# Weakly Supervised SVM for Chinese-English Cross-lingual Subcategorization Lexicon Acquisition

**Xiwu Han[1] Chengguo Lv[1] Tiejun Zhao[2]**

[1] School of Computer Science and Technology, Heilongjiang University
[2] School of Computer Science and Technology, Harbin Institute of Technology

## Abstract

This paper describes experiments and the results of Chinese-English cross-lingual subcategorization lexicon acquisition based on a weakly supervised SVM method. Previous similar researches are mostly focused on statistical filtering, and there is no supervised training for the generation of hypotheses. Therefore, all these methods are unsupervised, whereas in our experiment the unsupervised hypothesis generator is replaced with an SVM classifier. And the weakly supervised SVM method was also used successfully for bilingual corpus preparation. Results of experiments indicate statistically significant improvement in the general cross-lingual acquisition performance.

**Keywords**: weakly supervised SVM, cross-lingual subcategorization, lexicon acquisition

## 1. Introduction*

According to the definition of (Chomsky 1965), subcategorization is the process that further classifies a syntactic category into its subsets, and the function of strict subcategorization features is to appoint a set of constraints that dominate the selection of verbs and other arguments in deep structure. Subcategorization of verbs, as well as categorization of all words in a language, is often implemented by means of their involved functional distributions, which constitute different environments accessible for a verb or word. Such a distribution is often called one subcategorization frame (SCF), usually integrated with both syntactic and semantic information. Motivated by cross-lingual information processing tasks such as machine translation, (Han 2008) defined Chinese-English cross-lingual subcategorization in a more syntactic way other than semantic, and the acquisition process for basic Chinese-English SCF types was managed unsupervised, i.e. at first, cross-lingual SCF hypotheses were generated according to certain heuristic rules, then filtered via statistical methods, and at last reliable cross-lingual SCF types were selected. However, possible applications of cross-lingual subcategorization information should be based on concrete lexicons other than basic SCF types since such a lexicon usually involves given verbs and their specialized SCFs.

This paper describes a weakly supervised method for the acquisition of cross-lingual syntactic subcategorization lexicon from large corpus of Chinese and English sentence pairs. By this method, we can easily exploit unparsed corpus, while previously unsupervised hypothesis generation rules must depend on parsing

* This research is jointly supported by the National Natural Science Foundation of China under Grant No. 60773069, 608731069.

results. And at the same time the supervised experiment outperformed the unsupervised significantly on the same testing corpus.

Furthermore, section 2 introduces related work basis and an easily designed baseline experiment of unsupervised acquisition method. Section 3 describes our weakly supervised method of SVM and its application scheme in both potential sentence pair recognition and cross-lingual SCF lexicon acquisition. Section 4 lists and analyzes the experiment results on the same testing corpus as that of the baseline. And the conclusion is given in section 5 with some suggestion for further work.

## 2. Related work and baseline experiment

According to the summary in (Korhonen 2001), the full description of verb subcategorization generally consists of seven kinds of linguistic knowledge: a) the number and type of arguments that a particular predicate requires, b) predicate sense in question, c) semantic representation of the particular predicate-argument structure, d) mapping between the syntactic and semantic levels of representation, e) semantic selectional restrictions or preferences on arguments, f) control of understood arguments in predicative complements, and g) diathesis alternations. But in the practices of acquisition, the actual SCF definitions or formats are often various accordingly with the concerned languages, and it seems that there is no easy solution to the problem how much subcategorization should constitute syntactically or semantically.

(Han 2008) took an NLP-task-oriented viewpoint to adopt syntactic subcategorization frames, which involves 137 Chinese basic SCF types from (Han 2005) and 82 English basic syntactic SCFs manually composed from (Korhonen 2001)'s syntactic and semantic types. And then, by rule-based heuristic methods, a set of 654 basic cross-lingual SCF types was established for parallel Chinese and English predicates.

However, reasonable and practical cross-lingual subcategorization information should include concrete lexicons that include given verb pairs with special SCF types they could enter. A simple scheme for building such a lexicon is to acquire cross-lingual SCFs in a way of monolingual SCF acquisition.

### 2.1. Prepare the corpus

Most subcategorization researches focus on predicative verbs, and thus those sentences with parallel predicative verbs and the two parallel predicates are to be recognized and gathered for further experiment. Our total corpus consists of 1,000,000 bilingual sentence pairs of English and Chinese, which were gathered either from public and free Internet resources or our translation works, and parsed with (Collins 1999)'s head-driven parser and the head-driven parser of MI&TLAB in Harbin Institute of Technology (Cao 2006).

We used the same method of syntactic compatibility and bilingual verb dictionary as in (Han 2008) to fulfil the corpus preparation. First, the former was used to select possible parallel sentence pairs, and then the latter was applied to recognized potential predicates.

Syntactic compatibility is based on the hypothesis that common word classes, such as nouns, pronouns, verbs, adjectives, etc, mainly constitute the vocabularies of most natural languages and may well survive translation activities resulting in sentence pairs with parallel predicates. Hence the cross-lingual syntactic compatibility ***D*** is defined as follows.

$$D=\sum_{i=1}^{n}\lambda_i\frac{Min(|GE_i|,|GC_i|)+1}{Max(|GE_i|,|GC_i|)+1} \tag{1}$$

$GE_i$ is an English grammatical category, $|GE_i|$ is the number it occurs in the English sentence, and $GC_i$ is the Chinese counterpart. $\lambda_i$ is the weight for the concerned category, which was estimated by a simple gradient descent algorithm on a sample of 10,000 manually analysed sentence pairs.

For this task, a maximum likelihood estimation filtering method was employed with a threshold of 0.79 on $\boldsymbol{D}$. Candidate sentence pairs would be accepted if the syntactic compatibility between the related English and Chinese sentences surpasses the threshold. Evaluation on a sample of 5,000 sentence pairs showed a precision ratio of 79.4% and a recall ratio of 53.94%.

Our verb dictionary is organized as bilingual verbal synonym classes, and there are altogether 3,611 entries including 67,836 Chinese and English verbs, which were drawn from the bilingual dictionary of the Chinese-English machine translation system of CEMT2K developed by MI&TLAB, English WordNet v. 1.2 and Chinese Extended Tongyicicilin v. 1.0. The algorithm for recognizing parallel predicates is described as follows.

For each sentence pair

- specify the English predicate $\boldsymbol{V_e}$ according to the English parsing results;
- form the Chinese predicate candidate set $\boldsymbol{S_c}$ with all potential words, such as verbs or adjectives;
- for each candidate $\boldsymbol{V_c}$ in $\boldsymbol{S_c}$, accept the pair <$\boldsymbol{V_e}$, $\boldsymbol{V_c}$> as parallel predicates if they appear in one entry of the bilingual dictionary.

Manual analysis on 2,000 sentence pairs showed that nearly 88% parallel predicates were recalled and the precision was about 64.5%. And further analysis indicated that most errors and unrecalled parallel sentence pairs or predicates were mainly due to bad POS tagging or parsing results of the Chinese or English sentences.

## 2.2. Baseline experiment

Finally we got a corpus of about 360,000 sentence pairs for the baseline experiment. And generally, the typical framework for SCF acquisition consists of four parts, i.e. the pre-processor, the argument pattern extractor, the SCF hypothesis generator, and the statistic filter.

As mentioned above, here our corpus was pre-processed with (Collins 1999)'s parser and (Cao 2006)'s parser respectively. Argument patterns for Chinese sentences were extracted via the rule-based analyzer of (Han 2005) with an argument token precision estimated as 86.5%, while the 66 rules of (Han 2008) performed the similar task for English counterparts with a token precision of about 92.6%. And for hypothesis generation, we also adopted a heuristic method of ontological arguments as (Han 2008) did, which achieved a Chinese hypothesis token precision of 82.7%, and one of 84.44% for English, and the bilingual hypothesis token precision reached nearly 68%.

We used a two-fold MLE as filtering method, which is inspired by the theory of diathesis alternations. (Han 2006) used the method to filter SCF hypotheses for English verb lexicon building and (Han 2008) used it to acquire basic cross-lingual SCF types for Chinese and English verb pairs.

There are typically two MLE filters employed. For each verb involved, first a common MLE filter is used, but it employs a threshold $\theta_1$ that is much higher than usual, and those SCF hypotheses that satisfy the requirement are accepted.

Then, all of the remained hypotheses are checked by another MLE filter seeded with diathesis alternations as heuristic information and equipped with a much lower threshold $\theta_2$. Any hypothesis $scf_i$ left out by the first filter will be accepted if its probability exceeds $\theta_2$ and it is an alternative of an SCF type $scf_j$ that has been accepted by the first filter, which means that $p(scf_i|scf_j,v)>0$ and $scf_j \in \boldsymbol{SCF_{accepted}}$. The filtering process will be performed repeatedly for those unaccepted hypotheses until no more hypotheses can be accepted for the verb.

We modified (Han 2008)'s method a little to adjust it to the cross-lingual SCF lexicon acquisition. We used the Chinese SCF diathesis alternations described in (Han 2005) as heuristic information, and the algorithm may be written briefly as follows.

For hypotheses of a predicate pair ($V_c$, $V_e$) with an English SCF $escf_i$,

- if $p(escf_i, cscf_i) > \theta_1$, accept the hypothesis into set $\boldsymbol{S}$;
- else if $p(escf_i, cscf_i) > \theta_2$, and $p(cscf_i|cscf_j) > 0$, and $(escf_i, cscf_j) \in \boldsymbol{S}$, accept the hypothesis into set $\boldsymbol{S}$;
- go to the first step till $\boldsymbol{S}$ doesn't increase.

Here, ($escf_i$, $cscf_i$) is a bilingual SCF hypothesis, $p(cscf_i|cscf_j) > 0$ means that $cscf_i$ is a diathesis alternative of $cscf_j$, and $\boldsymbol{S}$ will be the acquired SCF lexicon for ($V_c$, $V_e$).

### 2.3. Baseline evaluation results

From the total 1,000,000 bilingual sentence pairs, we drew out 362,453 sentence pairs with possible parallel predicative verbs, on which the baseline acquisition experiment was performed.

We manually analyzed 150 sentence pairs for each pair of the 20 parallel predicates listed in Table 1, where **N** is the number of supporting sentence pairs for each verb pair. Against this gold standard, the baseline acquisition results were evaluated in terms of precision, recall and F-measure of bilingual SCF types. As in SCF acquisition for a single language, precision is the percentage of types that the system proposes correctly, while recall is the percentage of types in the gold standard that the system proposes.

$$\text{Precision} = \frac{|\text{True positives}|}{|\text{True positives}|+|\text{False positives}|} \tag{2}$$

$$\text{Recall} = \frac{|\text{True positives}|}{|\text{True positives}|+|\text{False negatives}|} \tag{3}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision } + \text{Recall}} \tag{4}$$

Here, true positives are correct cross-lingual SCF types proposed, false positives are incorrect types proposed, and false negatives are correct types that are not proposed.

Table 1: Parallel predicates for evaluation

| **N** | **E-v** | **C-v** | **N** | **E-v** | **C-v** |
|---|---|---|---|---|---|
| 322 | say | 说 | 199 | like | 喜欢 |
| 303 | have | 有 | 193 | write | 写 |
| 296 | know | 知道 | 191 | look | 看 |
| 229 | think | 想 | 188 | think | 认为 |
| 216 | find | 发现 | 185 | see | 看到 |
| 213 | want | 想 | 185 | ask | 问 |
| 213 | like | 想 | 183 | eat | 吃 |
| 204 | see | 看 | 181 | hope | 希望 |
| 203 | come | 来 | 180 | buy | 买 |
| 201 | tell | 告诉 | 178 | see | 看见 |

For the experiment, we empirically set $\theta_1$ to be 0.35, and $\theta_2$ 0.005. Table 2 lists the general performance of the baseline experiment.

Table 2: Acquisition performances

| Precision | Recall | F-measure |
|---|---|---|
| 88.3% | 61.2% | 72.3% |

Analysis on the testing corpus showed that there exist two causes to the wrongly proposed and unrecalled types. First, the rule-based hypothesis generator performed limitedly. Second, supporting corpus for each predicate pair was too small to simulate the actual distribution of cross-lingual SCF types.

## 3. Weakly supervised SVM scheme

In the baseline experiment, both corpus preparation and cross-lingual SCF lexicon acquisition were managed by means of unsupervised methods, limitations of which had obviously harmed the general performance according to our analysis. And supervised methods might well play a better role as a possible alternative, but supervised methods also call for large training corpus of standard labeling or classification that in turn requires too much manual work. However, a weak supervision might be derived from the outputs of supervised methods and thus manual work can be avoided.

We chose SVM as the practical model of our weakly supervised methods. This experiment consists of three parts, i.e. collecting parallel sentence pairs, recognizing predicates, and cross-lingual SCF hypothesis generation.

### 3.1. Parallel sentence pair collection

The training corpus for this SVM model is made up of two parts. One part consisting of 20,000 positive samples for parallel sentence pairs was drawn from the corpus of the output of a similar unsupervised experiment as described in Section 2 except that the two thresholds were set differently with $\theta_1$ as 0.4 and $\theta_2$ as 0.01. Another part consisting of 30,000 negative samples for non-parallel sentence pairs was randomly selected from the rejected sentence pairs by the unsupervised corpus preparation.

Only POS tags of the Chinese and English sentences were used to represent the training corpus, and for one sentence pair the POS tags were organized in a linear string with those of the English sentence following those of the Chinese, so here the sentences need not be parsed any more.

A string-input SVM[1] with an edit distance kernel was applied to this task. Evaluation on the classification results of the same testing corpus in Section 2.1 showed a precision ratio of 89.2% and a recall ratio of 76.4% for parallel sentence pair collection.

### 3.2. Predicate recognition

Previously the English predicate verbs were specified according to the outputs of Collin's parser, while now we used an SVM classifier to recognize them. The describing vectors were organized as 7-length windows of POS tags around English verbs in the sentences, i.e. $<x_{-3}, x_{-2}, x_{-1}, x_0, x_1, x_2, x_3>$ with $x_0$ denoting POS of the verb, $x_{-i}$ denoting POS of the preceding words, and $x_i$ of the following words.

Positive training samples for this model were draw from the 20,000 English sentences mentioned in Section 3.1, and 20,000 negative ones were randomly gathered from non-predicate verbs occurring in our English corpus. The same SVM tool kit was applied except that now it took vectors as input and used a polynomial kernel.

The trained classifier was then used to determine predicates for English sentences gathered in Section 3.1. Although this weakly supervised method only outperformed Collin's parser by a very little percentage, its application skipped the parsing step. Whereas, Chinese predicates were also recognized by means of the bilingual verb dictionary.

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/string/libsvm-2.84-string.zip

### 3.3. Hypothesis generation

Two SVM classifiers for Chinese and English were trained separately. For each mono-lingual SCF basic type, 1,000 sentences with automatic SCF labeling were randomly drawn from the output of the baseline experiment, and thus a Chinese training corpus of 137,000 sentences and an English training corpus of 82,000 were established. These SVM models also took vector inputs with description of 11-length windows of POS tags, i.e. <$x_{-3}$, $x_{-2}$, $x_{-1}$, $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$, $x_7$, $x_8$>, $x_{-i}$ denoting POS of the preceding words, and $x_i$ of the following words, while the predicate was not considered here.

Unlike the two previous binary classifiers, hypothesis generation involves multi-classes, for which we employed a pair-wise mechanism. Sampling analysis indicated that this method achieved a Chinese hypothesis token precision of 83.9%, and one of 86.6% for English, which was nearly 2% higher than that of the unsupervised counterpart.

## 4. Experiments and evaluation

Two weakly supervised experiments were performed. The first one was designed to test the weakly supervised SCF hypothesis generator. In this experiment we only replaced the rule-based generators respectively with the trained SVM classifiers for Chinese and English, and all the other parts were kept as the same of those of the baseline experiment. And on the same testing corpus listed in Table 1, with the same thresholds, acquisition performance was estimated as in Table 3. Compared with the baseline experiment, the precision ratio improved a little while the recall ratio increased significantly.

Table 3: Acquisition Performances

| Precision | Recall | F-measure |
|---|---|---|
| 90.2% | 74.6% | 81.7% |

The other experiment involved the total weakly supervised scheme. From the 1,000,000 sentence pairs, 448,623 pairs of useful corpus were drawn by the weakly supervised corpus preparation method, then possible predicates were recognized by the method described in Section 3.2, and at last all the rest parts were preceded as in the first experiment above.

The evaluation of this experiment was also made against the same gold standard as for the baseline experiment, but supporting corpus for each verb pair had increased by about 50 sentence pairs on average. And Table 4 gives the acquisition performances. This time the recall ration was promoted significantly again.

Table 4: Acquisition Performances

| Precision | Recall | F-measure |
|---|---|---|
| 92.3% | 79.8% | 86.1% |

## 5. Conclusion

This paper presents a weakly supervised SVM method to improve the acquisition performance of Chinese English cross-lingual subcategorization lexicon. Procedure of subcategorization acquisition mainly includes two typical steps: a. Subcategorization hypotheses are generated according to certain heuristic rules; b. Hypotheses are filtered via statistical methods and reliable subcategorization types are selected. Previous efforts to improve the acquisition performance are focused on statistical filtering, and there is no supervised training for the generation of hypotheses in relevant experiments, whereas our experiment replaced the unsupervised hypothesis generator with a weakly supervised SVM classifier. And this method was also used successfully for bilingual corpus preparation. Evaluation on the results of a few experiments indicates that statistically significant improvement has been achieved in the general cross-lingual acquisition performance.

In future work, we will improve our methods in several aspects. It might work better by trying more complicated methods to combine the rule-based method and the weakly supervised scheme. And larger corpus should also be exploited for more useful cross-lingual SCF information.

**Acknowledgement** We are obliged to IRLAB in Harbin Institute of Technology for the dictionary resources of Chinese Extended Tongyicicilin v. 1.0 they provided us generously.

## 6. References

[1] Chomsky, Noam. *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA, 1965.

[2] Han, Xiwu, Tiejun Zhao and Conhui Zhu. Cross-lingual Syntactic Subcategorization Analysis Based on Chinese and English Sentence Pairs. *Pro of First International Conference on Global Interoperability for Language Resources*, pp. 175-182 2008.

[3] Korhonen, Anna. *Subcategorization Acquisition*, Dissertation for PhD, Trinity Hall University of Cambridge, 2001.

[4] Han, Xiwu. *Research on Automatic Acquisition of Chinese Verb Subcategorization*, Dissertation for PhD, Harbin Institute of Technology, Harbin, 2005.

[5] Collins. M. *Head-Driven Statistical Models for Natural Language Parsing*. PhD Dissertation, University of Pennsylvania, 1999.

[6] Cao, Hailong. *Research on Chinese Syntactic Parsing Based on Lexicalized Statistical Model*, Dissertation for PhD, Harbin Institute of Technology, Harbin, 2006.

[7] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/string/libsvm-2.84-string.zip