Asian Journal of Information Technology 7 (8): 381-387, 2008
ISSN: 1682-3915

English and Chinese Verb Subcategorization Lexicon Construction

Xiwu Han
School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China

Abstract: Verb subcategorization information, mainly coding the types of distribution of predicative features, is indispensable knowledge for further development in the field of natural language processing and thus, subcategorization lexicon construction has become more and more important for especially corpus linguistics. This study describes, our techniques and experiments on subcategorization acquisition and relevant lexicon construction for English and Chinese verbs, which actually includes 4 parts. In the first 3 parts, we will introduce our research on English subcategorization, Chinese subcategorization and English-Chinese cross-lingual subcategorization, respectively and in the fourth part some simple applications of constructed lexicons to concrete NLP tasks will be discussed.

Key words: English verb, Chinese verb, subcategorization lexicon, cross-lingual, NLP tasks

INTRODUCTION

Subcategorization is the process that further classifies a syntactic category into its subsets. Chomsky (1965) defines the function of strict subcategorization features as appointing 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 or environment is called one subcategorization frame (SCF or SF), usually integrated with both syntactic and semantic information. Since verbs are mostly the central pegs for various syntactic relations to hang on, lexicons with verb subcategory specified have always been one of the most desirable outcomes of traditional linguistics. Since (Brent, 1991), there have been a considerable amount of researches focusing on verb lexicons with respective subcategorization information specified both in the field of traditional linguistics and that of computational linguistics. As for the former, subcategory theories illustrating the syntactic behaviours of verbal predicates are now much more systemically improved (Korhonen, 2001). And for auto-acquisition and relevant application, researchers have made great achievements not only in English (Briscoe and Carroll, 1997; Korhonen *et al*., 2003), but also in many other languages, such as Germany (Shulte im Walde, 2002) and Czech (Sarkar and Zeman, 2000).

This study is about, our efforts on the construction of English and Chinese verb subcategorization lexicons under supports from the national natural science foundation of China.

ENGLISH SUBCATEGORIZATION LEXICON CONSTRUCTION

The construction of English subcategorization is the earliest launched project and has always been the most thoroughly explored research. Automatically, acquired English verb lexicons with subcategorization information have already proved accurate and useful enough for some NLP purposes (Korhonen, 2001). Korhonen (2002) reported that semantically motivated SCF acquisition achieved a precision of 87.1%, an absolute recall of 71.2% and a relative recall of 85.27% and thus, making the acquired lexicon much more accurate and useful. Here, by absolute recall, we refer to the figure computed against the background of input corpus, while relative recall is against the set of generated hypotheses. However, the accuracy still shows room for improvement, especially for those SCF hypotheses with low frequencies. Detailed analysis on the acquisition system and some resulting data shows that three main causes should account for the comparatively unsatisfactory performance: the imperfect hypothesis generator, the Zipfian distribution of syntactic patterns, the incomplete partition over SCF types of a given verb. The first problem mainly comes from the inadequate parsing performance and noises existing in the corpus, while the other two problems are inherent to natural languages and should be solved in terms of acquisition techniques particularly during the process of hypothesis selection. Therefore, we proposed a new filtering method via diathesis alternation, which improved the performance of Korhonen's acquisition system remarkably, with the precision increased to 91.18% and recall unchanged, making the acquired lexicon much more practical for further manual proofreading and other NLP uses.

Corpus in use: The background of this experiment is the public resource for subcategorization acquisition of English verbs, provided by Korhonen (2001) in her personal home page. The data include 30 verbs, as shown in Table 1, their unfiltered SCF hypotheses, which were automatically generated via Briscoe and Carroll's (1997) SCF acquisition system and the manually established standard. For each verb, there is a corpus of 1000 sentences extracted from the BNC and all together 42 SCF types are involved in the corpus.

Diathesis alternations and filtering: Diathesis alternations are generally regarded as alternative ways, in which verbs express their arguments. Subcategorization of verbs has much to do with diathesis alternations and most SCF research regard information of diathesis alternation as an indispensable part of subcategorization (Korhonen, 2001; McCarthy, 2001). Therefore, one may conclude that, for subcategorization acquisition, the independence assumption supporting the MLE filter is not as appropriate as previously thought. For a given verb, the assumption will be appropriate and sufficient if and only if there is no diathesis alternation between all the SCFs it enters and formula (1) and (2) in the study are efficient enough to serve as a foundation for the MLE filtering method. Otherwise, if there are diathesis alternations between some of the SCFs that a verb enters, then formula (1) and (2) must be modified as illustrated in formula (3) and (4). In either case, for the sake of convenience, it would be better to combine the formulas as shown in (5) and (6).

$$\forall i, \forall j, i \neq j, p(scf_i \mid scf_j, v) = 0 \quad (1)$$

$$\sum_{i=1}^{n} p(scf_i \mid v) = 1 \quad (2)$$

$$\exists i, \exists j, i \neq j, p(scf_i \mid scf_j, v) > 0 \quad (3)$$

$$\sum_{i=1}^{n} p(scf_i \mid v) > 1 \quad (4)$$

$$\forall i, \forall j, i \neq j, p(scf_i \mid scf_j, v) \geq 0 \quad (5)$$

$$\sum_{i=1}^{n} p(scf_i \mid v) \geq 1 \quad (6)$$

According to this theory, we formed a new filtering method with diathesis alternations as heuristic information, which is, in fact, derived from the simple MLE filter and based on formula (5) and (6). The algorithm can be briefly expressed as shown in Table 2.

Table 1: English verbs in use

Add	Agree	Attach	Lend	Lock	Marry
Bring	Carry	Carve	Meet	Mix	Move
Chop	Cling	Clip	Offer	Provide	Visit
Fly	Cut	Travel	Push	Sail	Send
Drag	Look	Give	Slice	Supply	Swing

Table 2: The new filtering method

For hypotheses of a given verb v,

1. If $p(scf_i|v) > \theta_1$, accept scf_i into the output set S;
2. Else if $p(scf_i|v) > \theta_2$ & $p(scf_i|scf_j, v) > 0$ & $scf_j \in S$, Accept scf_i into set S;
3. Go to step 1 until S doesn't increase

Table 3: Performance comparison

Methods	No-f	MLE	Ours	Kor.
P (%)	47.85	67.89	91.18	87.10
Ab_R (%)	34.62	32.52	32.52	71.20
Re_R (%)	100.00	93.93	93.93	85.27
Ab_F	40.17	43.98	47.94	78.35
Re F	64.73	78.81	92.53	86.18

In our method, two filters are employed. For each verb involved, first a common MLE filter is used, but it employs a threshold θ_1 that is much higher than usual and those SCF hypotheses that satisfy the requirement are accepted. Then, all of the remainder of the hypotheses are checked by another MLE filter seeded with diathesis alternations as heuristic information and equipped with a much lower threshold θ_2. Any hypothesis scf_i left out by the first filter will be accepted if its probability exceeds θ_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 $scf_j \in S$. The filtering process will be performed repeatedly for those unaccepted hypotheses until no more hypotheses can be accepted for the verb.

Experimental evaluation: In the experiment we empirically set $\theta_1 = 0.2$, which is ten times of Korhonen's threshold for her MLE filter; $\theta_2 = 0.002$, which is one 10th of the Korhonen's. Table 3 lists the performances of the baseline method of non-filtering (No_f), MLE filtering with $\theta = 0.02$ and our filtering method on the evaluation corpus and also gives the best results (Kor) of Korhonen's semantic motivation to make a comparison. Here, Ab_R is the absolute recall ratio, Re_R the relative recall ratio, Ab_F the absolute F-measure that is calculated from Precision and Ab_R and Re_F the relative F-measure that is from precision and Re_R.

The evaluation shows that our new filtering method improved the acquisition performance remarkably: Compared with MLE, precision increased by 23.29%, recall ratio remained unchanged, absolute F-measure increased by 3.96 and relative F-measure increased by 13.72; Compared with Korhonen's best results, precision, Re_R and Re_F also increased, respectively. Thus, the general performance of our filtering method makes the acquired

lexicon much more practical for further manual proofreading and other NLP uses. What's more, the data also implies that there is little room left for improvement of the statistical filter, since the absolute recall ratio is only 2.1% lower than that of the non-filtering method. Whereas, detailed analysis of the evaluation corpus shows that the hypothesis generator accounts for about 95% of those unrecalled and wrongly recalled SCF types, which indicates, for the present time, more improvement efforts need to be made on the first step of subcategorization acquisition, i.e., hypothesis generation.

Other efforts to improve english subcategorization acquisition: We also tried some other methods for improvement of English verb subcategorization acquisition. For example, an approach based on clustering and classification was proposed, which improved the general performance of Korhonen's system. Precision of hypothesis generation was increased by 8.88% and that of MLE testing by 15.71%, with the recall rates unchanged. Another experiment used a weakly supervised method for English subcategorization acquisition, where the unsupervised hypothesis generator is replaced with an SVM classifier.

CHINESE SUBCATEGORIZATION LEXICON CONSTRUCTION

The construction of subcategorization lexicons generally involves tasks of predefining a basic SCF set for the concerned language and automatic acquisition of SCF information for individual verbs. Before our research, relevant researches on Chinese verbs are generally limited to case grammar, valency, some semantic computation theories and a few papers on manual acquisition or prescriptive designment of syntactic patterns. Due to irrelevant initial motivations, syntactic and semantic generalizabilities of the consequent outputs are not in such a harmony that satisfies the describing granularity for SCF. For instances, Chinese sentence patterns provided by Zhang (1999) and Hu *et al.* (1995) are generally based on case information with comparatively too strong semantic, while too weak syntactic generalization and those of Jin (2001) based on augmented valency features on the contrary are syntactically more generalized than required by the task. Therefore, we need to fulfil the predefinition task at first.

Predefinition for basic Chinese SCFs: Based on some relevant linguistic theories and our observation of real corpus, SCF for Chinese verbs can be described as a quintuple grammar <V, T_A, N_A, P_A, C_L>, which is context-sensitive to some extent. Herein and below, for Chinese verbs: V is a set of verbs capable of filling in the predicate slot, T_A is a set of argument types and T_A ={NP, VP, QP, BP, PP, BAP, BIP, TP, MP, JP, S} (Table 4), N_A is a set of numbers of argument slots, P_A is the set of positions for argument slots and C_L is the set of constant labels that may be added to some particular SCF and C_L ={"zhe"(着), "le"(了), "guo4"(过), "mei2"(没), "bu4"(不)}, where the first three furnish SCF aspects and the last 2 offer negation options. Besides, in conformance with our observation on linguistic behaviours of Chinese verbs and in order to distinguish arguments from adjuncts, we prescribe some constraints or rules over the occurrence of each of the five elements in an SCF, as shown in Table 5.

We then used Flexible Maximum Likelihood (FML), a variational filtering method of the simple maximum likelihood (ML) with observed relative frequencies, to the task of predefining a basic SCF set for Chinese verb subcategorization acquisition. By setting a flexible threshold for SCF probability distributions over 1774 Chinese verbs, we obtained 127 basic SCFs with a reasonably practical coverage of 98.64% over 43,000 Chinese sentences. After complementation of 11 manually observed SCFs, a both linguistically and intuitively acceptable basic SCF set was predefined for future SCF acquisition work.

Chinese subcategorization acquisition: There are 4 steps in the process of our automatic acquisition experiment for Chinese subcategorization. First, the corpus is processed with a cascaded HMM parser; second, every possible local patterns for verbs are abstracted and then, the verb patterns are classified into SCF hypotheses according to the predefined set; at last, hypotheses are filtered statistically and the respective frequencies are also recorded. The actual application program consists of 6 parts as shown in the following paragraphs.

Segmenting and tagging: The raw corpus is segmented into words and tagged with POS by the comprehensive segmenting and tagging processor developed by MI&TLAB of Computer Department in Harbin Institute of Technology. The advantage of the POS definition is that it describes some subsets of nouns and verbs in Chinese.

Parsing: The tagged sentences are parsed with a cascaded HMM parser, developed by MI&TLAB of HIT, but only the intermediate parsing results are used. The training set of the parser is 20,000 sentences in the Chinese Tree Bank of (Zhao, 2001).

Error-driven correction: Some key errors occurring in the former two parts are corrected according to manually obtained error-driven rules, which are generally about words or POS in the corpus.

Table 4: Chinese SCF argument types

T_A	Definition
NP	Nominal phrase
VP	Verbal phrase
QP	Tendency verbal complement
BP	Resulting verbal complement
PP	Positional phrase
BAP	Phrase headed by "ba3" (把)
BIP	Phrase headed by "bei4" (被) or other characters with passive sense
TP	Temporal phrase
MP	Quantifier complement
JP	Adjective or adverb or "de" (得) headed complement
S	Clause or sentence

Table 5: Relative rules for SCF description

Elements	Rules
V	Only one v (v∈V) except in repeating alternatives with one v but two slots
T_A	
NP	No more than two in a series and no more than three in one SCF
VP,S	No serial occurrences
QP, BP, JP	No serial occurrences and only occurring after v
BAP, BIP	No more than one occurrence
TP, PP	No co-occurrences with NP before v
MP	No serial occurrences nor presence in adjacency before NP
N_A and P_A	Shady elements without substantial presence in SCF
C_L	Notated as a binary vector, e.g. {01110}

Table 6: An example of auto-acquisition

No	Actions	Results
a)	Input	两个人在大伙儿的追·下·明了老人的身份。
b)	Tag and parse	BNP[BMP[两/m 个/q]人/ng]在/p NDE[大伙儿/r 的/usde]BVP[追·/vg 下/vq]BVP[·明/vg 了/ut]NP[老人/nc 的/usde 身份/ng]。/wj
c)	Correct errors	BNP[BMP[两/m 个/q]人/ng]在/p NDE[大伙儿/r 的/usde 追·/vg 下/vq]BVP[·明/vg 了/LE]NP[老人/nc 的/usde 身份/ng]。/wj
d)	Abstract patterns	BNP PP BVP[vg LE] NP
e)	Generate hypothesis	NP V NP {01000}
f)	Filter hypotheses	NP V NP {01111}

Table 7: Verbs in the testing set

Verbs	English	Tokens	Verbs	English	Tokens
·展	Develop	1,006	建立	Set up	1,186
表·	Behave	1,007	·持	Insist	1,200
决定	Decide	1,038	想	Think	1,200
·束	End	1,140	要求	Require	1,200
·始	Begin	1142	写	Write	2,000
·	Read	503	希望	Hope	620
··	Find	529	看	See	645
考·	Reckon	543	投入	Invest	679
拉	Pull	544	··	Know	722
反映	Report	612	送	Send	800

Pattern abstraction: Verbs with largest governing ranges are regarded as predicates, then local patterns, previous phrases and respective syntactic tags are abstracted and isolated parts are combined, generalized or omitted according to basic phrase rules in (Zhao, 2001).

Hypothesis generation: Based on linguistic restraining rules, e.g. no more than two NP's occurring in a series and no more than three in one pattern and no PP TP MP occurring with NP before any predicates, the patterns are coordinated and classified into the predefined SCF groups. In this part, about 5% unclassifiable patterns are removed.

Hypothesis filtering: According to the statistical reliability of each type of the SCF hypotheses and the linguistic principle that arguments occur more frequently with predicates than adjuncts do, the hypotheses are filtered by means of statistical methods, for this task which are Binomial Hypotheses Testing (BHT) and Maximum Likelihood Estimation (MLE).

In Table 6, for example, when acquiring SCF information for "证明" (prove) and a related sentence in the corpus is: our tagger and parser will return b) error-driven correction will return c) with errors of NDE and the 1st BVP corrected. Since, the governing range of "证明" is larger than that of "追问" (ask), the other verb in this sentence, the program abstracts its local pattern BVP [vg LE] and previous phrase BNP, generalizes BNP and NDE as NP, combines the second NP with isolated part "在/p" into PP and returns d). Then the hypothesis generator return e) as the possible SCF in which the verb may occurs. Actually in the corpus there are 621 hypothesis tokens generated and among them 92 ones are of same arguments with e) and thus, e) can pass the hypothesis testing, so we obtain one SCF for "证明" as f).

With this experiment, we acquired an SCF lexicon for 3,558 common Chinese verbs from the corpus of people's daily. In the lexicon, the minimum number of SCF tokens for a verb is 30 and the maximum is 20,000. In order to check, the acquisition performance of the used system, we evaluated a part of the lexicon against a manual gold standard. The testing set includes 20 verbs of multi syntactic patterns and for each verb there are 503~2,000 SCF tokens with the total number of 18,316 (Table 7). Table 8 gives the evaluation results for different filtering methods, including non-filtering, BHT and MLE with thresholds of 0.001, 0.005, 0.008 and 0.01. We calculated the type precision and recall as Korhonen (2001) did.

According to Table 8, all other filtering methods outperform non-filtering and MLE is better than BHT. Among the four MLE thresholds, 0.008 achieves the best comprehensive performance but its F-measure is only 0.74 larger than that of 0.01, while its precision drops by 2.4%. Hence, we chose 0.01 as the threshold for the whole experiment with purpose to meet the practical requirement of high precision and to avoid possible over-

Table 8: System performance for different filtering methods

Measures/ methods		Precision (%)	Recall (%)	F-measure
Non-filtering		37.43	85.90	52.14
BHT		50.00	57.20	53.36
MLE	0.001	39.20	85.90	53.83
	0.005	40.30	83.33	54.33
	0.008	58.20	54.50	56.30
	0.010	60.60	51.30	55.56

fit phenomena. Finally, with a confidence of 95%, we can estimate the general performance of the acquisition system with precision of 60.6±2.39% and recall of 51.3±2.45%.

CROSS-LINGUAL SUBCATEGORIZATION LEXICON CONSTRUCTION

Researches on subcategorization acquisition for a single language have met with considerable achievements since (Brent, 1991) pioneering research, e.g. those of English (Korhonen, 2001), German (Shulte im Walde, 2002), Spanish (Chrupala, 2003), Czech (Sarkar and Zeman, 2000) and Portuguese (Gamallo *et al*., 2002) and the above mentioned of our reaserch on Chinese. However, relevant cross-lingual phenomena are only dealt with theoretically in a few remarks lying sparsely in linguistic books for translation (Baker, 2000) or second language acquisition (Ellis, 2000).

Although, subcategorization frames (SCF) are generally regarded as functional distributions integrated with both syntactic and semantic information, both concrete definitions and formats for SCF vary greatly from one language to another. This may be the most severe obstacle in cross-lingual subcategorization research. Subcategories, nevertheless, exist universally in almost all natural languages for almost all linguistic categories. Thus, it remains an interesting and challenging question how much for subcategorization theories is common, comparable or compatible cross-lingually.

Syntactic subcategorization: We adopted syntactic subcategorization frames for our special task. Our 138 Chinese basic SCF types are purely syntactic. For English we used 82 English basic syntactic SCFs, 77 of which were manually regrouped from (Korhonen, 2001) types and 5 were formed according to real corpus with no counterparts in (Korhonen, 2001). The syntactic argument types are listed in Table 9, where AS stands for 'a', IT for 'it', RP for particles and so on.

Crosslingual subcategorization acquisition: We first used a combination method of bilingual verb dictionary and syntactic compatibility to abstract those with parallel predicates from 650,000 English-Chinese bilingual

Table 9: Syntactic argument types for English

AP	DP	SS	AS-NP	PASS-VP	VPING
AS-AP	IT	TO-VP	AS-PASS-VP	PP	WH-SS
AS-IF-SS	NP	VP	AS-VPING	RP	WH-TO-VP

Table 10: Acquisition performances

Methods	Precision (%)	Recall (%)	F-measure(%)
Baseline	75.38	65.20	69.92
Our	87.60	81.35	84.36

sentence pairs for a useful corpus. The results of 247,471 sentence pairs were estimated of a precision ratio of 81.3% and a recall ratio of 71.04%.

We modified the filtering method in Table 2 a little to adjust it to bilingual SCF acquisition. We used the Chinese SCF diathesis alternations as heuristic information and the algorithm may be written as follows. For hypotheses with an English SCF $escf_i$,

1. if $p(escf_i, cscf_i) > \theta_1$, accept the hypothesis into set S;
2. Else if $p(escf_i, cscf_i) > \theta_2$ and $p(cscf_i|cscf_j) > 0$ and $(escf_i, cscf_j) \in S$, accept the hypothesis into set S;
3. Go to step 1 until S doesn't increase.

Here, $(escf_i, cscf_i)$ is a bilingual SCF hypothesis and $p(cscf_i|cscf_j) > 0$ means that $cscf_i$ is a diathesis alternative of $cscf_j$.

For the experiment, we empirically set θ_1 to be 0.003 and θ_2 0.0001. Table 10 lists the performances of our filtering method and a baseline MLE method with a single threshold of 0.0005. We can see that our method outperformed the baseline a lot. The precision rate was improved by 12.22%, the recall rate by 16.15% and F-measure by 14.44%.

At last, we totally acquired 654 bilingual SCF types for Chinese and English predicative verbs. Against the background corpus, these subcategorization frames are syntactically compatible with probabilities from 0.0001-0.0746.

SIMPLE APPLICATIONS ON CONCRETE NLP TASKS

In order to further analyse, the practicability of the previously acquired lexicons, we performed some simple experiments with subcategorization information as heuristics. These experiments are actually task-oriented evaluation of our automatically constructed SCF lexicons.

On Chinese PCFG parsing: We applied the acquired Chinese SCF lexicon in a PCFG parser helping to choose from the n-best parsing results. The concerned parser was trained from 10,000 manually parsed Chinese sentences (Zhao, 2001). In this experiment, there are 664 verbs and their SCF information involved. The open testing set

Table 11: Parsing evaluation

Parsing methods	Phrase-based Precision (%)	Phrase-based Recall (%)	Sentence-based Precision=Recall (%)
One-best	57.50	55.00	13.64
5-best	65.28	64.59	26.20
With SCF	62.86	62.10	21.66

Table 12: SMT performances with weighted corpora

Corpus size	λ	BLEU	NIST	OOV
400,000	0	0.4001	6.9082	23
600,000	0.5	0.4138	7.0796	18
600,000	0.67	0.4259	7.2997	26
600,000	0.8	0.4243	7.2706	39
200,000	1	0.3835	7.0982	47

consists of 1,500 sentences, for each of which the PCFG parser outputs 5-best parsing results. Then SCF hypotheses are generated for each result by means of the mentioned technology in the study. Finally, the maximum likelihood between hypotheses and those SCF types for the related verb in the lexicon is calculated in the following way.

$$ML = Max_{i,j} \frac{|\{Arguments\_in\_h_i\} \cap \{Arguments\_in\_scf_j\}|}{|\{Arguments\_in\_h_i\} \cup \{Arguments\_in\_scf_j\}|} \quad (7)$$

Here, $i \leq 5$, h_i is one of the hypotheses generated for the parsing results and scf_j is the jth SCF type for the concerned verb. This calculation keeps the likelihood between 0 and 1. The parsing result with maximum likelihood is then regarded as the final choice. When two or more hypotheses hold the same likelihood, the one with larger or largest PCFG probability will be chosen. Table 11 shows the phrase-based and sentence-based evaluation results for the parser without and with SCF heuristic information. There are three cases included: The output is one-best, the output is 5-best and the best evaluation result is recorded and the 5-best output is checked again for the best syntactic tree by means of SCF information. The phrased-based evaluation follows the popular method for evaluating a parser, while the sentence-based depends on the intersection of the parsed trees and those in the gold standard. Since, the PCFG parser output at least one syntactic tree for every sentence in our testing corpus, the sentence-based precision and recall are equal to each other.

It shows that SCF information remarkably improved the performance of the PCFG parser: the phrase-based precision increased by 5.36% and recall by 7.1%, while the sentence-based precision and recall both increased by 8.04%. However, this doesn't reach the upper limit of the 5-best. The possible reasons are: the our present SCF lexicon remains to be improved and our method of applying SCF information to the parser is too simple, e.g., probabilities of PCFG parsing results haven't been exploited thoroughly.

On statistical machine translation: We applied bilingual subcategorization information as heuristics to training corpus weighting for statistical machine translation. Our training corpus includes 200,000 Chinese-English sentence pairs, all which strongly hold some kind of cross-lingual subcategorization relations according to the experiment in the study and 400,000 pairs without such relations. Our evaluation corpus includes 506 Chinese sentences and 16 reference English sentences for each Chinese one. We used the off-the-shelf phrase-based decoder Pharaoh. The phrase bilingual lexicon is derived from the intersection of bi-directional IBM Model 4 alignments, obtained with GIZA++. The language model of ISI ReWrite Decoder is trained with the CMU-cambridge statistical language modeling toolkit v.2. Given an SMT training data set X, which includes n bilingual sentence pairs, i.e. the input vector $X = \{x_1, x_2, Y, x_i, Y, x_n\}$. During the original training process, every sentence pair x_i contributes in the same way to the estimation of parameters in the translation model since the corpus has not been weighted. Now, we tried to adjust the contribution of x_i according to its bilingual subcategorization. If we set the weight vector to be $W = \{w_1, w_2, Y, w_i, Y, w_n\}^T$, the weighted corpus would become $X' = WX = \{w_1x_1, w_2x_2, Y, w_ix_i, Y, w_n x_n\}$, where, $w_i = \lambda$ when x_i holds cross-lingual subcategori-zation, or 1 B λ when otherwise. And λ is an empirical weighting parameter satisfying that $0 \leq \lambda \leq 1$. The module of GIZA++ was also modified to ensure that the weight imposed on sentence pairs could be effectively transmitted to smaller translation units. We trained five SMT models of different weights. Table 12 lists both the training parameters and relevant decoding performances of the five models. We can see that SMT model achieved the best performance when λ was set to be 0.67.

CONCLUSION

This study describes, our techniques and experiments on subcategorization acquisition and relevant lexicon construction for English and Chinese verbs. We have made some achievements on the improvement of English subcategorization, Chinese subcategorization lexicon construction and English-Chinese cross-lingual subcategorization acquisition. However, there still remains a lot for improvement and adjustment and approaches that are more complicated still exist theoretically. For instance, English diathesis alternations might as well promote the

bilingual acquisition performance to a certain degree, bilingual SCF types for individual pairs of parallel verbs are yet to be acquired and some SCF types unseen by the hypothesis generator might be recalled by integrating semantic verb-classification information into the system.

More essential aspects of our future research will focus on improving the performance of hypothesis generators for both single language and cross-lingual subcategorization and testing and applying the acquired subcategorization lexicon in some other concrete NLP tasks.

ACKNOWLEDGEMENTS

This research was supported by the National Natural Science Foundation of China (grant no. 60773069 and 60873169).

REFERENCES

Baker, M., 2000. In other words: A Coursebook on Translaton. Foreign Language Teaching and Research Press, Beijing, pp: 10-78. ISBN: 7-5600-1919-6. DOI: 34971.

Brent, M.R., 1991. Automatic acquisition of subcategorization frames from untagged text. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL), Berkeley, CA., pp: 209-214. http://citeseer.ist.psu.edu/ushioda93automatic.html.

Briscoe, T. and J. Carroll, 1997. Automatic extraction of subcategorization from corpora. In: Proceedings of the 5th ACL Conference on Applied Natural Language Processing, Washington, DC. http://acl.ldc.upenn.edu/A/A97/A97-1052.pdf.

Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, ISBN: 0262530074.

Ellis, R., 2000. Second language Acquisition. Shanghai Foreign Language Education Press, Hong Kong, pp: 3-14. ISBN 7-81046-794-8. DOI: 14564.

Gamallo, P., A. Agustini and L.P. Gabriel, 2002. Using co-composition for acquiring syntactic and semantic subcategorisation. Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX), Philadelphia, pp: 34-41. http://citeseer.ist.psu.edu/article/gamallo02using.html.

Hu, Y. *et al*., 1995. Research on Chinese Verb. Beijing Language College Press, Beijing, pp: 119-177. ISBN: 7-81041-118-7 (in Chinese).

Jin, G., 2001. Semantic computations for modern chinese verbs. Beijing University Press, Beijing. pp: 44-122. ISBN: 7-301-04993-5. DOI: 032127 (in Chinese).

Korhonen, A., 2001. Subcategorization Acquisition. Dissertation for Ph. D., Trinity Hall University of Cambridge. http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-530.html.

Korhonen, A., 2002. Subcategorization Acquisition. Technical Report Number 530, Trinity Hall University of Cambridge. http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-530.html.

Korhonen, A., Y. Krymolowski and Z. Marx, 2003. Clustering Polysemic Subcategorization Frame Distributions Semantically. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp: 64-71. http://www.cl.cam.ac.uk/~alk23/ACL-clustering.pdf.

McCarthy, D., 2001, Lexical Acquisition at the syntax-semantics interface: Diathesis alternations, subcategorization frames and selectional preferences. Ph. D. thesis, University of Sussex. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.9104.

Sarkar, A. and D. Zeman, 2000. Automatic extraction of subcategorization frames for czech. Proceedings of the 19th international conference on computational linguistics, saarbrucken, Germany. http://www.sfu.ca/~anoop/papers/pdf/coling0final. pdf.

Shulte im Walde, Sabine, 2002. Inducing German semantic verb classes from purely syntactic subcategorization information. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp: 223-230. http://www.aclweb.org/anthology-new/P/P02/P02-1029.pdf.

Zhang, W., 1999. A Semantic study of the unique syntactic structures in Chinese. Beijing Language College Press, Beijing, pp: 1-182. ISBN 7-5619-0750-8. DOI: 28944 (in Chinese).

Zhao, T., 2001. Knowledge Engineering Report for MTS2000. Machine Translation Laboratory, Harbin Institute of Technology. Harbin. http://mitlab.hit.edu.cn/index.php/2008-04-24-06-13-13/25-the-project/ker.pdf.