

Automatic Learning of Chinese English Semantic Structure Mapping

Pascale FUNG, WU Zhaojun, YANG Yongsheng, and Dekai WU

Human Language Technology Center
Department of Electronic and Computer Engineering
Department of Computer Science and Engineering
University of Science and Technology, Hong Kong
{pascale@ee.ust.hk}

Abstract

We present twin results on Chinese semantic parsing, with application to English-Chinese cross-lingual verb frame acquisition. First, we describe two new state-of-the-art Chinese shallow semantic parsers leading to an F-score of 82.01 on simultaneous frame and argument boundary identification and labeling. Subsequently, we propose a model that applies the separate Chinese and English semantic parsers to learn cross-lingual semantic verb frame argument mappings with 89.3% accuracy. The only training data needed by this cross-lingual learning model is a pair of non-parallel monolingual Propbanks, plus an unannotated parallel corpus. We also present the first reported controlled comparison of maximum entropy and SVM approaches to shallow semantic parsing, using the Chinese data.

1 Introduction

In this paper we present an approach of English-Chinese cross-lingual semantic verb frame acquisition: (1) we describe two new state-of-the-art Chinese shallow semantic parsers leading to upwards of 80% precision and recall on simultaneous frame and argument boundary identification and labeling, and (2) we propose a model that employs the separate Chinese and English semantic parsers to learn cross-lingual semantic verb frame argument mappings with 86.9% accuracy, given only a pair of non-parallel monolingual Propbanks plus an unannotated parallel corpus.

There are numerous potential applications for Chinese semantic parsing and for cross-lingual semantic frame acquisition. For instance, consider the following example taken from error analysis of a typical modern statistical MT system:

input 美国政府今天表示，有关美国要求澄清报导以色列意图在所占领的戈兰高地扩大犹太人的屯垦计画，以色列尚未给予满意的回答。

system The United States Government requested clarification of Israel's intention in the occupied Golan today, on the planned expansion of Jewish settlement, Israel has not yet given a satisfactory response.

reference The United States government said today that Israel had not provided a satisfactory answer to U.S. request

for clarification about the reported plans to expand Jewish settlement in the occupied Golan Heights.

Such errors of semantic role confusion – where the system is unaware of, for example, the correct argument structure for the nominalized “*intention*” verb frame (as well as numerous other complements) – are a common source of errors in current systems for both machine translation and machine-aided translation. Different languages realize semantic roles in different surface syntactic order, and the language models and word reordering models in SMT are not always sufficient to discriminate between alternative hypotheses that may well score equally well in fluency despite high variance in translation adequacy.

One approach toward attacking such problems involves the acquisition of *cross-lingual semantic verb frames*, which capture this sort of role correspondence information in a manner independently of word reordering. Cross-lingual semantic verb frames specify the conventional patterns of alignment of semantic argument structures between a pair of semantic verb frames (or valency frames, qualia structures, etc.) in translation.

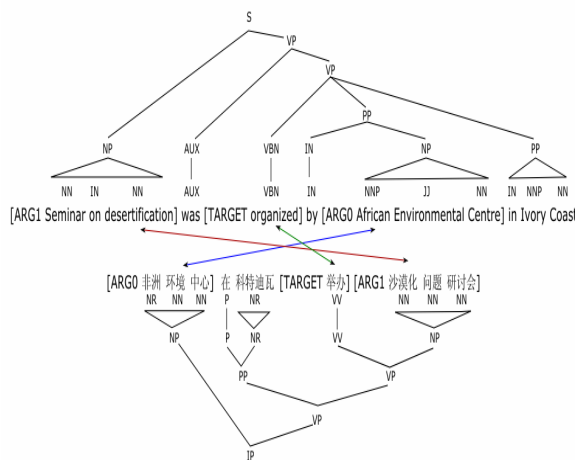


Figure 1. Bilingual Semantic Structure Mapping

While a fair amount of work on English shallow semantic parsers has been done resulting in models that turn in decent performance (Gildea and Jurafsky 2002; Kwon *et al.* 2004; Pradhan *et al.* 2004; Pradhan *et al.* 2005; Xue and Palmer 2004; Moschitti and Basili 2005), there is as yet very little research on shallow semantic parsing for languages like Chinese. The new Chinese shallow semantic parser described in this paper yields a state-of-the-art F-score of 82.01 on held-out test data from the Chinese Propo-

sition Bank. To our knowledge, the only other significant result to date on broad-coverage Chinese shallow semantic parsing is Xue and Palmer’s (2005) model which obtained an F-score of 61.3, and required greater amounts of training data (from pre-release Chinese Propbank) than the present model. For narrow-coverage Chinese shallow semantic parsing, Sun and Jurafsky (2004) described a model for 10 manually selected verbs.

As mentioned above, we employ our Chinese and English shallow semantic parsers in a new model that automatically acquires cross-lingual semantic verb frames that align semantic predicate arguments with 89.3% accuracy.

A challenge we faced is that (contrary to what one might first assume) even with semantic rather than syntactic arguments, the acquisition model still needs to be capable of dealing with the fact that predicate verb translations in English and Chinese often do *not* have the same semantic argument structure, due to cross-linguistic lexical and conceptual differences and translation idiosyncrasies. This might seem surprising since, in principle, it would seem that semantic role labels for translatable verbs ought to be preserved more closely than syntactic roles across languages, since the agents, patients, and so forth seem more likely to remain constant in translation independent of verb alternations – whereas in contrast, surface syntactic labels (subject, object, etc.) often do not survive translation, due to language-specific verb alternations. However, we will describe experimental results indicating that even semantic roles are not preserved across Chinese and English 8.7% of the time. Thus, our acquisition model cannot assume that the argument labels learned by our separately trained Chinese and English semantic parsers will necessarily correspond to each other cross-linguistically. To address this we introduce a cosine similarity model enabling our acquisition model to build and extract the bilingual semantic verb predicate-argument structure.

In the sections that follow, we first describe the Chinese shallow semantic parsers we constructed and evaluated, and report comparative experiments. In section 3 we describe the shallow semantic parsing of a Chinese-English parallel corpus. We then present an algorithm for acquiring cross-lingual semantic verb frames in section 4, with experimental results and discussion of relevant linguistic findings.

2 Learning Semantic Structures

Shallow semantic parsing uses syntax and lexical features to identify and classify predicate argument structures in a sentence. Figure 1 shows an example of a predicate argument structure constructed on top of a syntactic parse tree.

Recent advances in shallow semantic parsing have been focusing on using SVM classifiers (Gildea and Jurafsky 2003; Sun and Jurafsky 2004; Pradhan *et al.* 2004, 2005) and Maximum Entropy classifiers (Kwon *et al.* 2004) to identify and classify semantic roles in a sentence, based on syntactic and lexical features. Sun and Jurafsky (2004) showed Chinese shallow semantic parsing results on a selected set of verbs, by using gold standard syntax trees and SVM classifier. Xue and Palmer (2005) used a maximum entropy classifier with a similar set of features. They found that Chinese syntactic parsing errors, which tend to be higher than those in English, cause serious degradation in semantic parsing results.

We use the SVM-based English semantic shallow parser from Pradhan *et al.* (2004) for the English side. Since the most recent Chinese semantic parsing results from auto-

matic syntactic parsers are not comparable to the performance in English (Xue and Palmer 2005), we experimented and constructed our own Chinese semantic parsing system in order to investigate the impact of classifier choice and feature set selection. More importantly, since syntactic errors caused most of the errors in the otherwise high performing system by Xue and Palmer, we chose a high-accuracy maximum entropy based Chinese syntactic parser previously developed in-house (Fung *et al.* 2004).

Chinese sentences are first segmented, POS-tagged, chunked and syntactically parsed. The syntactic parser was trained with Penn Chinese Treebank Release 1.0, with section 001-270 (3369 sentences, approximately 90% of the CTB). The feature set we used is an extension of the standard set of features used in previous work on shallow semantic parsing. We compare using SVM and maximum entropy classifiers in the Chinese semantic parsing task. The 82.01 F-score performance achieved by our best Chinese semantic parser is superior to that reported in Xue and Palmer (2005) of 61.3. Although we did not have a common shared test set, it is reasonable to compare the numbers, since their system and both our SVM and maximum entropy systems use Propbank data for training and testing.

2.1 Feature extraction

Most recent work on English and Chinese shallow semantic parsing (Xue and Palmer 2004, 2005; Pradhan *et al.* 2004, 2005; Sun and Jurafsky 2004) employ a fairly standard set of features. We carried out further experiments on using additional features from English (Pradhan *et al.* 2005) and compared the results.

Standard features	Extended features
Predicate	Partial path
Current Parsing Word and its POS	Left path string
Head word & its POS	Parent string
Word before or after the predicate	Right path string
Path	Path trigrams
Path to the predicate	
Sub Category	
Lexical path	
First and last word of the constituent in focus	
Position	

Table 1. Feature set for Chinese semantic parsing

We compared a particular set of features used on English sentences (Pradhan *et al.* 2004) and found that they are not all useful for Chinese. For example, the *partial-path* and *path-string* features negatively impact system accuracy performance in Chinese, although the same features bring a positive effect for English. Other features, such as “Voice” and “Verb Sense Information” are either not particularly applicable to Chinese, or not available from the Chinese Propbank data we use. The final feature set we use for Chinese, as listed in Table 1, excepting the “partial path” and “path string” features, is a combination of useful features from Sun and Jurafsky (2004), Xue *et al.* (2005), and Pradhan *et al.* (2005).

2.2 Classifier Comparison

Xue and Palmer (2004) demonstrated that, by using a smaller set of features and a maximum entropy classifier, they were able to obtain semantic parsing performance comparable to that of Pradhan *et al.* (2004) which employed a SVM classifier. However, they did not compare the classifiers with the same feature set and using the same syntactic parse trees. For Chinese semantic parsing, we compared a maximum entropy based classifier developed in-house, to a SVM-based classifier, ported from the English parser (Pradhan *et al.* 2004), using the same feature set as described above in Section 2.1, and the same syntactic parse trees from our maximum entropy syntactic parser. We evaluated Chinese semantic parsing results on a held-out test set from the Chinese Propbank. Table 2 shows that using the same set of features, the SVM classifier outperforms the maximum entropy classifier on both precision and recall. Consequently, we chose to use the SVM based semantic parsing model in the subsequent experiments as described in the following sections.

	Precision	Recall	F-score
SVM	82.73%	81.30%	82.01
Maximum Entropy	77.09%	73.97%	75.70

Table 2. Comparison between SVM and ME classifiers for semantic parsing.

3 Bilingual Semantic Parsing of a Parallel Corpus

We are interested in learning the bilingual semantic structure given a predicate verb pair in English and Chinese, as in Figure 1. The predicate verb pair “*organized/举办*” have the operators *arg0* “*African Environmental Centre/非洲环境中心*”, and the operands *arg1* “*Seminar on desertification/沙漠化问题研讨会*”.

In the above example, the subject of the English sentence is *arg1*, the operand, whereas the object is *arg0*, the operator. On the other hand, the subject-object order is reversed in the Chinese sentence. The location “*Ivory Coast*” after the predicate verb and *arg1*, at the end of the English sentence, whereas the Chinese translation is before the predicate verb, after *arg0*, in the Chinese sentence.

We parsed the Xinhua Chinese-English Parallel News Text Version 1 beta 2 [LDC2002E18] to learn bilingual semantic structures. 1000 sentence pairs from the corpus were randomly chosen and parsed using the SVM-based semantic parsers for English and Chinese. These sentences contain from 8 to 25 words. Among these, 55 Chinese sentences and 97 English sentences have no identifiable predicate verb, and are eliminated from further processing. Finally, 582 sentence pairs containing at least two predicate verbs are used in our task. There are a total of 2339 predicate verbs in the Chinese sentences, and 1748 in English.

4 Bilingual Semantic Verb Frame Predicate Argument Mapping

Given semantically parsed English/Chinese sentence pairs, we use the semantic argument alignment algorithm shown in Figure 4 to acquire the bilingual predicate-argument mappings that give us the cross-lingual semantic verb frames.

With the aid of this alignment algorithm, 689 predicate-argument structure mappings are learned, with 167 unique

Chinese verbs and 352 unique English verbs. Many of these verbs are part of multiple context-dependent semantic structures. Human translation errors in the bilingual corpus, syntactic parsing and tagging errors account for some of the unmatched predicate-argument structures. Despite this, we obtained a very high accuracy of 86.9% in bilingual semantic structure mapping, as evaluated against the gold standard.

As an example of the learned structure mapping, we refer to Figure 1 again. There is a cross alignment between *arg0* and *arg1* across the translated sentence. In this example, we will learn that “*arg1[was]-organized-[by]arg0*” in English maps to “*arg0-举办-arg1*” in Chinese.

Semantic argument role labels may not be preserved across languages

Due to language differences and translation idiosyncrasies, predicate verbs in English versus

Semantic predicate argument structure alignment algorithm:

For each bilingual sentence pair:

For every source predicate verb, from left to right:

1. Extract its arguments *arg0*,...*argM*
2. Look for the translated target predicate verb in the target sentence using a bilingual lexicon, and extract *its* arguments *arg0*...*argN*
3. If we cannot find the predicate verb translation in target, then find the predicate pairs whose *arg0* translates to the source *arg0*, and extract the full set of *its* arguments *arg0*...*argN*
4. Align(*argi*, *argj*): align the source/target arguments (see below)
5. Go to step 1

Align (argument *i*, argument *j*) is done by using cosine similarity between the lexical constituents in arguments *i* and *j*, and choosing the most similar pairings:

$$\cos(\arg_i, \arg_j) = \frac{\arg_i \bullet \arg_j}{|\arg_i| |\arg_j|}$$

Figure 2. Algorithm to align previously identified and labeled semantic predicate arguments.

Chinese do not always have the same argument structure. In the example below, the English sentence is not a literal translation of the Chinese sentence, the latter meaning “*The 21st Moscow International Film Festival opens*”. The predicate verb pair is “*opens/开幕*”. The semantic structure mapping learned is “*arg1-opens-argm-loc/arg0-开幕*”. This is because the Chinese target verb 开幕 cannot be followed by an *arg1*.

English [ARG1 International Film Festival] [TARGET opens] [ARGM-LOC in Moscow]
Chinese [ARG0 第 21 届 莫斯科 国际 电影节] [TARGET 开幕]。
Gloss 21st Moscow International Film Festival opens

In another example below, the Chinese target verb is “*举行/hold*” and the argument 1 is “*会议/meeting*”, whereas the English target word is simply “*[to] meet*”

and there is no argument 1. In this case, we learned that `arg0-举行-[arg1 会议]` maps to “`arg0-meet`”.

English [ARG0 foreign ministers of the six countries] [ARGM-MOD will] [TARGET meet] [ARGM-TMP at least once a year]

Chinese [ARG0 6 国 外长 每年] [ARGM-ADV 将] [ARGM-ADV 至少] [TARGET 举行] [ARG1 一次 会议]

Gloss 6-country foreign ministers every year will at least hold one meeting)

The final mapping shows that 89.6% of the `arg0` semantic arguments are preserved across English and Chinese, 94.37% of `arg1`s are preserved, 77.8% of argument 2s, etc. This implies that, as mentioned in the introduction, the other 8.68% are mappings between arguments labeled with different semantic roles due to cross-linguistic semantic frame variation.

Further linguistic findings from the bilingual semantic structure mappings are discussed in the following subsections.

Parallel structures

Some English sentences are translated into parallel structures in Chinese. In particular, prepositional phrases in English are often translated as a separate clause, after a comma. Argument mapping is nevertheless correct despite this type of syntactic mismatching.

each year [ARG1 hundreds of people] were [TARGET killed]
 每年 , [ARG0 交通事故] [TARGET 造成] [ARG1 希腊数百人死亡], [ARG1 数千人受伤]。
 (each year, traffic accidents cause Greece hundreds of deaths, thousands of injuries.)

Ellipsis

The ellipsis of some syntactic elements, such as the subject, occurred in either English or Chinese in the parallel sentences and might lead to some null argument mapping in the other language:

English In the medium term [ARG1 a Buy Thai year] is being [TARGET considered] with a series of promotional campaigns spanning over several months.

Chinese 在中期阶段, [ARG0 政府] [ARG-ADV 打算] [TARGET 开展] [ARG1 “购买国货年” 运动] , 并准备了为期数月的一系列 宣传活动 。

Gloss In the medium term, the government plans to start the Buy Thai year campaign, and planned a several-months-series of promotional campaigns.

One-to-many argument mapping

In our proposed algorithm, arguments of the source predicate verb and those of the target predicate verb are mapped by a cosine similarity measure. In principle, it is possible for there to be one-to-many mappings. We found that when this occurs, the mapping is in fact correctly grouping an argument structure with its modifier together. In the following example, argument 1 and 2 in English are correctly mapped to argument 1 in Chinese.

English [ARG0:South Korea] to [TARGET amend] [ARG1:references to china] [ARG2:in textbooks]

Chinese [ARG0 韩国] 决定 [TARGET 修改] [ARG1 教科书中有关中国部分]

Gloss outh Korea decides amend textbook inside about China parts

5 Conclusion

We have presented new state-of-the-art results on Chinese shallow semantic parsing and English-Chinese cross-lingual verb frame acquisition. Our best Chinese shallow semantic parser yields an F-score of 82.01 on simultaneous frame and argument boundary identification and labeling, and is syntax-driven with SVM-based classifiers for semantic labeling. We have presented a model that employs the separate Chinese and English semantic parsers to learn cross-lingual semantic verb frames, producing argument mappings with 89.3% accuracy, while requiring as training data only a pair of non-parallel monolingual propbanks plus an unannotated parallel corpus. One obvious next step is to embed the shallow semantic parsers and the cross-lingual verb frame acquisition model in end-to-end MT or MAT applications. We would also like to acquire cross-lingual semantic frames for other categories besides verbs.

References

- Stephen Della Pietra and Vincent J. Della Pietra and John D. Lafferty, Inducing Features of Random Fields. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 19, number 4, pages 380-393, 1997.
- Pascale Fung, Grace Ngai, Yongsheng Yang, and Benfeng Chen, 2004. A Maximum-Entropy Chinese Parser Augmented by Transformation-Based Learning. *ACM Transactions on Asian Language Processing* 3(2): 159-168. June 2004.
- N. Kwon, M. B. Fleischman, and E.H. Hovy. 2004. FrameNet-based Semantic Parsing using Maximum Entropy Models. In *COLING 2004*.
- Alessandro Moschitti and Roberto Basili. 2005. Verb subcategorization kernels for automatic semantic labeling. In *ACL-SIGLEX Workshop on Deep Lexical Acquisition*. Ann Arbor: June 2005. 10-17.
- Sameer Pradhan. 2005. ASSERT : Automatic Statistical Semantic Role Tagger. <http://oak.colorado.edu/assert/>.
- Sameer Pradhan, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin, Daniel Jurafsky. 2005. Support Vector Learning for Semantic Argument Classification. *Machine Learning* 60(1-3): 11-39.
- Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin, and Daniel Jurafsky. 2004. Shallow Semantic Parsing using Support Vector Machines. In *Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2004)*. Boston, MA. May 2004.
- Honglin Sun and Daniel Jurafsky. 2004. Shallow Semantic Parsing of Chinese. In *Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2004)*. 249-256. Boston, MA. May 2004.
- Nianwen Xue and Martha Palmer. 2005. Automatic Semantic Role Labeling for Chinese Verbs. In *19th International Joint Conference on Artificial Intelligence*. Edinburgh, Scotland.