Singular or Plural? Exploiting Parallel Corpora for Chinese Number Prediction

Elizabeth Baran
Computer Science Department
Brandeis University
Waltham, MA 02453, USA
ebaran@brandeis.edu

Nianwen Xue
Computer Science Department
Brandeis University
Waltham, MA 02453, USA
xuen@brandeis.edu

Abstract

We explore a novel approach to automatically predict noun number in Chinese by using a word-aligned Chinese-English parallel corpus. We first map number information from English onto Chinese to create a dataset labeled with a POS tagset enhanced with number information, and then train a model to automatically predict noun number using a combination of lexical and syntactic features. We evaluate the quality of the automatically mapped data and show the mapping is largely adequate despite a small percentage of errors. Trained on a relatively small data set, our model achieves a 4% improvement in absolute accuracy over a majority baseline that considers all nouns to be singular.

1 Introduction

Chinese is poor in inflectional morphology, and noun number morphology is no exception to this generalization. While in English, the overwhelming majority of nouns have an inflectional suffix to indicate that a noun is plural, in Chinese, there is no such obligatory explicit indication on the noun itself, and we must rely on surrounding context to infer the number of the noun instead. For example, the word 人 can mean either "person" - singular or "people" - plural. But in the following example it is clear that 人 should be plural because of the number three that precedes it:

(1) 三 个 人
three M person/people
"three people"

Unlike the Penn English Treebank (Santorini, 1990), large-scale popular Chinese corpora such as the Penn Chinese Treebank (Xue et al., 2005) generally do not account for noun number in their prescribed part-of-speech tagset (Xia, 2000). One can imagine a reason for this is that while number information is readily available through English morphology making it a straightforward addition to the tagset, it is not in Chinese. Also, while verbal morphology is often dependent on the number of its noun subject in English, Chinese has little verbal morphology and does not have to worry about number agreement. Whatever the reason may be, past annotation frameworks have tended to ignore noun number in Chinese entirely.

Predicting noun number in Chinese, however, is interesting from both theoretical and practical viewpoints. From a theoretical point of view, it is interesting to model how native speakers of Chinese infer number without the help of such easily observable cues as inflectional morphology. It also has practical value for a number of natural language tasks. One such task is *anaphora resolution*. Unlike Chinese nouns, Chinese pronouns *do* have a number morpheme that is obligatory - the suffix 们 , is added to any one of the singular pronouns 我(I), 你(you), 他(he), 她(she), 它(it), to make them plural.[1] What follows is an interesting phenomenon where pronouns, which are explicitly plural or singular, are linked to antecedents that are grammati-

[1]It should be noted that 们 can sometimes be attached to people nouns as well, e.g. 朋友们 (friends), showing that nouns are not *entirely* void of number morphology. In any case, this only further shows the significance of noun number in Chinese.

cally number neutral. Determining the number of the antecedent would help resolve a pronoun to its antecedent, by virtue of the fact that the pronoun and its antecedent generally agree in number. Correctly predicting number also helps the more general *coreference resolution* task, where nouns and pronouns are partitioned into their equivalent classes. Number agreement is important in determining coreferentiality of two noun phrases, and predicting number is the prerequisite for determining number agreement in the absence of explicit morphological cues.

Predicting noun number also helps build better Machine Translation models that translate Chinese into a morphologically rich language. While producing the correct word order in the target language has been the focus of most MT research to date, rendering words in their correct morphological forms will have to be factored into the MT models in order to produce fluent translations. While it is easy to see that knowing the number information of a Chinese noun would help the MT system output a word with the correct number inflection in the target language, predicting noun number is also useful in less obvious ways. Chinese is a pro-drop language that allows pronouns to be dropped. When translated into a non-pro-drop language such as English, these pronouns will have to be recovered and this has been the topic of a few recent research efforts in recovering dropped pronouns (Yang and Xue, 2010; Chung and Gildea, 2010; Kong and Zhou, 2010; Cai et al., 2011). A part of recovering these dropped pronouns, for Chinese, would include determining whether the pronoun should be singular or plural. To determine this, it will be necessary to know the number of its antecedent, which is where automatic prediction of noun number in Chinese may prove to be very useful.

The task of deciding whether a noun is singular or plural is a mindless process for any fluent speaker of Chinese, but is not nearly as straightforward to determine algorithmically. This is where having number information readily available in Chinese becomes much less trivial. In this paper, we attempt to take on this task of predicting noun number in Chinese by exploiting a manually word-aligned parallel Chinese-English corpus. The number information is first mapped from English onto Chinese and is added to the Chinese Treebank tagset. We then frame the Chinese number prediction problem as a part-of-speech tagging task with an enhanced tagset that includes number. We trained a Maximum Entropy classifier on all word tokens and their parts-of-speech, and achieve significant gains over the majority class baseline that treats all nouns as singular.

The rest of the paper is organized as follows. In section 2 we explain the process of collecting data and mapping English number onto Chinese, as well as some of the issues encountered by going about it this way. In section 3 we describe the training process and outline the features we used in our algorithm. In section 4 we report our results. We discuss related work in 5, and in section 6 we conclude and discuss future work.

2 Data Preparation

In order to train a supervised machine learning model to predict number, we needed to establish some sort of gold standard training and testing data. As mentioned above, noun number is not specified in the Chinese Treebank (CTB), but a portion of the CTB has English translations that are POS-tagged and parsed according to the Penn English Treebank annotation specifications (Santorini, 1990; Bies et al., 1995). The Chinese side and the English side of the parallel corpus are also manually word-aligned. As alluded to above, the Penn English Treebank POS tag set encodes the number information. For example, a common noun receives the NN tag if it is singular and the NNS tag if it is plural. Similarly, a proper noun receives the NNP tag if it is singular and the NNPS tag if it is plural. We decided to exploit this fact, and use the number information from the parallel English data to determine noun number in the Chinese data. Specifically, we used data files from the newswire section of the Chinese Treebank (Xue et al., 2005) that had corresponding English translations, parses, and alignments for our training, development, and testing sets. See Table 1 for the data split.

Basically, if a plural noun in English was aligned to a noun in Chinese, we added an 'S' to the current noun tag, to make it plural. If more than one noun in English was aligned to a single Chinese noun, and any of those aligned English nouns were plural, then the Chinese noun would also be tagged as plural. At

Data	Train	Test
CTB	8, 11-14, 17-20, 23-24, 26, 28, 30-33, 35-37, 43-44, 46-49, 51, 53-64, 66, 68, 71, 73-74, 76, 79, 81-84, 86-87, 89, 91, 93-95, 97-98, 101-104, 107-109, 111, 113, 115-116, 123, 126, 130-132, 134-138, 142-143, 146-150, 153-156, 159-169, 208-215, 217-218, 221-223, 229-230, 232-234, 236-242, 245-246, 249-251, 255-256, 258-259, 261, 263, 265, 267, 268-269	301, 304, 306, 311-314, 316-318, 320, 323

Table 1: The training and testing data set divisions.

the end of the process, we doubled the size of the current CTB noun tagset so that {NN, NT, NR} became {NN, NNS, NT, NTS, NR, NRS}.

2.1 Issues with Mapping

Ideally, a gold standard corpus would be created through manual annotation, but for this first proof-of-concept attempt, mapping English number onto aligned Chinese nouns served as a reasonable substitute. Nevertheless, the reader should be conscious of the types of errors that arose by going about it this way.

To make sure the error rate was low enough to proceed, we looked at 10 files that had been mapped in the way described above. We scanned through all nouns, and determined whether or not the number mappings from English were correct. If we determined a tag to be incorrect, we looked back at the Chinese and English aligned sentences to see why this occurred. Errors generally fell into 5 different categories:

Incorrect Translation - An incorrect or inaccurate translation is a case in which the English noun that was aligned to the Chinese noun was not the best translation or was just completely wrong, and led to an incorrect number mapping. One can imagine that for this type of task, the more literal the English translation, the more faithful the number alignment for its corresponding Chinese noun. However, as we all know, literal translations are not always the best in terms of overall aesthetic quality and overall conveyance of the source sentence. With that said, inaccurate or incorrect translations, here, should be understood to mean inaccurate or incorrect within the framework of this number mapping task. Put simply, a translation deemed incorrect here may be acceptable for other purposes. Example 2 shows an incorrect translation. Note that the gloss comes straight from the English aligned sentence and the English words are placed under the Chinese words to which they are aligned as specified in the alignment file.

(2) 从 内资 和 外资
from of national capital and foreign capital
两 个 **方面**
both NA the aspect
"from the aspect of both national capital and foreign capital"

The bolded word in 2 is the noun in question. The word it is aligned with in English, "the aspect", is singular and therefore 方面 is tagged as singular. But really, a more accurate translation would have been something along the lines of "from the two aspects of national capital and foreign capital", where "aspects" is plural and aligns with 方面. 两 is aligned with "both" but it literally means "two" and is directly modifying 方面 (aspect) so 方面 is undeniably plural and is wrongfully tagged due to this inaccurate translation.

Incorrect Word Alignment - An incorrect word alignment is when the Chinese noun in question is aligned to the wrong English noun resulting in an incorrect number mapping. In other words, the correctly translated word does exist, but it was either incorrectly aligned or not aligned at all. Example 3 illustrates this type of error.

(3) 业内 **人士** 认为
insiders NA feel
"insiders feel"

As we can see, the bolded word in 3 was not aligned to any word in English, which is incorrect. Although 业内 and 人士 are both considered nouns, 人士 is the head noun meaning something close to the words "people" or "public figures" in English, and 业内 is a modifying noun that means "inside of the industry". When put together, this could reasonably translate to "insiders" in English, but even so, it makes more sense to align "insiders"

with the head noun meaning “people”. This incorrect word alignment resulted in two number errors, where 业内 was incorrectly tagged as plural and 人士 was incorrectly tagged as singular.

Incorrect Part-of-Speech Alignment - An incorrect part-of-speech alignment means that the Chinese noun in question was aligned with a word that was semantically similar but belonged to a different word class in English. These types of pairings, are considered correct according to the alignment guidelines (Li et al., 2009), and arguably so since English tends to more easily derive adjectives from nouns. Because of this it is not rare to see a modifying noun in Chinese aligned with an adjective in English as in Example 4 .

(4) 其中 五十七 项 被
among them 57 items were approved
批准 列入 国家 、 省
NA to be listed in national , provincial
、 市 火炬 计划
and municipal Torch the plan
“among them , 57 items were approved to be listed in the national , provincial and municipal Torch Plan”

Ignoring some other odd mappings, e.g. “the plan” was pulled from another part of the sentence, not shown here, we can see that 省 , which means “provinces” and 市 which means “cities”, are paired with the adjectives “provincial” and “municipal” in English. Because they were not associated with any nouns in English, they were left as singular, when it is clear that they should be plural in Chinese.

Incorrect English Part-of-Speech Tag - An incorrect English part-of-speech tag is simply an error in the English annotation that resulted in an erroneous number mapping. For example, on one occasion “statistics” was tagged as a singular proper noun in English, when it should have been a plural common noun, and therefore the corresponding Chinese noun, 统计, was also incorrectly labeled as singular.

Other - This category encapsulates all other issues and errors that do not fall under the categories listed above. Most notably, this includes words for which it was difficult to determine number like 外资 (foreign investment/s) and 成本 (cost/s). To decide whether these types of nouns are singular or plural is difficult because they are mass nouns, meaning they represent a plurality of things but are grammatically singular. It would seem that this would answer our question and we should call them singular, but as is well-documented in the literature any mass noun can shift to be countable given the right context; the atomic unit just becomes that original mass. With sums of money one could argue that “investment” becomes “investments” when the money comes from different $sources$, and “cost” becomes “costs” when there are different $types$ of costs. Even given these definitions, it is not entirely clear in many contexts which explanation is in effect in English, a language with a rich number morphology, let alone Chinese. Another similar issue occurred with words that had more than one translation in English, both of which were acceptable, but were different in number. For example, the word 联合 was translated as “integration” (singular) in some contexts, and “ties” (plural) in others. We decided that in Chinese 联合 was singular and a translation like “ties” in English was a mere idiosyncrasy. 境内 was another challenging word that means “within the border/s”. Should it be “within the border” or “within the borders”? The latter translation sounds slightly more acceptable in English but it is not as clear in Chinese.

Despite these types of errors, out of the 1202 nouns we looked at, only about 49 of those had incorrect noun tags (about 4.1%) with respect to number. The relatively low error rate confirmed our confidence in using aligned English number information to establish a Chinese noun number gold standard data set.

3 Features

We used the Stanford Maximum Entropy classifier (Manning and Klein, 2003) to train a model on all word tokens and their part-of-speech tags. Conceptually, we sought to create an enhanced part-of-speech tagger that had a new layer for noun number with a new set of noun tags to match. This can be viewed as a two-pass process to assign to each word a POS tag that encodes

number information. In the first pass, the sentence is automatically parsed with a syntactic structure based on the original POS tag set. The second pass is to add number information to the POS tag of each word based on features extracted from the word tokens and from the syntactic parse of the sentence (obtained automatically), which incorporates the original numberless POS tags. It is conceivable that this new enhanced tagger can be trained in one step. However, in that scenario, we would not be able to use the syntactic information as features. Although we are training and testing for all word classes in Chinese, the results will be most interesting for number, since much of the original part-of-speech tagging will be done indirectly through features from our automatic parse file. Note that the English side of the parallel data is only used for mapping the number information onto Chinese to create a gold standard "numbered" corpus and is not used for feature extraction. The full list of features used in our model is described below:

Lexical Features:

1. **word** - the current word token (w_0)
2. **left_word** - the word to the left of the current word (w_{-1})
3. **right_word** - the word to the right of the current word (w_{+1})
4. **common_number** - 'p' if the current noun is most often plural (>= 55% plural) and 's' otherwise.

The plural/singular frequencies were obtained in the same manner as our gold standard data set by using English number information. We used about 100,000 sentences from a large Chinese-English parallel corpus that was automatically word-aligned and parsed to calculate number frequency for each word. This parallel corpus is a resource that is separate from the data files we used for our experiment. The 55% was the optimal ratio obtained by manually tweaking on the development data set.

Syntactic Features:

To obtain the following syntactic features, we trained the Berkeley Parser (Petrov and Klein, 2006) on the set of CTB files that were complement to the ones we had in our data set. Then we used that model to automatically parse our data files and extract the following features:

5. **word_pos** - the current word's part-of-speech tag (p_0)
6. **left_pos** - the part-of-speech tag of the word to the left (p_{-1})
7. **right_pos** - the part-of-speech tag of the word to the right (p_{+1})

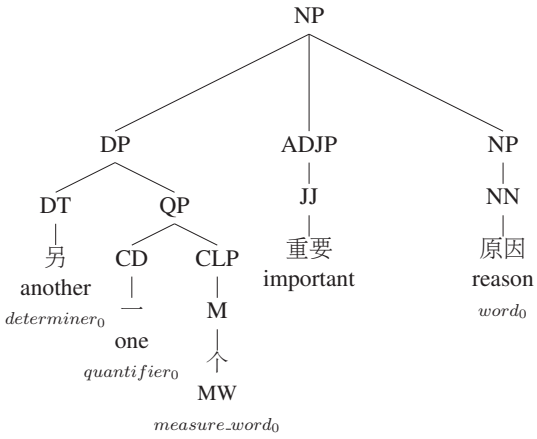

Figure 1: An example tree structure displaying the current $word$ and the following associated features: $determiner$, $quantifier$, and $measure\_word$.

The following set of syntactic features are relevant only to **head nouns**. We define a head noun as any noun that has no right siblings or whose immediate right sibling is a conjunction (CC) or some form of punctuation (PU).

8. **quantifier** - the quantifier that precedes the head noun, if one exists. The quantifier is the child of CD or an OD embedded within a QP and/or a DP that is a left sibling of the current noun's parent NP. See Figure 1 for an example.
9. **measure_word** - the measure word that precedes the head noun if one exists. The measure word is the child of an M node that is embedded in a QP and/or DP that is a left sibling of the current noun's parent NP. See Figure 1 for an example.
10. **determiner** - the determiner that precedes the current head noun, if one exists. The determiner is the child of a DT that is embedded in

a DP that is a left sibling to the current noun's parent NP. See Figure 1.

The $quantifier$, $measure\_word$, and $determiner$ features reveal information about the noun they precede in both obvious and subtle ways. The $quantifier$ feature is the most direct representation of number, when it exists, for obvious reasons in that it represents the actual number amount of the noun. However, the way these features interplay reveals much about noun number too. For example, in the absence of a quantifier, the determiners 这 (this/these) and 那 (that/those) followed by a measure word, almost always signal a singular noun, e.g. 这个人 (this person).[2]

11. **adverb_following** - if the noun is the subject of a verb, the adverb that precedes that verb, if one exists. This is to capture adverbs that tend to follow plural nouns, e.g. 分别 (respectively), 均 (all).
12. **np_is_np_number** - if the noun is the subject of a VC (e.g. 是 , 为 be), and the object is an NP, the $common\_number$ of the head noun of that NP. If the noun is the object of a VC, and the subject is an NP then the $common\_number$ of the head noun of that NP, if one exists. If the head noun for any of these cases is a pronoun, 's' for a singular pronoun and 'p' for a plural pronoun. Or if the head noun for any of these cases is part of a list of items, then 'p'.
13. **np_is_np_pos** - the part-of-speech of the word token obtained through $np\_is\_np\_number$. See Figure 2.
14. **appositive_number** - if the current noun is in an appositive construction, i.e. (NP (NP ...) (PU ,) (NP ...)), the $common\_number$ of the head noun of the corresponding NP. See Figure 3.
15. **appositive_pos** - the part-of-speech parent node of the word token obtained through $appositive\_number$. See Figure 3.

The np_is_np and $appositive$ features seek to exploit number information from nouns that are in parallel structures. For example, if something is something else, then those two things should be equal in number, at least the majority of the time.[3]

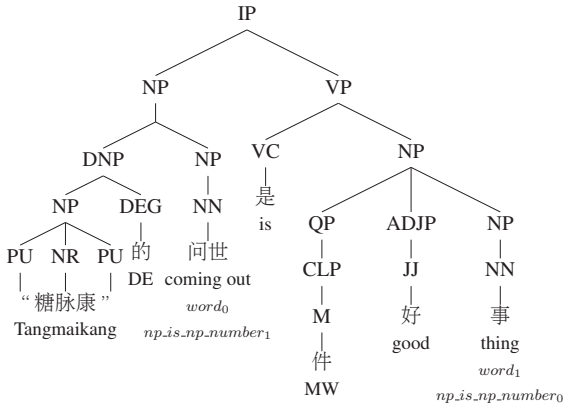

Figure 2: An example tree structure displaying two $word$ features and their associated $np\_is\_np\_number$ features.

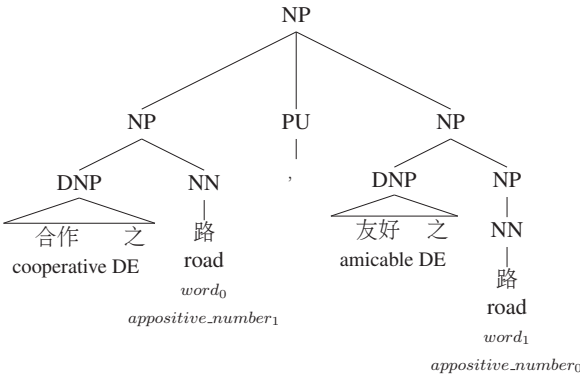

Figure 3: An example tree structure displaying two $word$ features and their associated $appositive\_number$ features.

4 Results

In order to establish a baseline measure, we calculated precision and recall for all nouns by comparing the part-of-speech tags obtained by automatically parsing the test data (these do *not* contain plural noun tags) against our gold standard corpus - a manually tagged corpus with the addition of plural noun tags obtained through the mapping process described in Section 2. You will see that there are 4

[2] An exception to this rule is when the determiner occurs with a *group* measure word, like 群 (group). If we substitute 群 (group) into the phrase above we get 这群人 (this group of people) where 人 (people) is plural.

[3] An example of an exception to this, in English, is the expression "we are the world" where "we" is plural and "world" is singular.

types of precision and recall scores that will carry through the rest of this paper .They will be referred to as **noun**, **number**, **plural**, and **singular** precision/recall/F1 scores henceforth.

Noun scores refer to complete noun tag matches (total string matches), whereas **number** scores are only concerned with whether or not the noun tag ends in 'S' (e.g. NRS == NNS and NR == NN). **Plural** scores refer to plural noun instances, ignoring noun types (i.e. NRS == NTS == NNS). Similarly, **singular** scores refer to singular noun instances, ignoring noun types (i.e. NR == NT == NN).

The results, using all of the features, are displayed in Table 2 .

	Baseline			Learning		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Noun	75.4	75.7	75.5	79.4	79.5	79.4
Number	80	80.1	80.0	84.1	84.2	84.2
Plural	0	0	0	68.8	39.3	50.0
Singular	80.0	97.4	87.8	85.8	93.9	89.7

Table 2: Test results from the Maximum Entropy classifier compared with the baseline measure obtained through the automatic parse file.

From Table 2 we see that our algorithm achieved 79.4% noun precision and 79.5% noun recall, as well as 84.1% number precision and 84.2% number recall showing about a 4% improvement for noun scores, and a little over a 4% improvement for number scores. Our algorithm achieved 68.8% plural precision and 39.3% plural recall, leading to an F1 score of 50%. And finally, singular recall decreased, but singular precision increased leading to an F1 score of 89.7%, an overall gain of close to 2% compared with the baseline.

Table 3 shows how effective each feature was in achieving these results. You will see that, in terms of noun and number scores, no feature was hurtful. A couple features hurt some of the plural and singular precision/recall measures, but still helped the overall F1 scores. For example, $left\_pos$ although helpful overall, actually negatively affected plural precision.

We can see that the $word\_pos$ feature was by far the most effective feature in terms of noun and number scores. This was expected since we are essentially feeding the Berkeley parser's best guess at the part-of-speech tag to the new algorithm. However, it was not the most effective feature in terms of plural scores. We can see that $common\_number$ significantly beats out all other features in this regard, suggesting that certain nouns tend to be either singular or plural and leveraging that information is useful. The $right\_word$ feature placed second in overall effectiveness, and it was also the second most effective feature for plural scores.

Both the $appositive\_number$ and $appositive\_pos$ features had no effect on the development data set just as they didn't on the test data set, but we assume this is due to a lack of instances of this feature as opposed to a lack of effectiveness. Conceptually, it is very similar to the $np\_is\_np\_pos$ and $np\_is\_np\_number$ features and the addition of both those features showed modest improvements.

Features	$Noun$	$Number$	$Plural$	$Singular$
all	79.4	84.2	50.0	89.7
-word_pos	-12	-5.9	-2.3	-6.3
-right_word	-1.3	-1.2	-5.5	-0.7
-common_number	-1.1	-1.3	-10.6	-0.6
-measure_word	-0.7	-0.8	-3.4	-0.4
-word	-0.6	-1.1	-5.4	-0.6
-right_pos	-0.5	-0.6	-3.3	-0.3
-quantifier	-0.4	-0.4	-0.7	-0.2
-determiner	-0.4	-0.5	-2.5	-0.2
-adverb_following	-0.3	-0.4	-1.7	-0.2
-left_word	-0.2	-0.2	-1.5	-0.1
-np_is_np_number	-0.2	-0.3	-1.2	-0.1
-left_pos	-0.1	-0.2	-2.1	-0.1
-np_is_np_pos	-0.1	-0.2	-1.0	-0.1
-appositive_number	-0.0	-0.0	-0.0	-0.0
-appositive_pos	-0.0	-0.0	-0.0	-0.0

Table 3: The F1 scores associated with the removal of each individual feature. Numbers are displayed as percent change relative to the all row.

5 Related Work

To our knowledge, no work has dealt specifically with predicting noun number in Chinese by using number information from an English parallel corpus.

Our training process is somewhat similar to that of Baldwin (2003) in predicting the countability of English nouns which is related to noun number. They too used a suite of lexical and syntactic features from English parses and train it on a classifier to predict the countability of nouns in English. Number information, which is available in English, was used extensively in their feature set to predict noun count-

ability. This is related to our motivation to extract features that were related to noun countability in Chinese, e.g. the $measure\_word$ feature, to predict noun number.

6 Conclusions and Future Work

The goal of this paper is to bring to the forefront a characteristic of Chinese that has been largely brushed over in the past, at least in the field of NLP - noun number. Nouns in Chinese, for the most part, are not morphologically marked for number as they are in English, so number is inferred through context instead. We have shown that by essentially expanding the noun tagset of the CTB to include a corresponding plural tag for each singular tag type, and then retraining a part-of-speech tagger with this new tagset and features from an automatic parse, it is feasible to make predictions for noun number in Chinese.

Using aligned English files to create a gold standard number corpus for Chinese worked as a proof-of-concept, but the errors we encountered were still unsettling. There are two ways we would like to mitigate these errors in the future. One would be to train our algorithm on a much larger data set, which would hopefully help weed out some of this noise and also make some of the features, like the $appositive$ features, more relevant. Also, using this method brought to light many inconsistencies and questionable mappings that currently exist in the translation and alignment files. When using alignment information on a single token level, those mappings need to be precise and it seemed clear at times that the mappings we had were not as tight as they could have been. Ideally, we hope that our model can be tested and improved on a larger data set with more reliable mappings, or of course, on a manually annotated corpus with tagsets that account for number in Chinese. It would also be interesting to delve deeper into the relationship between quantifiers, determiners, and measure words with respect to noun number and supply more discrete and informed versions of the related features from these experiments.

Acknowledgements

This work is supported by the National Science Foundation via Grant No. 0910532 entitled "Richer Representations for Machine Translation". All views expressed in this paper are those of the authors and do not necessarily represent the view of the National Science Foundation.

References

Ann Bies, Mark Ferguson, Karen Katz, Robert Macintyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for treebank ii style penn treebank project.

Shu Cai, David Chiang, and Yoav Goldberg. 2011. Language-independent parsing with empty elements. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 212–216, Portland, Oregon, USA, June. Association for Computational Linguistics.

Tagyoung Chung and Daniel Gildea. 2010. Effects of empty categories on machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 636–645, Cambridge, MA.

Timothy Baldwin Csli and Timothy Baldwin. 2003. Learning the countability of english nouns from corpus data. In *in Proc. of the 41st Annual Meeting of the ACL*, pages 463–470.

Fang Kong and Guodong Zhou. 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*, MIT, Massachusetts.

Xuansong Li, Niyu Ge, and Stephanie Strassel. 2009. Guidelines for chinese-english word alignment.

Christopher Manning and Dan Klein. 2003. Optimization, maxent models, and conditional estimation without magic.

Slav Petrov and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *COLING-ACL*.

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the penn treebank project.

Fei Xia. 2000. The part-of-speech tagging guidelines for the penn chinese treebank (3.0).

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*, pages 207–238.

Yaqin Yang and Nianwen Xue. 2010. Chasing the Ghost: Recovering Empty Categories in the Chinese Treebank. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, Beijing, China.