Chinese Overt Pronoun Resolution: A Bilingual Approach

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

Much research has been done on the problem of English pronoun resolution, but there has been relatively little work on the corresponding problem of Chinese pronoun resolution. While pronoun resolution in both languages remains a challenging task, Chinese pronoun resolution is further complicated by (1) the lack of publicly available Chinese word lists or dictionaries that can be used to look up essential mention attributes such as gender and number; and (2) the relative dearth of Chinese coreference-annotated data. Existing approaches to Chinese pronoun resolution are *monolingual*, training and testing a pronoun resolver on Chinese data. In contrast, we propose a *bilingual* approach to Chinese pronoun resolution, aiming to improve the resolution of Chinese pronouns by leveraging the publicly available English dictionaries and coreference annotations. Experiments on the OntoNotes 5.0 corpus demonstrate that our bilingual approach to Chinese pronoun resolution significantly surpasses the performance of state-of-the-art monolingual approaches.

Introduction

Recent years have seen a surge of interest in multilingual anaphora and coreference resolution, as evidenced by the organization of shared tasks that focus on multilingual coreference, such as the series of ACE evaluations, the SemEval-2010 shared task on Coreference Resolution in Multiple Languages (Recasens et al. 2010), and the CoNLL-2012 shared task on Modeling Multilingual Unrestricted Coreference in OntoNotes (Pradhan et al. 2012). An important outcome of these shared tasks is the creation of annotated coreference data in multiple languages. For example, a corpus composed of coreference-annotated documents in six European languages was released as part of the SemEval-2010 shared task, and both the OntoNotes 5.0 corpus and the ACE corpus, which are released as a consequence of the CoNLL-2012 shared task and the ACE evaluations respectively, are composed of coreference-annotated documents in English, Chinese, and Arabic. These corpora have enabled the development of corpus-based approaches to anaphora and coreference resolution in languages other than English.

Our goal in this paper is to improve the state of the art in Chinese *overt* pronoun resolution.[1] Given the availability of Chinese coreference-annotated data, a natural question is: how can a learning-based Chinese pronoun resolver be built? Perhaps the simplest solution, which is also the solution adopted by virtually all the participants of the CoNLL-2012 shared task and the ACE evaluations, is to build it in exactly the same way as an English resolver, meaning that we can just train on the available Chinese training data and employ the same kind of features that are typically used for English pronoun resolution. In principle, there is no problem with constructing a Chinese pronoun resolver using this *monolingual* approach, because basic linguistic constraints on coreference, such as agreement in gender, number, and semantic class, are applicable to both English and Chinese.

In practice, however, such a monolingual approach to Chinese pronoun resolution may not work as well as it does for English pronoun resolution for at least two reasons. First, while there exist publicly available English word lists that can be used to look up essential mention attributes such as gender and number, such resources are not available for Chinese. In fact, gender and number word lists are provided by the CoNLL-2012 shared task organizers for English but not Chinese. Second, the amount of coreference-annotated data available for training Chinese resolvers is far less than that available for training English resolvers. For example, the English training data used for the CoNLL-2012 shared task consists of 1.6 million tokens, whereas the Chinese training data consists of only 950,000 tokens.

Motivated by these observations, we propose a *bilingual* approach to Chinese pronoun resolution, seeking to improve the state of the art by leveraging two kinds of English resources, namely (1) publicly available word lists and dictionaries for computing essential mention attributes such as gender and number; and (2) existing English coreference annotations. Experiments on the OntoNotes 5.0 corpus demonstrate that our bilingual approach significantly surpasses the performance of state-of-the-art monolingual approaches.

[1] There are two kinds of pronouns in Chinese, namely overt and zero pronouns. In this paper, we will focus exclusively on the resolution of overt pronouns. For the sake of simplicity, we will use the term *pronouns* to refer to *overt pronouns* throughout the paper.

Approach: An Overview

At the core of our approach is a combination of two methods for exploiting English word lists and coreference annotations to resolve Chinese pronouns, as described below.

In the first method, we begin by machine-translating all the Chinese training documents into English, so that each Chinese mention is mapped to its English counterpart in the translated text.[2] Then, we create the instances for training a Chinese resolver as in the typical monolingual approach, meaning that the feature vector will be composed of features computed based on the Chinese pronoun to be resolved (call it p) and one of its candidate antecedents (call it c), as well as features capturing the relationship between p and c. We then *augment* the feature vector with features computed for the English pronoun to which p is mapped and the English candidate antecedent to which c is mapped. Finally, we train a Chinese pronoun resolver on the feature-augmented training instances as in the monolingual approach. Note that feature augmentation allows us to exploit English word lists and dictionaries for computing essential attributes of a mention such as gender and number, which in turn helps us to resolve the Chinese pronouns. Of course, the usefulness of these English features for Chinese pronoun resolution depends in part on whether there is a natural correspondence between gender and number in English and Chinese. Fortunately, there is a direct correspondence, meaning that an English name/nominal always has the same number and gender as the corresponding Chinese name/nominal. Also note that to employ this method, both the Chinese training documents and the Chinese test documents need to be machine-translated into English, as feature augmentation is needed in the creation of both the training and test instances.

Unfortunately, the aforementioned method does not enable us to exploit English coreference-annotated data. To address this shortcoming, we introduce a second method wherein we train an English pronoun resolver on only the English coreference-annotated training data. Each training instance is represented using features computed based on the English pronoun to be resolved and one of its candidate antecedents, as well as features that capture the relationship between the two. During testing, we first machine-translate the Chinese document whose pronouns are to be resolved into English, so that each Chinese mention is mapped to its English counterpart in the translated text. Then, we apply the English resolver acquired in the training step to resolve the English mentions in the translated text. Finally, we project the resolution results back to the Chinese side to resolve Chinese pronouns. Unlike the first method, this method exploits both the English coreference annotations and the English word lists for computing the attributes of an English mention. In addition, we only need to machine-translate the Chinese test documents, as the training step merely involves training an English resolver on the English training documents.

A natural question is: which method should we employ to resolve Chinese pronouns? We hypothesize that an *ensemble* method that combines both of the methods above would work better than any of these methods alone. The reason is that the first method exploits only the Chinese coreference-annotated data in the training step, whereas the second method exploits only the English coreference-annotated data in the training step. Hence, only when we combine both methods can we exploit the coreference-annotated data for both English and Chinese.

[2] In practice, word alignment errors do not permit each Chinese mention to be mapped to an English mention. We will address this issue when detailing our bilingual approach in the next section.

Approach: Implementation Details

In this section, we describe the details involved in implementing the bilingual approach to Chinese pronoun resolution that we outlined in the previous section. As we will see, our approach requires the training of three pronoun resolvers, namely a Chinese pronoun resolver that is trained using the typical monolingual approach, an English resolver trained using the second method described above, and a mixed resolver trained using feature augmentation, as described in the first method above. Before we describe the details of how these three resolvers are trained, recall that the English resolver requires that a test Chinese document be machine-translated into English so that each Chinese mention can be mapped to an English mention, and the mixed resolver requires that both the training and the test Chinese documents be machine-translated into English.[3] We therefore begin by describing the document preprocessing step, which includes details on how the Chinese mentions are mapped to the English mentions after machine translation (MT).

Document Preprocessing

Recall that the output of MT is a pseudo parallel corpus consisting of Chinese-English sentence pairs that are translations of each other. The goal of the preprocessing step is to align (1) the words in each sentence pair and (2) the mentions in each sentence pair.

Word alignment. We align the words in each pair of sentences using BerkeleyAligner[4]. Note that BerkeleyAligner outputs a posterior probability P_a for each aligned word pair indicating the probability that the two words in the pair should be aligned. We filter those aligned pairs whose probability is below a predefined threshold.

Mention alignment. Next, we extract the mentions from each Chinese document and each English document using Chen and Ng's (2012) and Björkelund and Farkas's (2012) mention detectors, respectively.[5] We then align each Chinese mention M_c with an English mention using the two rules:

Head rule. Let H_{M_c} be M_c's head word, and H'_{M_e} be the highest-probability English word to which H_{M_c} should be mapped according to the word aligner. If H'_{M_e} is a head word for some mention M_e, we will align M_c with M_e.

Boundary rule. If the head rule fails, we employ the boundary rule. Let L_{M_c} be the leftmost word of M_c and

[3] In our experiments, we use Google Translate (translate.google.com) for machine translation.

[4] http://code.google.com/p/berkeleyaligner/

[5] These two systems were the top-performing systems in the CoNLL-2012 shared tasks (Pradhan et al. 2012).

R_{M_c} be the rightmost word of M_c. We (1) find the highest-probability English words L'_{M_c} and R'_{M_c} to which L_{M_c} and R_{M_c} should be aligned respectively; (2) create an English mention starting with L'_{M_c} and ending with R'_{M_c}; and (3) align M_c with this English mention.

Classifier Training

Next, we train our three resolvers, the Chinese resolver, the English resolver, and the mixed resolver. Each resolver is a binary classifier that determines whether a pronoun m_k and one of its candidate antecedents m_j are coreferent or not (Soon, Ng, and Lim 2001), where the set of candidate antecedents of m_k is the set of mentions preceding m_k in the associated text. All these *mention-pair* models are trained on the training data using the SVM learning algorithm implemented in the LIBSVM software package (Chang and Lin 2011), which returns a value between 0 and 1 that indicates the probability that m_k and m_j are co-referring.

Training an English pronoun resolver, PR^E. Training instances for PR^E are created from the English training texts using Soon, Ng, and Lim's (2001) method. Specifically, we create (1) a positive instance for each anaphoric pronoun m_k and its closest antecedent m_j; and (2) a negative instance for m_k paired with each of the intervening mentions, $m_{j+1}, m_{j+2}, \ldots, m_{k-1}$. Each instance is represented using a set of features employed by Björkelund and Farkas's (2012) coreference resolver, one of the top performing systems in the CoNLL-2012 shared task. Linguistically, the feature set consists of lexical, grammatical, syntactic, semantic, and positional features, as well as conjunctions of these features.[6]

Training a Chinese pronoun resolver, PR^C. Training instances for PR^C are created from the Chinese training texts using Soon, Ng, and Lim's (2001) method. Each instance is represented using the features employed by Björkelund and Farkas's Chinese coreference resolver. Note that the features used by their English and Chinese resolvers are essentially the same, except that (1) the gender and number agreement features are only present in the English feature set, since these features were not provided for the Chinese documents by the CoNLL-2012 shared task organizers; and (2) features are conjoined differently in the two languages owing to the use of different heuristics.

Training a mixed pronoun resolver, PR^M. Next, we train a mixed pronoun resolver that employs a stronger feature set composed of information extracted from both languages. The training instances for PR^M are a subset of those created for PR^C. Specifically, while the training instances for PR^C are created from *all* anaphoric Chinese pronouns, those for PR^M are created only from those anaphoric Chinese pronouns that have been aligned to some English pronoun. The features representing the instance, however, fall into two groups. The first group consists of features computed from m_j and m_k by PR^C, where m_k is a Chinese anaphoric overt pronoun and m_j is one of its candidate antecedents. The second group consists of features computed from $m_{j'}$ and $m_{k'}$ by PR^E, where $m_{j'}$ and $m_{k'}$ are the English counterparts of m_j and m_k according to our mention alignment algorithm.

[6]The complete list of features can be found in the source code of the resolver. See http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/IMSCoref.en.html. The linguistic annotations used to compute these features, such as POS tags and syntactic parse trees, are automatically created.

Classification of Test Instances

After training, each of these resolvers can be applied independently to classify test instances. Recall that PR^C is used to classify test instances created directly from the Chinese test documents. PR^M is also used to classify test instances created directly from the Chinese test documents, except that each instance is represented using features computed from both the Chinese mention and the English mention involved. On the other hand, PR^E is used to classify test instances created from the translated Chinese documents.

What this implies is that given a pronoun m_k to be resolved in a Chinese test document, PR^C will always produce some result for m_k (i.e., either resolving it to some preceding mention or determining that it has no antecedent). On the other hand, PR^M and PR^E may not always produce a result for m_k. More specifically, recall that for PR^M, a test instance will be created from a pronoun if and only if it can be mapped to some English mention. Hence, if m_k is not mapped to any English mention, no instances will be created for it, and as a result, PR^M will not produce any result for it. The situation is similar for PR^E. Specifically, since PR^E classifies instances created from the translated Chinese documents, if m_k is not mapped to any English mention in the translated text, PR^M will not produce any result for it.

In sum, given a test instance composed of a Chinese pronoun, m_k, and one of its candidate antecedents, m_j, if both m_k and m_j can be mapped to some English mentions, then each of the three resolvers will return a value that indicates the probability that m_j and m_k are coreferent. However, if m_k or m_j is not mapped, then we assume that the value returned by PR^M and PR^E for the test instance formed from m_k and m_j is NA (**N**ot **A**pplicable).

Resolution Methods

Next, we describe how we combine the three resolvers above to resolve a Chinese pronoun in a test text. We investigate four resolution methods. For notational convenience, we assume that m_k is the pronoun to be resolved.

Method 1. We resolve m_k to the closest preceding mention whose coreference probability with m_k according to PR^E is at least 0.5. If PR^E returns NA or a value of less than 0.5 for all of m_k's preceding mentions, then we resolve m_k to the closest preceding mention whose coreference probability with m_k is at least 0.5 using PR$_C$.

Method 2. Method 2 is the same as Method 1 except that PR^E is replaced with PR^M.

Method 3. Unlike the previous two resolution methods, which exploit only two of the three resolvers, the next two resolution methods exploit all three resolvers.

Recall that each of the three resolvers independently returns a value for each of m_k's candidate antecedents that indicates the probability that it is coreferent with m_k (though

the returned value might be NA for PR^M and PR^E). In Method 3, we take the unweighted average of these three probabilities for each candidate antecedent, and resolve m_k to the closest candidate antecedent whose average coreference probability is at least 0.5. Note that NA values will be ignored when computing the unweighted average.

Method 4. Let us first define some notation. Let m_j denote a candidate antecedent of m_k, and P^C_{jk}, P^E_{jk}, and P^M_{jk} be the coreference probabilities between m_j and m_k according to PR^C, PR^E and PR^M, respectively. In Method 4, we resolve m_k to the closest preceding mention m_j if at least one of four conditions is satisfied: (1) $P^C_{jk} > t_C$; (2) $P^M_{jk} > t_M$; (3) $P^E_{jk} > t_E$; and (4) $P^{norm}_{jk} \geq 0.5$, where t_C, t_M, and t_E are thresholds to be tuned, and P^{norm} is a probability that we will define shortly. In essence, the first three conditions say that if any of the three mention-pair models is confident that m_j is the correct antecedent of m_k (by virtue of the fact that the corresponding coreference probability is above some confidence threshold), then we resolve m_k to m_j. Otherwise, we check the fourth condition, $P^{norm} \geq 0.5$, where

$$P^{norm}_{jk} = \frac{P^C_{jk} + P^E_{jk} w_e(P_a)^{w_{ae}} + P^M_{jk} w_m(P_a)^{w_{am}}}{1 + w_e(P_a)^{w_{ae}} + w_m(P_a)^{w_{am}}}.$$

In this formula, the numerator is a weighted combination of P^C, P^E, and P^M (i.e., the coreference probabilities returned by the three pronoun resolvers), and the denominator normalizes P^{norm}_{jk} to a value between 0 and 1. A closer inspection of the formula reveals that the weight associated with P^E has two components, w_e and $P_a^{w_{ae}}$. The weight parameter w_e, which can be thought of as indicating the relative importance of PR^E in the decision-making process, is fine-tuned by $P_a^{w_{ae}}$, where P_a is the probability returned by the alignment model[7], and w_{ae} adjusts the degree of influence of P_a on the weight associated with P^E. The two weights associated with P^M can be interpreted similarly.

As we can see, the four conditions have seven tunable parameters. We jointly tune them to maximize performance (which in this case is F-score) on held-out development data using a hill-climbing local search algorithm, where we tune one parameter at a time while holding the remaining parameters fixed.

Finally, note that Method 4 can be applied only if all three resolvers return coreference probabilities for each of m_k's candidate antecedents. Hence, if PR^E or PR^M returns NA for a given candidate antecedent, we replace NA with the value returned by PR^C.

Evaluation

Experimental Setup

Corpus. We use the OntoNotes 5.0 corpus that we obtained from the CoNLL-2012 shared task organizers for evaluating our bilingual approach to Chinese pronoun resolution.

[7]We define P_a as the minimum of the probability of aligning the head words of the Chinese and English pronominal anaphors and the probability of aligning the head words of the Chinese and English candidate antecedents.

	Type	Train	Dev	Test	Total
CH	Docs	1,391	172	166	1,729
	Words	750K	110K	90K	950K
	Chains	28,257	3,875	3,559	35,691
	Mentions	102,854	14,183	12,801	129,838
EN	Docs	1,940	222	222	2,384
	Words	1.3M	160K	170K	1.6M
	Chains	35,143	4,546	4,532	44,221
	Mentions	155,560	19,156	19,764	194,480

Table 1: Statistics on the Chinese (CH) and English (EN) data in the OntoNotes 5.0 corpus.

	Resolution Method	R	P	F	A^a	A^{na}	A^o
1	Closest-first	71.7	65.3	68.4	71.7	59.3	67.4
2	Best-first	72.0	65.6	68.7	72.0	59.3	67.6
3	Best Shared Task	63.8	67.5	65.6	63.8	**76.7**	68.2
4	Rahman and Ng	64.3	65.2	64.7	64.3	68.5	65.8
5	Method 1	65.6	64.4	65.0	65.6	66.0	65.8
6	Method 2	**73.0**	65.1	68.8	**73.0**	56.7	67.4
7	Method 3	71.5	70.5	71.0	71.5	67.6	70.2
8	Method 4	71.1	**71.5**	**71.3**	71.1	70.4	**70.8**

Table 2: Resolution results on the OntoNotes 5.0 test set. The strongest result in each column is boldfaced.

Statistics on the Chinese and English coreference-annotated data that we employed are shown in Table 1. We follow the shared task's train-test partition of the documents, performing training and parameter tuning on the training and development documents and reserving the test documents solely for evaluation purposes. Specifically, when Methods 1−3 are employed, we train the three resolvers on the combined training-development set; on the other hand, since Method 4 requires parameter tuning, we train the resolvers on the training set and tune the parameters on the development set.

Evaluation metrics. We report results in terms of recall (R), precision (P), and F-score (F) on resolving *anaphoric* pronouns. Hence, P and R increase with the number of correctly resolved pronouns. Also, P increases with the number of unresolved non-anaphoric pronouns.

We additionally report results in terms of (1) the percentage of anaphoric pronouns correctly resolved (A^a), which is equivalent to the recall number mentioned above; (2) the percentage of non-anaphoric pronouns *not* resolved (A^{na}); and (3) the overall accuracy (A^o), which is computed as the sum of the number of correctly resolved anaphoric pronouns and the number of unresolved non-anaphoric pronouns divided by the total number of pronouns.

Results and Discussion

Results are shown in Table 2.

Baseline 1: Monolingual approach. This baseline has two versions. Both versions are trained in the same manner as PR^C, and differ only in how the pronouns are resolved. Baseline 1a employs closest-first clustering, resolving a pronoun to the closest coreferent candidate antecedent (row 1). Baseline 1b employs best-first clustering, resolving a pronoun to the candidate with the highest coreference probability (row 2). As we can see, they are statistically indistinguish-

able with respect to all six evaluation metrics.[8]

Baseline 2: Best shared task system. To gauge the performance of Baseline 1, we run the best-performing Chinese coreference resolver in the shared task on this test data.[9] Since this resolver outputs coreference chains, we assume that a pronoun m_k is correctly resolved if its closest antecedent in the system chain appears in the gold chain containing m_k (row 3). As we can see, this resolver performs significantly worse than Baseline 1b, owing to large drops in recall accompanied by smaller drops in precision. This information is also reflected in its lower A^a and higher A^{na}. Its A^o is slightly higher than but not statistically distinguishable from that of Baseline 1b.

Baseline 3: Rahman and Ng's (2012) approach. The resolution approach closest to ours is Rahman and Ng's. They assume a setting in which coreference-annotated data is only available in one language (which in their case is English), and their goal is to resolve a pronoun by applying a three-step *annotation projection* approach where they (1) machine-translate the texts in the target language into English, (2) apply the coreference resolver trained on the English coreference-annotated data to each translated text, and (3) project the coreference chains back to the target language. In essence, their approach is the same as our Method 1, except that no backoff model will be used to resolve a pronoun p if p or any of its candidate antecedents cannot be mapped to an English mention.

In an attempt to obtain stronger baseline results, we applied our reimplementation of Rahman and Ng's approach to our test data. Results are shown in row 3. As we can see, Baseline 1b significantly outperforms this baseline with respect to both F-score and A^o. Consequently, we will compare our approach against Baseline 1b, the best baseline.

Our bilingual approach. Rows 5−8 show the results of our approach.

Row 5 shows the results of Method 1, where we resolve a Chinese pronoun m_k using PR^E and backoff to PR^C only if m_k is not mapped to any English pronoun. As we can see, this method yields significantly worse results than the baseline with respect to both F-score and A^o as a result of a large drop in recall.

Row 6 shows the results of Method 2, which differs from Method 1 in that PR^E is replaced with PR^M. Two points deserve mention. First, in comparison to Method 1, we see that F-score increases significantly by more than 3.8% and A^o increases significantly by 1.6%, owing to a considerable rise in recall. These results suggest that PR^M, which combines features from both languages, is indeed a stronger resolver than the monolingual PR^E.

Second, its F-score and accuracy are indistinguishable from those of the best baseline. Since the major difference between the baseline (PR^C) and Method 2 ($PR^M + PR^C$) lies in whether the features from the English side (e.g., gender, number) are used, these statistically indistinguishable results seem to suggest that English gender and number are not useful for resolving Chinese pronouns. To understand whether this is indeed the case, we examine the pairwise R, P, and F of PR^C and PR^M on the test set. The R/P/F scores are 35.8/63.1/45.7 for PR^C and 40.0/57.8/47.2 for PR^M. These results show that PR^M performs better than PR^C at the pairwise level, suggesting that the English features are indeed useful for resolving Chinese pronouns.

Rows 7 and 8 show the results of Methods 3 and 4, where resolution decisions are made by taking the unweighted and weighted averages of the probabilities returned by the three models, respectively. The F-score and A^o of Method 4 are statistically indistinguishable from those of Method 3. In addition, the F-score and A^o of both methods are significantly better than those of Method 2, owing to a substantial increase in precision. These results suggest that (1) even the simple unweighted averaging scheme can perform as well as weighted averaging via automatic parameter tuning; and (2) employing all three models offers better results than employing only two models.

Analysis of Results

Comparison with the Monolingual Baseline

To better understand why our best-performing approach (the one employing Method 4) is superior to the best baseline (the monolingual baseline using best-first clustering), we analyze some cases where our approach makes the correct decision and the best baseline fails.

The NEUTER attribute. Recall that the CoNLL-2012 shared task organizers provided the participants with a semantic resource that can be used to label an English mention with its gender (i.e., *Masculine*, *Feminine* or *Neuter* (Bergsma and Lin 2006), but such a resource is not available for Chinese. As a result, our monolingual baseline may incorrectly link a pronoun to a candidate antecedent that is incompatible with respect to the *NEUTER* attribute. Consider the following example[10]:

(牙根) 已经愈合了。所以 (他) 笑起来很漂亮，而且牙齿也很好。
(Tooth root) has healed. So (he) laughs very beautiful, and the teeth are very good.

The pronoun to be resolved, 他 (he), is *Masculine*, but its candidate antecedent, 牙根 (Tooth root), is *Neuter*. The baseline posits them as coreferent, largely owing to the fact that 牙根 (Tooth root) is the pronoun's nearest subject. However, using the aforementioned English semantic resource, our approach knows that the candidate antecedent's English counterpart has the attribute *Neuter* and therefore correctly classifies this pair as not coreferent.

The NUMBER attribute. Unlike English nouns, Chinese nouns do not inflect in number. Thus, if a noun is not premodified by a number indicator such as 一些 (some), 许多

[8] All significance tests are paired t-tests, with $p < 0.05$.

[9] The system is available from www.hlt.utdallas.edu/~yzcchen/coreference.

[10] For all the examples shown in this subsection, the first line is the original Chinese sentence, whereas the second line is its English translation obtained via Google Translate.

	Resolution Method	Machine Trans. R	P	F	Human Trans. R	P	F
1	Closest-first	63.0	62.7	62.8	63.0	62.7	62.8
2	Best-first	62.3	62.0	62.2	62.3	62.0	62.2
3	Best Shared Task	55.2	65.8	60.1	55.2	65.8	60.1
4	Rahman and Ng	54.7	58.1	56.4	46.1	59.9	52.1
5	Method 1	55.6	57.4	56.5	56.3	59.8	58.0
6	Method 2	**65.6**	59.8	62.5	**65.7**	61.7	63.7
7	Method 3	61.7	**66.9**	64.2	63.6	66.7	65.1
8	Method 4	63.8	65.3	**64.5**	64.5	**67.0**	**65.7**

Table 3: Resolution results on the 400-document parallel corpus obtained via five-fold cross validation. The strongest result in each column is boldfaced.

(many) and the suffix 们, it is generally hard to determine whether a noun is *Plural* or not. However, when a MT system translates a plural Chinese noun into English, it may be able to correctly turn the noun into its plural form by exploiting context. The following example illustrates this case.

萨达姆在二十世纪八十年代残酷镇压 (什叶派教徒)。(这) 为拉姆斯菲尔德和美国在伊拉克的历史性失败奠定基础。
Saddam 's brutal repression of the 1980s (Shiites). (It) was Rumsfeld and U.S. failure in Iraq, the historic foundation.

In this example, the pronoun to be resolved is 这 (It), and one of its candidate antecedents is 什叶派教徒 (Shiites). Since 这 (It) is *Singular*, all *Plural* candidate antecedents, including 什叶派教徒 (Shiites), should be removed from consideration. However, in the monolingual baseline, as the 什叶派教徒 (Shiites) is not marked as *Plural* (because number information is absent in the Chinese corpus), the baseline wrongly posits this candidate antecedent as coreferent with 这 (It). In contrast, the English counterpart of this candidate antecedent has a plurality marker (i.e., the suffix 's'), so our approach correctly determines that this candidate antecedent is not coreferent with the pronoun.

Impact of Machine Translation Quality

Recall that two of our classifiers, PR^M and PR^E, rely on automatically translated English text. A natural question, then, is: to what extent does MT quality impact pronoun resolution performance? To answer this question, we compare the results of our approach when it is used in combination with human-translated text and with machine-translated text. Specifically, we repeat the experiments in Table 2 on the 400 document-parallel corpus in OntoNotes 5.0, first using the English documents produced by machine-translating the 400 Chinese documents, and then using the human-translated documents taken from the document-parallel corpus.[11]

Five-fold cross-validation results of different resolution methods using MT and human translation (HT) are shown in Table 3. The systems in the first three rows of the table do not rely on translated text, so their results under both the MT and HT columns are identical. The Rahman and Ng results shown in row 4 are somewhat counter-intuitive: F-score drops by 6% when MT is replaced with HT. An examination of the output reveals that the errors in sentence and mention alignment have contributed to the poorer HT results. While we also need to perform mention alignment in the case of MT, the mention alignment rate on the machine-translated text is higher than that on the human-translated text (94.5% vs. 90.3%). The reason why it is easier to align mentions in machine-translated text is simple: the structures of the sentences in a machine-translated sentence pair are more similar to each other than those in a human-translated pair.

Rows 5−8 of Table 3 show the results of the four resolution methods employed by our approach. Note that any differences between the MT and HT results should be attributed to not only differences in translation quality, but also differences in sentence and mention alignment rate. As expected, when MT is replaced with HT, the F-scores of all four methods increase significantly by 0.9−1.5%.

Note that the relative performance of the resolution methods remains the same when MT is replaced with HT. This observation also holds true when we compare the results in Table 2 and Table 3.

[11] Note that the document-parallel corpus is not sentence-parallel. To align Chinese and English sentences, we employ the Champollion Tool Kit (http://champollion.sourceforge.net/) and heuristically adjust its output to improve the recall rate of the alignment.

Related Work

Outside the ACE evaluations and the CoNLL-2012 shared task, there has been little work on Chinese anaphora and coreference resolution, and all of these works used monolingual approaches. Luo and Zitouni (2005) employ Chinese specific syntactic features for coreference resolution. Wang and Ngai (2006) apply clustering to Chinese coreference resolution, employing features commonly-used for English coreference resolution. Wei et al. (2008) employ syntactic features and word senses to resolve third-person Chinese pronouns. Kong and Zhou (2012) employ tree kernels to resolve Chinese pronouns. Kong and Ng (2013) exploit zero pronouns to improve Chinese coreference resolution.

To our knowledge, our bilingual approach is the only approach to pronoun resolution that exploits coreference-annotated data in two languages (in our case English and Chinese) to help improve the resolution of pronouns in the poorer-resourced language of the two (in our case Chinese). The approaches that are most closely related to ours are those that attempt to resolve pronouns in a resource-poor language via annotation projection. Specifically, these approaches operate under the setting where coreference-annotated documents are only available in a resource-rich language (e.g, English) for training a resolver, but no coreference-annotated data is available in the target language. As noted before, the idea is to resolve the mentions in the target language via an annotation projection approach, where the documents in the target language can be translated automatically (using machine translation, see Rahman and Ng (2012)) or manually (possibly via a parallel corpus, see Harabagiu and Maiorano (2000) and de Souza and Orasan (2011)).

Conclusions

We investigated a novel bilingual approach to Chinese pronoun resolution that exploits English resources. Results on OntoNotes 5.0 show that our approach significantly outper-

forms its monolingual counterparts with respect to both F-score and accuracy. To our knowledge, our results for this task are the best results reported to date on this dataset.

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.

Björkelund, A., and Farkas, R. 2012. Data-driven multilingual coreference resolution using resolver stacking. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Shared Task*, 49--55.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3):27.

Chen, C., and Ng, V. 2012. Combining the best of two worlds: A hybrid approach to multilingual coreference resolution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Shared Task*, 56--63.

de Souza, J. C., and Orasan, C. 2011. Can projected chains in parallel corpora help coreference resolution? In *Anaphora Processing and Applications*. Springer. 59--69.

Harabagiu, S., and Maiorano, S. 2000. Multilingual coreference resolution. In *Proceedings of the Sixth Applied Natural Language Processing Conference*, 142--149.

Kong, F., and Ng, H. T. 2013. Exploiting zero pronouns to improve Chinese coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 278--288.

Kong, F., and Zhou, G. 2012. Pronoun resolution in English and Chinese language based on tree kernel. *Journal of Software* 23(5):1085--1099.

Luo, X., and Zitouni, I. 2005. Multi-lingual coreference resolution with syntactic features. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, 660--667.

Pradhan, S.; Moschitti, A.; Xue, N.; Uryupina, O.; and Zhang, Y. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Shared Task*, 1--40.

Rahman, A., and Ng, V. 2012. Translation-based projection for multilingual coreference resolution. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 720--730.

Recasens, M.; Màrquez, L.; Sapena, E.; Martí, M. A.; Taulé, M.; Hoste, V.; Poesio, M.; and Versley, Y. 2010. SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 1--8.

Song, W.; Qin, B.; Lang, J.; and Liu, T. 2008. Combining syntax and word sense for Chinese pronoun resolution. *Journal of Chinese Information Processing* 22(6):8--13.

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

Wang, C.-S., and Ngai, G. 2006. A clustering approach for unsupervised Chinese coreference resolution. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 40--47.