Syntactic Parsing of Clause Constituents for Statistical Machine Translation

Jianjun Ma

School of Foreign Languages,
Dalian University of Technology, Dalian, China
E-mail: majian@dlut.edu.cn

Jiahuan Pei, Degen Huang, Dingxin Song

School of Computer Science and Technology,
Dalian University of Technology, Dalian, China
E-mail: p_sunrise@mail.dlut.edu.cn

Abstract: Clause is considered as the basic unit of grammar in linguistics, which is a structure between a chunk and a sentence. Clause constituents, therefore, are one important kind of linguistically valid syntactic phrases. This paper adopts the CRFs model to recognize English clause constituents with their syntactic functions, and testifies their effect on machine translation by applying this syntactic information to an English-Chinese PBSMT system, evaluated on a corpus of business domain. Clause constituents are mainly classified into six kinds: subject, predicator, complement, adjunct, residues of predicator, and residues of complement. Results show that our rich-feature CRFs model achieves an F-measure of 93.31%, a precision of 93.26%, and a recall of 93.04%. This syntactic knowledge in the source language is further combined with the NiuTrans phrasal SMT system, which slightly improves the English-Chinese translation accuracy.

Keywords: syntactic parsing; clause constituents; PBSMT

Reference to this paper should be made as follows: Ma, J.J, Pei, J.H, Huang, D.G. and Song, D.X. (2016) 'Syntactic Parsing of Clause Constituents for Statistical Machine Translation', *Int. J. Computational Science and Engineering*, Vol. X, No. Y4, pp.000–000.

Biographical notes: Jianjun Ma, professor of Dalian University of Technology. Her current research interests focus on chunking and machine translation.

Jiahuan Pei is a PhD. candidate student of Dalian University of Technology. Her main research interests include sentence similarity computation, temporal intent understanding, keyword extraction and machine translation.

Degen Huang, professor of Dalian University of Technology. His major research interests include machine translation, information retrieval and natural language processing.

Dingxin Song is a PhD. candidate student of Dalian University of Technology. His major research interests include machine translation.

1 Introduction

Clause, a structure between a chunk and a sentence, is considered as the basic unit of grammar in linguistics according to Halliday (2008). A clause is a sequence of words in a sentence that contains a subject and a predicate, as in

(CL1 A text must be segmented into clauses) before (CL2 the detailed functional annotation (CL3 the theory describes) can be applied).

The tasks of syntactic parsing of clause constituents can be classified into three kinds: identifying the types of constituents such as NP or VP, the dependency relationship between constituents and the syntactic functions of constituents like subject or predicate. Syntactic information of the first two tasks has been combined with phrase-based Statistical Machine Translation (PBSMT). One simple means of incorporating syntax into SMT is by re-ranking the n-best list of a baseline SMT system using various syntactic models, but Och et al. (2004) have found very little positive impact on this approach. However, Quirk et al. (2005) exploit dependency information of the source language in a phrasal translation model, which produces better outputs than phrase-based systems when evaluated on relatively small scale, domain specific corpora. Li et al. (2013) introduce linguistically motivated constraints like NP and VP into a hierarchical model, which improves Chinese-to-English translation accuracy. As to the task of identifying

phrases with their syntactic functions, research is mainly restricted to the Chinese language, referred to as Chinese functional phrase chunking in Zhou and Zhao (2007); Liu and Huang (2011). Actually, like chunks with semantic information such as terms (Zhang et al., 2016) and topics (Hashimoto et al., 2012; Sainani et al., 2012), the clause constituents identified by their syntactic functions are also linguistically valid chunks, for what really matters, in human translation and machine translation as well, particularly in the step of reordering, is whether it is a subject or a predicator or an adjunct, instead of a noun phrase (NP) or a verb phrase (VP). Therefore, this paper aims to identify English clause constituents with their syntactic functions, and incorporate this syntactic knowledge in the source language with a phrasal SMT system. The main purpose is to testify whether this kind of linguistically motivated syntactic information can improve the English-Chinese machine translation.

Statistical methods, especially machine learning approaches, are found to be universally applicable parsing methods. Kim et al. (2002) use Maximum Entropy (ME) model to determine segmentation positions and simplify parsing. Molina and Pla (2002) tackle shallow parsing work as tagging problems and apply a specialized Hidden Markov Model (HMM) to the process. Support Vector Machines (SVMs) is also introduced to the parsing tasks via supervised learning like Pradhan et al. (2004) and semi-supervised learning like Kate and Mooney (2007). Conditional Random Fields (CRFs) is another widely used system, and Sha and Pereira (2003) report that CRFs offers advantages over HMMs and ME model. Moreover, some fusion methods have been put forward, which combine linguistic rules with multiple statistical models. Ram and Devi (2008) identify clause boundaries using CRFs model and add linguistic rules to its features. Marinčič and Šef (2012) decompose complex parsing problems to simple ones using types of machine learning methods and heuristic rules. And Fan et al. (2014) present a Hidden Markov Support Vector Machines (HM-SVMs) approach combining SVMs and HMMs for chunking.

This paper, therefore, adopts the CRFs model to recognize English clause constituents with their syntactic functions, and testifies their effect on machine translation by applying this syntactic information to an English-Chinese PBSMT system, evaluated on a corpus of business domain.

2 Definition of Clause Constituents

According to Functional Grammar proposed by Halliday (2008), clause constituents can be divided into five kinds in terms of syntactic functions: subject (S), finite (F), predicator (P), complement(C) and adjunct (D). In this paper, for the convenience of translation, some changes are made as follows:

(1) We mix finite and predicator as Predicator (P).

Eg1: *[S We/PRP] [P would/MD appreciate/VB] it/IT if/INC [S you/PRP] [P could/MD send/VB] [C1 us/PRP] [C2 a/DT list/NN of/INP your/PRP$ merchandise/NN and/CC a/DT price/NN list/NN] ./.*

(2) We introduce function tags C1/C2/C3/C4, in case a clause contains two or more than two complements. Usually 4 complements will be the most.

Eg2: *[S We/PRP] [P make/VBP] [C1 you/PRP] [C2 the/DT offer/NN] [D subject/JJ to/INP your/PRP$ reply/NN] [P reaching/VBG] [C us/PRP] [D not/RB later/RBR than/INP noon/NN December/NNP 23/CD] ./.*

(3) A new function tag CR is introduced, which is the short form for the residues of complement.

Eg3: *[S We/PRP] [P have/VBP received/VBN] [C your/PRP$ letter/NN] [P dated/VBN] [D 19/CD August/NNP] [CR in/INP connection/NN with/INP the/DT above/JJ subject/NN] ./.*

(4) Similarly, PR, the residues of predicator, is another new tag. In a verb + noun fixed phrase, the noun phrase should be considered as a part of the predicator, and translated together with the verb.

Eg4: *Therefore/RB, we/PRP request/VBP you/PRP [P do/VBP] [PR your/PRP$ utmost/JJ] to/TO send/VB the/DT overdue/JJ goods/NNS without/INP any/DT delay/NN.*

Altogether, there are 7 types of the syntactic functions: S, P, C, D, C1/C2/C3/C4, CR, PR, as is shown in Table 1.

Table 1 Syntactic Functions

Function types	English meaning
S	Subject
P	Predicator
C	Complement
D	Adjunct
C1/C2/C3/C4	The 1st/2nd/3rd/4th complement
CR	The residues of complement
PR	The residues of predicator

3 Chunking Method

3.1 Conditional Random Fields (CRFs) Modeling

CRFs are undirected graphical models (also known as random fields) used to calculate the conditional probability of values on designated output nodes given values assigned to other designated input nodes. As McCallum (2002) stated, CRFs have achieved empirical success recently in POS tagging, noun phrase segmentation and table extraction from government reports. CRFs introduced here are the simple linear-chain CRFs, which can solve the problem of sequence labeling and compose different features using feature templates more flexibly.

We tackle the identification task as a sequence labeling problem, and model the process by linear-chain CRFs as follows:

Let $Y=y_1, y_2, \cdots, y_T$ be a finite state sequence, which refers to the boundary-type labels of a clause constituent, $X = x_1, x_2, \cdots, x_T$ an observation sequence of words, $\Lambda = \lambda_1, \lambda_2, \cdots, \lambda_n$ the parameters of the model we should learn in the training process, f_k a feature function and $Z(X)$ an input-dependent normalization factor. Then a linear-chain CRFs is a distribution $P_\Lambda(Y|X)$, which can be written as Eq.1:

$$P_{\Lambda}(Y|X) = \frac{1}{Z(X)} exp\left\{\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(y_t, y_{t-1}, X_t)\right\} \quad (1)$$

Our object is to get the parameters Λ, which maximizes the objective function as Eq.2:

$$Y^* = \arg\max P_{\Lambda}(Y \mid X) \quad (2)$$

More detailed information about linear-chain CRFs model, please see introduction in Sutton and McCallum (2011), and the open source toolkit CRF++ (http://taku910.github.io/crfpp/) is used to implement our recognition task.

3.2 Feature Engineering

Feature engineering is an essential work for designing a good set of feature functions f_k for CRFs models. Too many features may lead to data redundancy and conflict; therefore, feature selection and template tailoring are the main points that influence results. Firstly, we choose sixteen candidate features as Table 2 shows and make values of these features for training and testing. Then, we customize some templates to select candidate features and extract the combined features. The basic template is released at CRF++ toolkit website and other templates form by adding some new features to the basic one.

To filter bad features from candidate set, we explore greedy algorithm by the method of sequential backward selection. In our experiment, our final feature set contains four elements: the current word w_i, the POS tag of current word t_i, the stem of the current word s_i, and the morpheme of the current word m_i.

Table 2 Original Candidate Features of CRFs

No.	Feature	Definition
0	w_i	current word
1	pi	POS tag of w_i
2	s_i	stem of w_i
3	m_i	morpheme of w_i
4	l_i	length class of w_i
5	f_i	frequency class of w_i
6	v_i	generalization of w_i with vowel
7	ty_i	generalization of w_i with word form
8	tya_i	replace capital and lowercase with X and x respectively
9	$pr2_i$	first two letters of w_i
10	$po2_i$	last two letters of w_i
11	$pr3_i$	first three letters of w_i
12	$po3_i$	last three letters of w_i
13	$pr4_i$	first four letters of w_i
14	$po4_i$	last four letters of w_i
15	isd_i	w_i is digit or not

Based on the features selected above, we make the template "step forward" to adjust window-size from 3 to 5, which is usually no more than 5, and then get the best window-size according to the results. In this paper, the best window-size turns out to be 4 and Table 3 shows the best template used in the experiments.

3.3 Parameter Estimation

In our experiments, we estimate parameter *a* and *c* in CRFs model to improve the performance of recognition.

- *-a*

For parameter *a*, we choose L2 regularization algorithm, which performs slightly better than L1 in our tests.

- *-c*

Value of *c* decides the hyper-parameter for the model, which significantly influences the performance. Therefore, we conduct 100 groups of experiments based on different *c* values from 0.1 to 10 and find an optimal value through five cross validation tests.

Table 3 Feature Templates of CRFs

No.	Features
0	w_i, t_i, m_i, s_i; $i\epsilon$[-3,3]
1	w_i w_{i+1}, m_i m_{i+1}, s_i s_{i+1}; $i\epsilon$[-2,1]
2	t_i t_{i+1}; $i\epsilon$[-3,2]
3	t_i t_{i+1} t_{i+2}; $i\epsilon$[-3,1]
4	t_i t_{i+2}; $i\epsilon$[-1,0]
5	$w_{-1}t_0$, w_1t_0
6	$t_{-1}t_0w_i$; $i\epsilon$[-2,1]
7	$t_1t_2t_i$; $i\epsilon$[-1,0]

4 Experiments and Results

4.1 Evaluation Metrics

The performance of the system is evaluated in terms of precision (P), recall (R) and F-measure (F), which are the standard measures for the chunk recognition. The equations are shown as Eq. 3, Eq.4 and Eq.5:

$$P = \frac{number\ of\ correct\ clauses}{number\ of\ clauses\ in\ the\ system\ output} \quad (3)$$

$$R = \frac{number\ of\ correct\ clauses}{number\ of\ clauses\ in\ the\ test\ corpra} \quad (4)$$

$$F_{\beta} = \frac{2*P*R}{P+R} \quad (5)$$

In our experiments, only when both the IOB tags and the order of IOB tags are correct, will the chunks are considered valid.

4.2 Experiment Corpus

An English-Chinese corpus in business domain with 10,059 sentence pairs, 200,000 English words, and 270,000 Chinese characters, is used as experiment corpus in this paper. In order to have a reliable result, five-fold cross tests are performed. The corpus is evenly divided into five groups, with each group equally covering all the business situations. In each test, one group is chosen as the testing data and the rest four are combined as the training data.

4.3 Results of Clause Constituent Recognition

The Precision, Recall, and F-measure of the five tests are shown in Table 4, and the Average and the Standard Deviation are calculated.

According to Table 4, Approach No.4 achieves the best result in this paper with a F-measure of 93.31% and a precision of 93.26%, which indicates that the system has a good performance, robustness and stability. Approach No. 2 has achieved a precision of 93.26 and F-measure of 92.98,

which performs 2.37% and 1.1% higher than the baseline. The approaches No.3-5 show the results of different window-size based on the best features in our experiments, which denotes that window-size four is the best choice.

Table 4 Overall Open Test Results

No	Approach	P (%)	R(%)	F(%)
1	Baseline: CRFs+ basic features	90.89±0.50	92.90±0.50	91.88±0.51
2	Baseline +func tags	93.26±0.35	92.71±0.48	92.98±0.42
3	CRFs+best features(win3)+ func tags	93.40±0.58	92.92±0.67	93.16±0.62
4	CRFs+best features(win4)+ func tags	93.57±0.54	93.04±0.62	93.31±0.57
5	CRFs+best features(win5)+ func tags	93.54±0.69	92.97±0.68	93.26±0.68

Then, we evaluate test results of specific function types of clause constituents and the average results are shown in Table 5.

Table 5 F-measure of Specific Function Types of Clause Constituents

	C	C1	C2	C3	CR	D	P	PR	S
Baseline	85.03	91.13	81.52	26.09	4.17	84.13	98.09	77.60	97.42
Best result	87.48	91.89	84.00	24.66	4.26	86.04	97.61	83.16	97.99

The results indicate that as to the identification of specific function types, the system works extremely well in identifying P and S, with F-measure over 97 percent. The second-best identified functional type is C1, with F-measure of 91.89 percent. The identification results of D, C, C2, CR, PR are comparatively low, especially the identification of C3 and CR, the Recall and F-measure are the lowest with only 24.66 percent and 4.26 percent.

Based on the best result, we estimate the parameter c and plot the precision, recall and F-measure as Figure 1.

Figure 1 Performance based on different c values

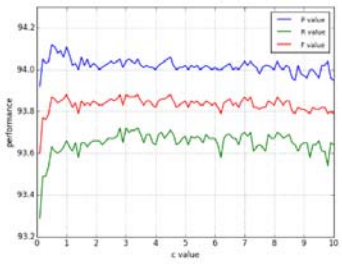


Figure 2 Learning curve of CRFs

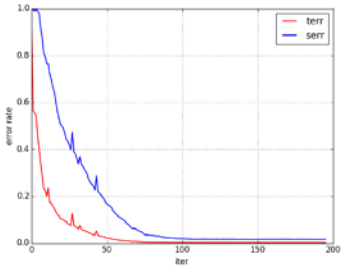


Figure 2 is the learning curve of our best CRFs model. When processing iterations, the error rate of tags and sentences tends to be convergent.

5 Application to PBSMT

The identified syntactic information of the source language is exploited in a phrased-based statistical machine translation (PBSMT) system in order to improve the translation performance. Sun et al. (2015) reports their best increase of 0.64 BLEU score of NiuTrans SMT system (Xiao et al., 2012) by adding a new bilingual syntactic phrase feature to translation phrase table. Inspired by their work, we construct a baseline PBSMT system using NiuTrans, evaluated on our corpora, and then we add our syntactic features to improve the baseline system. Three features are designed as follows:

5.1 Maximum Similarity Score (MSS)

MSS is a feature to measure the similarity between the phrases generated by SMT system and the clause constituents recognized by our model. Compared with the simple feature "has" or "has not" designed by previous work, our MSS is more general. Firstly, we extract a dictionary of clause constituents from the output and name it as OUR_CLAUSE_DICT and the translation phrase table generated from our baseline system as NIU_PHRASE_TABLE. Second, in order to reduce the unnecessary calculation of similarity, we construct indexes for OUR_CLAUSE_DICT and NIU_PHRASE_TABLE respectively by their first English word. Then, we design an algorithm to calculate the value of MSS (See Table 6).

Table 6 MSS Algorithm

```
def MSS
    for en_phrase in NIU_PHRASE_TABLE
    start_id, end_id = get_first_word(en_phrase)
    mss = 0.0
    for our_clause in OUR_CLAUSE_DICT [start_id:end_id]
       sim = SIM(en_phrase, our_clause)
          if sim>mss:
             mss =  sim
    if mss < theta:                          # we set theta =0.7
       mss = 0.0
    return mss

def SIM(str1,str2)
    v1,v2 = convert en_phrase, our_clause to vectors
    return cos_distance(v1, v2)
```

5.2 Type of Clause Constituent (TOCC)

While MSS represents the probability of a phrase fragment in translation phrase table that can be regarded as a clause constituent, TOCC adds more grammar message to the phrase table and is useful for the SMT system in the reordering process.

5.3 Bilingual Length Similarity (BLS)

For an item of a translation phrase table, the length of English fragment and Chinese fragment may be unbalanced. That is, a short English fragment may be aligned to a quite long Chinese one. It may bring degradation in quality of

translation phrase table, and affect the decoding process in SMT.

We firstly design the BLS as the absolution distance between length ratio, see Eq.6.

$$lr_{len1,len2} = \left\| \frac{len1}{len2} - 1 \right\| \tag{6}$$

Then, to normalize Eq.6, we introduce the sigmoid function as Eq.7 and construct a linear function transformation to get final BLS measure, see Eq.8.

$$s(x) = \frac{1}{1 + exp\{-x\}} \tag{7}$$

$$\mathrm{BLS}_{len1,len2} = 2*\left(1 - s\left(lr_{len1,len2}\right)\right) = \frac{2exp\left\{\left\| \frac{len1}{len2} - 1 \right\|\right\}}{1 + exp\left\{\left\| \frac{len1}{len2} - 1 \right\|\right\}} \tag{8}$$

5.4 Results of Improvement for SMT

To testify the performance of our approach, we adopt NiuTrans SMT to build our baseline system and then we add our syntactic features to its translation phrase table to improve the system. The BLEU scores of our experiments are listed in Table 7. The best score is 10.67, which is slightly higher than that of the baseline (9.87).

Table 7 Translation Results

Groups	BLEU(%)
Baseline	9.87
+MSS	10.38
+MSS+TOCC	10.42
+MSS+TOCC+BLS	10.67

6 Conclusion

This paper adopts a rich-feature CRFs model to recognize English clause constituents with their syntactic functions, and testifies their effect on a phrasal SMT system, evaluated on a corpus of business domain. Clause constituents are mainly classified into six kinds: subject, predicator, complement, adjunct, residues of predicator, and residues of complement. Results show that our rich-feature CRFs model achieves an F-measure of 93.31%, a precision of 93.26%, and a recall of 93.04%. As to the specific function type, the system works best in identifying S and P, the F-scores of which both reach over 97 percent, while the identification of adjunct phrases (D) needs further improvement. The combination of this syntactic knowledge in the source language (English) with the NiuTrans phrasal SMT system proves a slight improvement on the English-Chinese translation accuracy, which shows that clause constituents, like noun phrases, can be one important kind of linguistically valid syntactic phrases for machine translation and other tasks of natural language processing.

Acknowledgements

This work was supported by Humanities and Social Science Research Projects in Ministry of Education, China (No.13YJAZH062).

References

Fan, S. X., Chen, L. D., Wang, X. and Tang, B. Z. (2014) 'Shallow parsing with Hidden Markov Support Vector Machines', *Proceedings of the 2014 International Conference on Machine Learning and Cybernetics*, Vol.2, pp.827–830.

Halliday, M.A.K. (2008) 'An Introduction to Functional Grammar', *Foreign Language Teaching and Research Press*, pp.106–158.

Hashimoto, T., Chakraborty, B. and Shirota, Y. (2012) 'Social media analysis – determining the number of topic clusters from buzz marketing site', *Int. J. Computational Science and Engineering,* Vol.7, No.1, pp.65-72.

Kate, R.J. and Mooney, R.J. (2007) 'Semi-supervised learning for semantic parsing using support vector machines', *Human Language Technologies 2007: the Conference of the North American Chapter of the Association for Computational Linguistics,* pp.81–84.

Kim, S.D., Zhang, B.T. and Kim, Y.T. (2000) 'Reducing parsing complexity by intra-sentence segmentation based on maximum entropy model', *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora,* pp.164–171.

Li, J., Resnik, P. and Daumé III, H. (2013) 'Modeling Syntactic and Semantic Structures in Hierarchical Phrase-based Translation', *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pp.540–549.

Liu, H. and Huang, D. (2011) 'Chinese Functional Chunk Parsing Employing CRF and Semantic Information', *Journal of Chinese Information Processing,* Vol.25, No.5, pp.53–59.

Marinčič, D., Šef, T. and Gams, M. (2012) 'Parsing With Clause and Intra-clausal Coordination Detection', *Computing and Informatics,* Vol. 31, No.2, pp.299–329.

McCallum, A. (2002) 'Efficiently inducing features of conditional random fields', *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence,* pp.403–410.

Molina, A. and Pla, F. (2002) 'Shallow parsing using specialized hmms', *The Journal of Machine Learning Research*, Vol.2, pp.595–613.

Och, F. J., Gildea, D., et al. (2004) 'A smorgasbord of features for statistical machine translation', *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* pp.161–168.

Pradhan, S.S., Ward, W., et al. (2004) 'Shallow Semantic Parsing using Support Vector Machines', *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* pp.233–240.

Quirk, C., Menezes, A. and Cherry, C. (2005) 'Dependency Treelet Translation: Syntactically- informed Phrasal SMT', *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics,* pp.271–279.

Ram, R.V.S. and Devi, S.L. (2003) 'Clause boundary identification using conditional random fields', *International Conference on Intelligent Text Processing and Computational Linguistics*, pp.140– 150.

Sainani, A., Krishna Reddy, P. and Maheshwari, S. (2012) 'Mining special features to improve the performance of e-commerce product selection and resume processing', *Int. J. Computational Science and Engineering*, Vol.7, No.1, pp.82-95.

Sha, F. and Pereira, F. (2003) 'Clause boundary identification using conditional random fields', *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp.134–141.

Sun, S., Ding, P. and Huang D. (2015) 'An Improved Syntactic Phrase Extraction Approach for Statistical Machine Translation', *Journal of Chinese Information Processing*, Vol. 29, No. 2, pp. 95–102.

Sutton, C. and McCallum, A. (2011) 'An introduction to conditional random fields', *Machine Learning*, pp.267–373.

Xiao, T., Zhu J.B., Zhang, H. and Li, Q. (2012). 'NiuTrans: An Open Source Toolkit for Phrase- based and Syntax-based Machine Translation', *Proceedings of ACL*, pp.19-24.

Zhang, S.X., Du, Y.Y. and Lu, K. (2016) 'A dynamic window split-based approach for extracting professional terms from Chinese courses', *Int. J. Computational Science and Engineering*, Vol.12, No.4, pp.341-351.

Zhou, Q. and Zhao, Y. Z. (2007) 'Automatic Parsing of Chinese Functional Chunks', *Journal of Chinese Information Processing*, Vol. 21, No. 5, pp.18–24.

Appendix: Equations, figures and tables

1. Equations:

$$P_{\Lambda}(Y|X)=\frac{1}{Z(X)}exp\left\{\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(y_t,y_{t-1},X_t)\right\} \quad (1)$$

$$Y^{*}=\arg\max P_{\Lambda}(Y\mid X) \quad (2)$$

$$P=\frac{number\ of\ correct\ clauses}{number\ of\ clauses\ in\ the\ system\ output} \quad (3)$$

$$R=\frac{number\ of\ correct\ clauses}{number\ of\ clauses\ in\ the\ test\ corpra} \quad (4)$$

$$F_{\beta}=\frac{2*P*R}{P+R} \quad (5)$$

$$lr_{len1,len2}=\left\|\frac{len1}{len2}-1\right\| \quad (6)$$

$$s(x)=\frac{1}{1+exp\{-x\}} \quad (7)$$

$$\mathrm{BLS}_{len1,len2}=2*\left(1-s\left(lr_{len1,len2}\right)\right)=\frac{2exp\left\{\left\|\frac{len1}{len2}-1\right\|\right\}}{1+exp\left\{\left\|\frac{len1}{len2}-1\right\|\right\}} \quad (8)$$

2. Figures:

Figure 1 Performance based on different c values

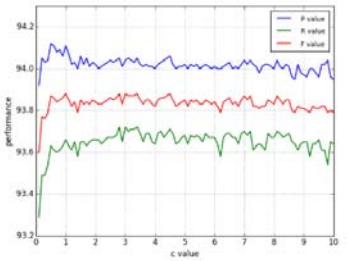
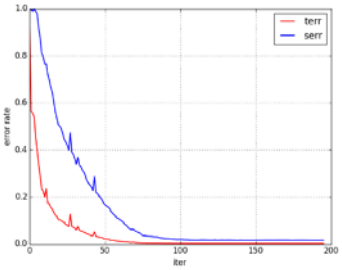


Figure 2 Learning curve of CRFs



3. Tables:

Table 1 Syntactic Functions

Function types	English meaning
S	Subject
P	Predicator
C	Complement
D	Adjunct
C1/C2/C3/C4	The 1st/2nd/3rd/4th complement
CR	The residues of complement
PR	The residues of predicator

Table 2 Original Candidate Features of CRFs

No.	Feature	Definition
0	w_i	current word
1	pi	POS tag of w_i
2	s_i	stem of w_i
3	m_i	morpheme of w_i
4	l_i	length class of w_i
5	f_i	frequency class of w_i
6	v_i	generalization of w_i with vowel
7	ty_i	generalization of w_i with word form
8	tya_i	replace capital and lowercase with X and x respectively
9	$pr2_i$	first two letters of w_i
10	$po2_i$	last two letters of w_i
11	$pr3_i$	first three letters of w_i
12	$po3_i$	last three letters of w_i
13	$pr4_i$	first four letters of w_i
14	$po4_i$	last four letters of w_i
15	isd_i	w_i is digit or not

Table 3 Feature Templates of CRFs

No.	Features
0	w_i, t_i, m_i, s_i; $i\epsilon[-3,3]$
1	$w_i\ w_{i+1}$, $m_i\ m_{i+1}$, $s_i\ s_{i+1}$; $i\epsilon[-2,1]$
2	$t_i\ t_{i+1}$; $i\epsilon[-3,2]$
3	$t_i\ t_{i+1}\ t_{i+2}$; $i\epsilon[-3,1]$
4	$t_i\ t_{i+2}$; $i\epsilon[-1,0]$
5	$w_{-1}t_0$, w_1t_0
6	$t_{-1}t_0w_i$; $i\epsilon[-2,1]$
7	$t_1t_2t_i$; $i\epsilon[-1,0]$

Table 4 Overall Open Test Results

No	Approach	P (%)	R(%)	F(%)
1	Baseline: CRFs+ basic features	90.89±0.50	92.90±0.50	91.88±0.51
2	Baseline +func tags	93.26±0.35	92.71±0.48	92.98±0.42
3	CRFs+best features(win3)+ func tags	93.40±0.58	92.92±0.67	93.16±0.62
4	CRFs+best features(win4)+ func tags	93.57±0.54	93.04±0.62	93.31±0.57
5	CRFs+best features(win5)+ func tags	93.54±0.69	92.97±0.68	93.26±0.68

Table 5 F-measure of Specific Function Types of Clause Constituents

	C	C1	C2	C3	CR	D	P	PR	S
Baseline	85.03	91.13	81.52	26.09	4.17	84.13	98.09	77.60	97.42
Best result	87.48	91.89	84.00	24.66	4.26	86.04	97.61	83.16	97.99

Table 6 MSS Algorithm

```
def MSS
    for en_phrase in NIU_PHRASE_TABLE
    start_id, end_id = get_first_word(en_phrase)
    mss = 0.0
    for our_clause in OUR_CLAUSE_DICT [start_id:end_id]
       sim = SIM(en_phrase, our_clause)
          if sim>mss:
              mss =  sim
    if mss < theta:                          # we set theta =0.7
       mss = 0.0
    return mss

def SIM(str1,str2)
    v1,v2 = convert en_phrase, our_clause to vectors
    return cos_distance(v1, v2)
```

Table 7 Translation Results

Groups	BLEU(%)
Baseline	9.87
+MSS	10.38
+MSS+TOCC	10.42
+MSS+TOCC+BLS	10.67