

To determine the impact of the improved resolution procedure, which exploits ZP coreference links, on resolution performance, we repeat our experiments *without* using it. Row 1 of Table 5 shows the overall resolution results of the Chen and Ng baseline, the ranking model, and the ILP method *without* using the improved resolution procedure. For ease of comparison, the corresponding results obtained with the improved resolution procedure are shown in row 2 of the table. As we can see, employing the improved procedure significantly boosts the F-score across the board.

Qualitative Error Analysis

In an attempt to better understand our approach, we perform a qualitative analysis of its errors. Our analysis reveals that it makes two major types of errors, as discussed below.

Incorrect choice of overt pronouns as gap fillers. To determine which overt pronoun *op* should be used to fill the gap of AZP zp, ILP considers the verb v governing zp. Though v offers a strong hint for determining the attribute values of zp, there are circumstances where we may need more contextual information to make the right decision. Filling the gap with an incorrect overt pronoun may in turn cause the zp to be incorrectly resolved, as shown in the following example:

(我) 想念 (你)。*pro1* 祝 *pro2* 平安。
(I) miss (you). *pro1* hope *pro2* are safe.

The first sentence has two overt pronouns, 我 (I) and 你 (singular you), whereas the second sentence has two AZPs, which are marked as *pro1* and *pro2* respectively. Assume that *pro1* has been correctly resolved to 我 (I), and that the current AZP to be resolved is *pro2*. As we include resolved AZPs as additional candidate antecedents, the candidate antecedents set for *pro2* contains three mentions, i.e., 我 (I), 你 (you) and *pro1*, which has been filled by its coreferential mention, 我 (I). The word governing *pro2*, 平安 (are safe), has the part-of-speech VA, which is a kind of Chinese verb. According to the gold standard, the overt pronoun that fills the gap of *pro2* should be the second person 你 (singular you), and the correct antecedent for *pro2* should be 你 (you). However, ILP incorrectly fills the gap with the first person 我 (I) and mistakenly links *pro2* to *pro1*. Since *pro1*'s antecedent is 我 (I), *pro2* is further resolved to 我 (I). The reason why ILP makes the wrong decision is that given the verb 平安 (are safe), it has no preference between using a first-person pronoun and using a second-person pronoun to fill the gap of *pro2*. If ILP could prefer a second-person pronoun to a first-person pronoun, then it might be able to correctly resolve *pro2*. For this to happen, it may need to take more context into account, e.g., the verb 祝 (hope) before *pro2* in this example. If ILP knew that the subject after 祝 (hope) was more likely to be second person, it might fill the gap of *pro2* with a second-person pronoun and subsequently resolve *pro2* correctly.

Incorrect resolution of ZPs with long-distance antecedents. Because AZPs and their closest antecedents are usually close to each other in the training data, our approach has acquired the *recency* preference (i.e., the preference for candidate antecedents that are closer to the AZP under consideration). Such preference has contributed in part to the poor resolution of AZPs whose closest antecedents are far away from them, as shown in the following example:

(八里乡) 位于 (((台北) 盆地) 西北端)。(行政区) 隶属于 (台北县)，*pro* 为台北县廿九个乡镇市之一。
(Bali Town) is located in the (Northwest of ((Taipei) Basin)). (Its administrative area) is affiliated with (Taipei County), *pro* is one of the 29 towns and cities.

Although our approach correctly fills the ZP gap marked as *pro* with 它 (it), it incorrectly resolves it to 行政区 (Its administrative area). The reason is that the correct antecedent, 八里乡 (Bali Town), is far away from *pro*: there are five candidate antecedents between *pro* and 八里乡 (Bali Town). Note, however, that it is easy for a human to resolve *pro* to 八里乡 (Bali Town) because the whole passage is discussing 八里乡 (Bali Town). Hence, to correctly handle such cases, one may construct a topic model over the passage and assign each candidate mention a prior probability so that the resulting system favors the selection as antecedents those mentions representing the topics.

Conclusions and Future Work

We investigated an unsupervised approach to Chinese zero pronoun resolution, exploiting the novel idea of training a ranker on overt pronoun coreference annotations and enhancing it with ILP constraints. Our approach achieves the best result reported to date on the OntoNotes 5.0 dataset. In future work, we plan to (1) improve our ILP method by using additional context to predict pronoun attributes; (2) apply our approach to other pro-drop languages; and (3) evaluate our approach on automatically identified AZPs.

Acknowledgments

We thank the three anonymous reviewers for their detailed and insightful comments on an earlier draft of the paper. This work was supported in part by NSF Grants IIS-1147644 and IIS-1219142. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of NSF.

References

Bergsma, S., and Lin, D. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 33--40.

Chen, C., and Ng, V. 2013. Chinese zero pronoun resolution: Some recent advances. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1360--1365.

Connolly, D.; Burger, J. D.; and Day, D. S. 1994. A machine learning approach to anaphoric reference. In *Proceedings of International Conference on New Methods in Language Processing*, 255--261.

Converse, S. 2006. *Pronominal Anaphora Resolution in Chinese*. Ph.D. Dissertation, University of Pennsylvania.

Denis, P., and Baldridge, J. 2007. A ranking approach to pronoun resolution. In *Proceedings of the Twentieth International Conference on Artificial Intelligence*, 1588--1593.

Denis, P., and Baldridge, J. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 660--669.

Ferrández, A., and Peral, J. 2000. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 166--172.

Graff, D., and Chen, K. 2003. Chinese Gigaword. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Grosz, B. J.; Joshi, A. K.; and Weinstein, S. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2):203--226.

Han, N.-R. 2006. *Korean zero pronouns: Analysis and resolution*. Ph.D. Dissertation, University of Pennsylvania.

Hobbs, J. 1978. Resolving pronoun references. *Lingua* 44:311--338.

Iida, R., and Poesio, M. 2011. A cross-lingual ILP solution to zero anaphora resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 804--813.

Iida, R.; Inui, K.; Takamura, H.; and Matsumoto, Y. 2003. Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the EACL Workshop on The Computational Treatment of Anaphora*, 23--30.

Iida, R.; Inui, K.; and Matsumoto, Y. 2006. Exploting syntactic patterns as clues in zero-anaphora resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 625--632.

Iida, R.; Inui, K.; and Matsumoto, Y. 2007. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Transactions on Asian Language Information Processing* 6(4):Article 12.

Isozaki, H., and Hirao, T. 2003. Japanese zero pronoun resolution based on ranking rules and machine learning. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 184--191.

Kong, F., and Zhou, G. 2010. A tree kernel-based unified framework for Chinese zero anaphora resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 882--891.

Rahman, A., and Ng, V. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, 968--977.

Seki, K.; Fujii, A.; and Ishikawa, T. 2002. A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In *Proceedings of the 19th International Conference on Computational linguistics - Volume 1*, 1--7.

Yang, X.; Zhou, G.; Su, J.; and Tan, C. L. 2003. Coreference resolution using competitive learning approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 176--183.

Yeh, C.-L., and Chen, Y.-C. 2007. Zero anaphora resolution in Chinese with shallow parsing. *Journal of Chinese Language and Computing* 17(1):41--56.

Zhao, S., and Ng, H. T. 2007. Identification and resolution of Chinese zero pronouns: A machine learning approach. In *Proceedings of the 2007 Joint Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning*, 541--550.

Zhou, G.; Kong, F.; and Zhu, Q. 2008. Context-sensitive convolution tree kernel for pronoun resolution. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, 25--31.