Shallow Semantic Parsing of Chinese

Honglin Sun[1]
Center for Spoken Language Research
University of Colorado at Boulder

Daniel Jurafsky[2]
Center for Spoken Language Research
University of Colorado at Boulder

Abstract

In this paper we address the question of assigning semantic roles to sentences in Chinese. We show that good semantic parsing results for Chinese can be achieved with a small 1100-sentence training set. In order to extract features from Chinese, we describe porting the Collins parser to Chinese, resulting in the best performance currently reported on Chinese syntactic parsing; we include our head-rules in the appendix. Finally, we compare English and Chinese semantic-parsing performance. While slight differences in argument labeling make a perfect comparison impossible, our results nonetheless suggest significantly better performance for Chinese. We show that much of this difference is due to grammatical differences between English and Chinese, such as the prevalence of passive in English, and the strict word order constraints on adjuncts in Chinese.

1 Introduction

Thematic roles (AGENT, THEME, LOCATION, etc) provide a natural level of shallow semantic representation for a sentence. A number of algorithms have been proposed for automatically assigning such shallow semantic structure to English sentences. But little is understood about how these algorithms may perform in other languages, and in general the role of language-specific idiosyncracies in the extraction of semantic content and how to train these algorithms when large hand-labeled training sets are not available. In this paper we address the question of assigning semantic roles to sentences in Chinese. Our work is based on the SVM-based algorithm proposed for English by Pradhan et al (2003). We first describe our creation of a small 1100-sentence Chinese corpus labeled according to principles from the English and (in-progress) Chinese PropBanks. We then introduce the features used by our SVM classifier, and show their performance on semantic parsing for both seen and unseen verbs, given hand-corrected (Chinese TreeBank) syntactic parses. We then describe our port of the Collins (1999) parser to Chinese. Finally, we apply our SVM semantic parser to a matching English corpus, and discuss the differences between English and Chinese that lead to significantly better performance on Chinese.

2 Semantic Annotation and the Corpus

Work on semantic parsing in English has generally related on the PropBank, a portion of the Penn TreeBank in which the arguments of each verb are annotated with semantic roles. Although a project to produce a Chinese PropBank is underway (Xue and Palmer 2003), this data is not expected to be available for another year. For these experiments, we therefore hand-labeled a small corpus following the Penn Chinese Propbank labeling guidelines (Xue, 2002). In this section, we first describe the semantic roles we used in the annotation and then introduce the data for our experiments.

2.1 Semantic roles

Semantic roles in the English (Kingsbury et al 2002) and Chinese (Xue 2002) PropBanks are grouped into two major types:

(1) arguments, which represent central participants in an event. A verb may require one, two or more arguments and they are represented with a contiguous sequence of numbers prefixed by arg, as arg0, arg1.

(2) adjuncts, which are optional for an event but supply more information about an event, such as time, location,

[1] Currently at Department of Computer Science, Queens College, City University of New York. Email: sunh@qc.edu.

[2] Currently at Department of Linguistics, Stanford University. Email: jurafsky@stanford.edu.

reason, condition, etc. An adjunct role is represented with argM plus a tag. For example, argM-TMP stands for temporal, argM-LOC for location.

In our corpus three argument roles and 15 adjunct roles appear. The whole set of roles is given at Table 1.

Table1 The list of semantic roles

Role	Freq train	Freq Test	Note
arg0	556	63	
arg1	872	91	
arg2	23	5	
argM-ADV	191	32	adverbial
argM-BFY	26	2	beneficiary(e.g. give support [to the plan])
argM-CMP	35	3	object to be compared
argM-CND	14	1	condition
argM-CPN	7	3	companion (e.g. talk [with you])
argM-DGR	53	4	degree
argM-FRQ	3	0	frequency
argM-LOC	207	31	location
argM-MNR	10	1	manner
argM-PRP	11	0	purpose or reason
argM-RNG	7	2	range(e.g. help you [in this aspect])
argM-RST	15	1	result(e.g. increase [to $100])
argM-SRC	11	1	source(e.g. increase [from $50] to $100)
argM-TMP	376	45	temporal
argM-TPC	12	2	topic

2.2 The training and test sets

We created our training and test corpora by choosing 10 Chinese verbs, and then selecting all sentences containing these 10 verbs from the 250K-word Penn Chinese Treebank 2.0. We chose the 10 verbs by considering frequency, syntactic diversity, and word sense. We chose words that were frequent enough to provide sufficient training data. The frequencies of the 10 verbs range from 41 to 230, with an average of 114. We chose verbs that were representative of the variety of verbal syntactic behavior in Chinese, including verbs with one, two, and three arguments, and verbs with various patterns of argument linking. Finally, we chose verbs that varied in their number of word senses.

In total, we selected 1138 sentences. The first author then labeled each verbal argument/adjunct in each sentence with a role label. We created our training and test sets by splitting the data for each verb into two parts: 90% for training and 10% for test. Thus there are 1025 sentences in the training set and 113 sentences in the test set, and each test set verb has been seen in the training set. The list of verbs chosen and their number of senses, argument numbers and frequencies are given in Table 2.

Table 2 List of verbs for experiments

Verb	# of senses	Arg number	Freq
成立/set up	1	2	106
出现/emerge	1	1	80
发表/publish	1	2	113
给予/give	2	3/2	41
建成/build into	2	2/3	113
进入/enter	1	2	123
举行/take place	1	2	230
通过/pass	3	2	75
希望/hope	1	2	90
增加/increase	1	2	167

3 Semantic Parsing

3.1 Architecture and Classifier

Following the architecture of earlier semantic parsers like Gildea and Jurafsky (2002), we treat the semantic parsing task as a 1-of-N classification problem. For each (non-aux/non-copula) verb in each sentence, our classifier examines each node in the syntactic parse tree for the sentence and assigns it a semantic role label. Most constituents are not arguments of the verb, and so the most common label is *NULL*. Our architecture is based on a Support Vector Machine classifier, following Pradhan et al. (2003). Since SVMs are binary classifiers, we represent this 1-of-19 classification problem (18 roles plus *NULL*) by training 19 binary one-versus-all classifiers.

Following Pradhan et al. (2003), we used tinySVM along with YamCha (Kudo and Matsumoto 2000, 2001) as the SVM training and test software. The system uses a polynominal kernel with degree 2; the cost per unit violation of the margin, C=1; tolerance of the termination criterion e=0.001.

3.2 Features

The literature on semantic parsing in English relies on a number of features extracted from the input sentence and its parse. These include the constituent's syntactic phrase type, head word, and governing category, the syntactic path in the parse tree connecting it to the verb, whether the constitutent is before or after the verb, the subcategorization bias of the verb, and the voice (active/passive) of the verb. We investigated each of these features in Chinese; some acted quite similarly to English, while others showed interesting differences.

Features that acted similarly to English include the *target* verb, the *phrase type*, the syntactic category of the constituent. (NP, PP, etc), and the subcategorization of the target verb. The sub-categorization feature represents the phrase structure rule for the verb phrase containing the target verb (e.g., VP -> VB NP, etc). Five features (path, position, governing category, headword, and voice) showed interesting patterns that are discussed below.

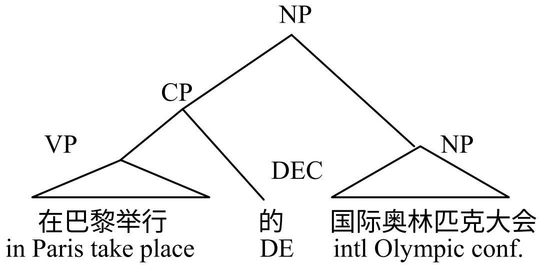
3.2.1 Path in the syntactic parse tree. The path feature represents the path from a constituent to the target verb in the syntactic parse tree, using "^" for ascending a parse tree, and "!" for descending. This feature manifests the syntactic relationship between the constituent and the target verb. For example the path “NP^IP!VP!VP!VV” indicates that the constituent is an “NP” which is the subject of the predicate verb. In general, we found the path feature to be sparse. In our test set, 60% of path types and 39% of path tokens are unseen in the training. The distributions of paths are very uneven. In the whole corpus, paths for roles have an average frequency of 14.5 while paths for non-roles have an average of 2.7. Within the role paths, a small number of paths account for majority of the total occurrences; among the 188 role path types, the top 20 paths account for 86% of the tokens. Thus, although the path feature is sparse, its sparsity may not be a major problem in role recognition. Of the 291 role tokens in our test set, only 9 have unseen paths, i.e., most of the unseen paths are due to non-roles.

Table 3 The positional distribution of roles

Role	**Before verb**	**After verb**	**Total**
arg0	547	72	619
arg1	319	644	963
arg2		28	28
argM-ADV	223		223
argM-BFY	28		28
argM-CMP	38		38
argM-CND	15		15
argM-CPN	10		10
argM-DGR		57	57
argM-FRQ		3	3
argM-LOC	233	5	238
argM-MNR	11		11
argM-PRP	11		11
argM-RNG	9		9
argM-RST		16	16
argM-SRC	12		12
argM-TMP	408	13	421
argM-TPC	14		14
Total	1878	838	2716

3.2.2 Position before or after the verb. The position feature indicates that a constituent is before or after the target verb. In our corpus, 69% of the roles are before the verb while 31% are after the verb. As in English, the position is a useful cue for role identity. For example, 88% of arg0s are before the verb, 67% of arg1s are after the verb and all the arg2s are after the verb. Adjuncts have even a stronger bias. Ten of the adjunct types can **only** occur before the verb, while three are always after the verb. The two most common adjunct roles, argM-LOC and argM-TMP are almost always before the verb, a sharp difference from English. The details are shown seen in Table 3.

3.2.3 Governing Category. The governing category feature is only applicable for NPs. In the original formulation for English in Gildea and Jurafsky (2002), it answers the question: Is the NP governed by IP or VP? An NP governed by an IP is likely to be a subject, while an NP governed by a VP is more likely to be an object. For Chinese, we added a third option in which the governing category of an NP is neither IP nor VP, but an NP. This is caused by the “DE” construction, in which a clause is used as a modifier of an NP. For instance, in the example indicated in Figure 1, for the last NP, “国际奥林匹克大会”(“international Olympic conference”) the parent node is NP, from where it goes down to the target verb “举行”(“taking place”).



“the international Olympic Conference held in Paris”
Figure 1 Example of DE construction

Since the governing category information is encoded in the path feature, it may be redundant; indeed this redundancy might explain why the governing category feature was used in Gildea & Jurafsky(2002) but not in Gildea and Palmer(2002). Since the “DE” construction caused us to modify the feature for Chinese, we conducted several experiments to test whether the governing category feature is useful or whether it is redundant with the path and position features. Using the paradigm to be described in section 3.4, we found a small improvement using governing category, and so we include it in our model.

3.2.4 Head word and its part of speech. The head word is a useful but sparse feature. In our corpus, of the 2716 roles, 1016 head words (type) are used, in which 646 are used only once. The top 20 words are given in Table 4.

Table 4 Top 20 head words for roles

Word	Freq	Word	Freq
在/in	214	中国/China	25
会议/meeting	43	对/for	23
今天/today	41	声明/statement	19
于/at	40	讲话/speech	18
已/already	38	阶段/stage	17
企业/enterprise	35	政府/government	16
公司/company	32	目前/present	16
比/than	31	银行/bank	15
将/will	30	日前/recently	14
仪式/ceremony	28	基地/base	14

In the top 20 words, 4 are prepositions ("在/in，于/at，比/than，对/for") and 3 are temporal nouns("今天/today，目前/present，日前/recently") and 2 are adverbs("已/already, 将/will"). These closed class words are highly correlated with specific semantic roles. For example,"对/for" occurs 195 times as the head of a constituent, of which 172 are non-roles, 19 are argM-BFYs, 3 are arg1s and 1 is an argM-TPC."在/in" occurs 644 times as a head, of which 430 are non-roles, 174 are argM-LOCs, 24 are argM-TMPs, 9 are argM-RNGs, and 7 are argM-CND. "已/already" occurs 135 times as a head, of which 97 are non-roles and 38 are argM-ADVs. "今天/today" occurs 69 times as a head, of which 41 are argM-TMPs and 28 are non-roles.

Within the open class words, some are closely correlated to the target verb. For example, "会议/meeting; conference" occurs 43 times as a head for roles, of which 24 are for the target "举行/take place" and 19 for "通过/pass". "仪式/ceremony" occurs 28 times and all are arguments of "举行"(take place)."声明/statement" occurs 19 times, 18 for "发表/release; publish" and one for "希望/hope".

These statistics emphasize the key role of the lexicalized head word feature in capturing the collocation between verbs and their arguments. Due to the sparsity of the head word feature, we also use the part-of-speech of the head word, following Surdeanu et al (2003). For example, "7月26日/July 26" may not be seen in the training, but its POS, NT(temporal noun) , is a good indicator that it is a temporal.

3.2.5 Voice. The passive construction in English gives information about surface location of arguments. In Chinese the marked passive voice is indicated by the use of the preposition "被/by" (POS tag LB in Penn Chinese Treebank). This passive, however, is seldom used in Chinese text. In our entire 1138-sentence corpus, only 13 occurrences of "LB" occur, and only one (in the training set) is related to the target verb. Thus we do not use the voice feature in our system.

3.3 Experimental Results for Seen Verbs

We now test the performance of our classifier, trained on the 1025-sentence training set and tested on the 113-sentence test set introduced in Section 2.2. Recall that in this 'stratified' test set, each verb has been seen in the training data. The last row in Table 5 shows the current best performance of our system on this test set. The preceding rows show various subsets of the feature set, beginning with the path feature.

Table 5 Semantic parsing results on seen verbs

feature set	P	R	F
	(%)	(%)	(%)
path	71.8	59.4	65.0
path + pt	72.9	62.9	67.5
path + position	72.5	60.8	66.2
path + head POS	77.6	63.3	69.7
path + sub-cat	80.8	63.6	71.2
path + head word	85.0	66.0	74.3
path + target verb	85.8	68.4	76.1
path + pt + gov + position + subcat + target + head word + head POS	91.7	76.0	83.1

As Table 5 shows, the most important feature is path, followed by target verb and head word. In general, the lexicalized features are more important than the other features. The combined feature set outperforms any other feature sets with less features and it has an F-score of 83.1. The performance is better for the arguments (i.e., only ARG0-2), 86.7 for arg0 and 89.4 for arg1.

3.4 Experimental Results for Unseen Verbs

To test the performance of the semantic parser on unseen verbs, we used cross-validation, selecting one verb as test and the other 9 as training, and iterating with each verb as test. All the results are given in Table 6. The results for some verbs are almost equal to the performance on seen verbs. For example for "发表" and "通过", the F-scores are over 80. However, for some verbs, the results are much worse. The worst case is the verb "出现", which has an F-score of 11. This is due to the special syntactic characteristics of this verb. This verb can only have one argument and this argument most often follows the verb, in object position. In the surface structure, there is often an NP before the verb working as its subject, but semantically this subject cannot be analyzed as arg0. For example:

(1)中国/China 不/not 会/will 出现/emerge 粮食/food 危机/crisis. (A food crisis won't emerge in China.)

(2)芬兰/Finland 经济/economy 出现/emerge 了/AUX 战后/post-war 最/most 严重/serious 的/AUX 萧条/depression. (The most severe post-war depression emerged in the Finland economy.)

The subjects, “中国/China” in (1) and “芬兰/Finland 经济/economy”, are locatives, i.e. argM-LOC, and the objects, “粮食/food 危机/crisis” in (1) and “战后/post-war 最/most 严重/serious 的/AUX 萧条/depression” in (2), are analyzed as arg0. But the parser classified the subjects as arg0 and the objects as arg1. These are correct for most common verbs but wrong for this particular verb. It is difficult to know how common this problem would be in a larger, test set. The fact that we considered diversity of syntactic behavior when selecting verbs certainly helps make this test set reflect the difficult cases.

If most verbs prove not to be as idiosyncratic as “出现 /emerge”, the real performance of the parser on unseen verbs may be better than the average given here.

Table 6 Experimental Results for Unseen Verbs

target	P(%)	R(%)	F(%)
发表/publish	90.7	72.9	80.8
增加/increase	49.6	34.3	40.5
举行/take place	90.1	63.3	74.4
建成/build into	65.2	55.5	60.0
给予/give	65.7	37.9	48.1
通过/pass	85.9	77.0	81.2
出现/emerge	12.6	10.2	11.3
进入/enter	81.9	58.8	68.4
成立/set up	79.0	61.1	68.9
希望/hope	77.7	35.9	49.1
Average	69.8	50.7	58.3

Another important difficulty in processing unseen verbs is the fact that roles in PropBank are defined in a verb-dependent way. This may be easiest to see with an English example. The roles *arg2*, *arg3*, *arg4* have different meaning for different verbs; underlined in the following are some examples of *arg2*:

(a) The state **gave** CenTrust 30 days to sell the Rubens.

(b) Revenue **increased** 11 to 2.73 billion from 2.46 billion.

(c) One of Ronald Reagan 's attributes as President was that he rarely gave his blessing to the claptrap that **passes** for consensus in various international institutions.

In (a), arg2 represents the goal of “give”, in (b), it represents the amount of increase, and in (c) it represents yet another role. These complete different semantic relations are given the same semantic label. For unseen verbs, this makes it difficult for the semantic parser to know what would count as an *arg2*.

4 Using Automatic Parses

The results in the last section are based on the use of perfect (hand-corrected) parses drawn from the Penn Chinese Treebank. In practical use, of course, automatic parses will not be as accurate. In this section we describe experiments on semantic parsing when given automatic parses produced by an automatic parser, the Collins (1999) parser, ported to Chinese. We first describe how we ported the Collins parser to Chinese and then present the results of the semantic parser with features drawn from the automatic parses.

4.1 The Collins parser for Chinese

The Collins parser is a state-of-the-art statistical parser that has high performance on English (Collins, 1999) and Czech(Collins et al. 1999). There have been attempts in applying other algorithms in Chinese parsing (Bikel and Chiang, 2000; Chiang and Bikel 2002; Levy and Manning 2003), but there has been no report on applying the Collins parser on Chinese.

The Collins parser is a lexicalized statistical parser based on a head-driven extended PCFG model; thus the choice of head node is crucial to the success of the parser. We analyzed the Penn Chinese Treebank data and worked out head rules for the Chinese Treebank grammar (we were unable to find any published head rules for Chinese in the literature). There are two major differences in the head rules between English and Chinese. First, NP heads in Chinese are rigidly rightmost, that is to say, no modifiers of an NP can follow the head. In contrast, in English a modifier may follow the head. Second, just as with NPs in Chinese, the head of ADJP is rigidly rightmost. In English, by contrast, the head of an ADJP is mainly the leftmost constituent. Our head rules for the Chinese Treebank grammar are given in the Appendix.

In addition to the head rules, we modified the POS tags for all punctuation. This is because all cases of punctuation in the Penn Chinese Treebank are assigned the same POS tag “PU”. The Collins parser, on the other hand, expects the punctuation tags in the English TreeBank format, where the tag for a punctuation mark is the punctuation mark itself. We therefore replaced the POS tags for all punctuation marks in the Chinese data to conform to the conventions in English.

Finally, we made one further augmentation also related to punctuation. Chinese has one punctuation mark that does not exist in English. This commonly used mark, ‘semi-stop’, is used in Chinese to link coordinates within a sentence (for example between elements of a list). This function is represented in English by a comma. But the comma in English is ambiguous; in addition to its use in coordination and lists, it can also represent the end of a clause. In Chinese, by contrast the semi-stop has only the conjunction/list function. Chinese thus uses the regular comma only for representing clause boundaries. We investigated two ways to model the use of the Chinese semi-stop: (1) just converting the semi-stop to the comma, thus conflating the two functions as in English; and (2) by giving the semi-stop the POS tag “CC”, a conjunction. We compared parsing results with these two methods; the latter (conjunction) method gained 0.5% net improvement in F-score over the former one. We therefore include it in our Collins parser port.

We trained the Collins parser on the Penn Chinese Treebank(CTB) Release 2 with 250K words, first removing from the training set any sentences that occur in the test set for the semantic parsing experiments. We then tested on the test set used in the semantic parsing which includes 113 sentences(TEST1). The results of the syntactic parsing on the test set are shown in Table 7.

Table 7 Results for syntactic parsing, trained on CTB Release 2, tested on test set in semantic parsing

	LP(%)	LR(%)	F1(%)
overall	81.6	82.1	81.0
len<=40	86.1	85.5	86.7

To compare the performance of the Collins parser on Chinese with those of other parsers, we conducted an experiment in which we used the same training and test data (Penn Chinese Treebank Release 1, with 100K words) as used in those reports. In this experiment, we used articles 1-270 for training and 271-300 as test(TEST2). Table 8 shows the results and the comparison with other parsers.

Table 8 only shows the performance on sentences ≤ 40 words. Our performance on all the sentences TEST2 is P/R/F=82.2/83.3/82.7. It may seem surprising that the overall F-score on TEST2 (82.7) is higher than the overall F-score on TEST1 (81.0) despite the fact that our TEST1 system had more than twice as much training as our TEST2 system. The reason lies in the makeup of the two test sets; TEST1 consists of randomly selected long sentences; TEST2 consists of sequential text, including many short sentences. The average sentence length in TEST1 is 35.2 words, vs. 22.1 in TEST2. TEST1 has 32% long sentences (>40 words) while TEST2 has only 13%.

Table 8 Comparison with other parsers: TEST2

	≤ 40 words		
	LP(%)	LR(%)	F1(%)
Bikel & Chiang 2000	77.2	76.2	76.7
Chiang & Bikel 2002	81.1	78.8	79.9
Levy & Manning 2003	78.4	79.2	78.8
Collins parser	86.4	85.5	85.9

4.2 Semantic parsing using Collins parses

In the test set of 113 sentences, there are 3 sentences in which target verbs are given the wrong POS tags, so they can not be used for semantic parsing. For the remaining 100 sentences, we used the feature set containing eight features (path, pt, gov, position, subcat, target, head word and head POS) , the same as that used in the experiment on perfect parses. The results are shown in Table 9.

Table 9 Result for semantic parsing using automatic syntactic parses

	P(%)	R(%)	F(%)
110 sentences	86.0	70.8	77.6
113 sentences	86.0	69.2	76.7

Compared to the F-score using hand-corrected syntactic parses from the TreeBank, using automatic parses decreases the F-score by 6.4.

5 Comparison with English

Recent research on English semantic parsing has achieved quite good results by relying on the large amounts of training data available in the Propbank and Framenet (Baker *et al.* 1998) databases. But in extending the semantic parsing approach to other languages, we are unlikely to always have large data sets available. Thus it is crucial to understand how small amounts of data affect semantic parsing. At the same time, there have been no comparisons between English and other languages with respect to semantic parsing. It is thus not clear what language-specific issues may arise in general with the automatic mapping of syntactic structures to semantic relations. In this section, we compare English and Chinese by using the same semantic parser, similar verbs and similar amounts of data. Our goals are two-folds: (1) to compare the performance of the parser on English and Chinese; and (2) to understand differences between English and Chinese that affect automatic mapping between syntax and semantics. At first, we introduce the data used in the experiments and then we present the results and give analysis.

5.1 The English data

In order to create an English corpus which matched our small Chinese corpus, we selected 10 English verbs which corresponded to our 10 Chinese verbs in meaning and frequency; exact translations of the Chinese when possible, or the closest possible word when an extract translation did not exist. The English verbs and their Chinese correspondents are given in Table 10.

Table 10 English verbs chosen for experiments

English	Freq	Chinese	English	Freq	Chinese
build	46	建成	hold	120	举行
emerge	30	出现	hope	63	希望
enter	108	进入	increase	231	增加
found	248	成立	pass	143	通过
give	124	给予	publish	77	发表

Table 12 The comparison between adjuncts in English and Chinese

	English				Chinese			
Role	Before verb	After verb	Freq in test	P R F (%)	Before verb	After verb	Freq in test	P R F (%)
argM-ADV	22	43	5	0 0 0	223	0	37	91.3 56.8 70
argM-LOC	25	82	11	80 36.4 50	233	5	31	90.0 87.1 88.5
argM-MNR	22	75	14	0 0 0	11	0	1	0 0 0
argM-TMP	119	164	37	66.7 27 38.5	408	13	44	96.7 65.9 78.4

After the verbs were chosen, we extracted every sentence containing these verbs from section 02 to section 21 of the Wall Street Journal data from the Penn English Propbank. The number of sentences for each verb is given in Table 10.

5.2 Experimental Results

As in our Chinese experiments, we used our SVM-based classifier, using N one-versus-all classifiers. Table 11 shows the performance on our English test set (with Chinese for comparison), beginning with the path feature, and incrementally adding features until in the last row we combine all 8 features together.

Table 11 Experimental results of English

	Chinese	English
feature set	R/F/P	P/R/F
path	71.8/59.4/65.0	78.2/48.3/59.7
path + pt	72.9/62.9/67.5	77.4/51.2/61.6
path + position	72.5/60.8/66.2	75.7/50.9/60.8
path + hd POS	77.6/63.3/69.7	79.1/49.7/61.0
path + sub-cat	80.8/63.6/71.2	79.9/45.3/57.8
path + hd word	85.0/66.0/74.3	84.0/47.7/60.8
path + target	85.8/68.4/76.1	85.7/49.1/62.5
COMBINED	91.7/76.0/83.1	84.1/62.2/71.5

It is immediately clear from Table 11 that using similar verbs, the same amount of data, the same classifier, the same number of roles, and the same features, the results from English are much worse than those for Chinese. While some part of the difference is probably due to idiosyncracies of particular sentences in the English and Chinese data, other aspects of the difference might be accounted for systematically, as we discuss in the next section.

5.3 Discussion: English/Chinese differences

We first investigated whether the differences between English and Chinese could be attributed to particular semantic roles. We found that this was indeed the case. The great bulk of the error rate difference between English and Chinese was caused by the 4 adjunct classes argM-ADV, argM-LOC, argM-MNR, and argM-TMP, which together account for 19.6% of the role tokens in our English corpus. The average F-score in English for the four roles is 36.7, while in Chinese the F-score for the four roles is 78.6. Why should these roles be so much more difficult to identify in English than Chinese? We believe the answer lies in the analysis of the *position* feature in section 3.2.2. This is repeated, with error rate information in Table 12. We see there that adjuncts in English have no strong preference for occurring before or after the verb. Chinese adjuncts, by contrast, are well-known to have an extremely strong preference to be preverbal, as Table 12 shows. The relatively fixed word order of adjuncts makes it much easier in Chinese to map these roles from surface syntactic constituents than in English.

If the average F-score of the four adjuncts in English is raised to the level of that in Chinese, the overall F-score on English would be raised from 71.5 to 79.7, accounting for 8.2 of the 11.6 difference in F-scores between the two languages.

We next investigated the one feature from our original English-specific feature set that we had dropped in our Chinese system: *passive*. Recall that we dropped this feature because marked passives are extremely rare in Chinese. When we added this feature back into our English system, the performance rose from P/R/F=84.1/62.2/71.5 to 86.4/65.1/74.3. As might be expected, this effect of voice is mainly reflected in an improvement on arg0 and arg1, as Table 13 shows below:

Table 13. Improvement in English semantic parsing with the addition of the voice feature

	-voice	+voice
	P R F	P R F
arg0	88.9 75.3 81.5	94.4 80 86.6
arg1	86.5 82.8 84.6	88.5 86.2 87.3

A third source of English-Chinese differences is the distribution of roles; the Chinese data has proportionally more adjuncts (ARGMs), while the English data has proportionally more oblique arguments (ARG2, ARG3, ARG4). Oblique arguments are more difficult to process than other arguments, as was discussed in section 3.4. This difference is most likely to be caused by labeling factors rather than by true structural differences between English in Chinese.

In summary, the higher performance in our Chinese system is due to 3 factors: the importance of passive in

English; the strict word-order constraints of Chinese adverbials, and minor labeling differences.

6 Conclusions

We can draw a number of conclusions from our investigation of semantic parsing in Chinese. First, reasonably good performance can be achieved with a very small (1100 sentences) training set. Second, the features that we extracted for English semantic parsing worked well when applied to Chinese. Many of these features required creating an automatic parse; in doing so we showed that the Collins (1999) parser when ported to Chinese achieved the best reported performance on Chinese syntactic parsing. Finally, we showed that semantic parsing is significantly easier in Chinese than in English. We show that this counterintuitive result seems to be due to the strict constraints on adjunct ordering in Chinese, making adjuncts easier to find and label.

Acknowledgements

This work was partially supported by the National Science Foundation via a KDD Supplement to NSF CISE/IRI/Interactive Systems Award IIS-9978025. Many thanks to Ying Chen for her help on the Collins parser port, and to Nianwen Xue and Sameer Pradhan for providing the data. Thanks to Kadri Hacioglu, Wayne Ward, James Martin, Martha Palmer, and three anonymous reviewers for helpful advice.

Appendix: Head rules for Chinese

Parent	Direction	Priority List
ADJP	Right	ADJP JJ AD
ADVP	Right	ADVP AD CS JJ NP PP P VA VV
CLP	Right	CLP M NN NP
CP	Right	CP IP VP
DNP	Right	DEG DNP DEC QP
DP	Left	M(r) DP DT OD
DVP	Right	DEV AD VP
IP	Right	VP IP NP
LCP	Right	LCP LC
LST	Right	CD NP QP
NP	Right	NP NN IP NR NT
PP	Left	P PP
PRN	Left	PU
QP	Right	QP CLP CD
UCP	Left	IP NP VP
VCD	Left	VV VA VE
VP	Left	VE VC VV VNV VPT VRD VSB VCD VP
VPT	Left	VA VV
VRD	Left	VVl VA
VSB	Right	VV VE

References

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkekey FrameNet Project. In Proceeding of COLING/ACL.

Bikel, Daniel and David Chiang. 2000. Two Statistical Parsing models Applied to the Chinese Treebank. In *Proceedings of the Second Chinese Language Processing Workshop*, pp. 1-6.

Chiang, David and Daniel Bikel. 2002. Recovering Latent Information in Treebanks. In *Proceedings of COLING-2002*, pp.183-189.

Collins, Michael. 1999. Head-driven Statistical Models for Natural Language Parsing. Ph.D. dissertation, University of Pennsylvannia.

Collins, Michael, Jan Hajic, Lance Ramshaw and Christoph Tillmann. 1999. A Statistical Parser for Czech. In *Proceedings of the 37th Meeting of the ACL*, pp. 505-512.

Gildea, Daniel and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245-288.

Gildea, Daniel and Martha Palmer. 2002. The Necessity of Parsing for Predicate Argument Recognition, In *Proceedings of the 40th Meeting of the ACL*, pp. 239-246.

Kingsbury, Paul, Martha Palmer, and Mitch Marcus. 2002. Adding semantic annotation to the Penn Treebank. In *Proceedings of HLT-02.*

Kudo, Taku and Yuji Matsumoto. 2000. Use of support vector learning for chunk Identification. In *Proceedings of the 4th Conference on CoNLL*, pp. 142-144.

Kudo, Taku and Yuji Matsumoto. 2001 Chunking with Support Vector Machines. In *Proceeding of the 2nd Meeting of the NAACL*. pp.192-199.

Levy, Roger and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? ACL 2003, pp. 439-446.

Pradhan, Sameer, Kadri Hacioglu,. Wayne Ward, James Martin, and Daniel Jurafsky. 2003. "Semantic Role Parsing: Adding Semantic Structure to Unstructured Text". In the *Proceedings of the International Conference on Data Mining* (ICDM-2003), Melbourne, FL, 2003

Surdeanu, Mihai, Sanda Harabagiu, John Williams and Paul Aarseth. 2003. Using Predicate-Argument Structures for Information Extraction, In *Proceedings of ACL.*

Xue, Nianwen. 2002. Guidelines for the Penn Chinese Proposition Bank (1st Draft), UPenn.

Xue, Nianwen, Fu-Dong Chiou and Martha Palmer. 2002. Building a large-scale annotated Chinese corpus. In *Proceedings of COLING-2002.*

Xue, Nianwen, Martha Palmer. 2003. Annotating the propositions in the Penn Chinese Treebank. In *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing.*