How Well Conditional Random Fields Can be Used in Novel Term Recognition *

Xing ZHANG[a], Yan Song[b], and Alex Chengyu Fang[b]

[a] Department of Chinese, Translation and Linguistics, City University of Hong Kong
Kowloon, Hong Kong SAR
zxing2@student.cityu.edu.hk

[b] Department of Chinese, Translation and Linguistics, City University of Hong Kong
Kowloon, Hong Kong SAR
{yansong, acfang} @cityu.edu.hk

Abstract. In this paper, we describe the construction of a machine learning framework that exploit syntactic information in the recognition of biomedical terms and present the limits of machine learning in generating a novel term candidate list. Conditional random fields (CRF), is used as the basis of this framework. We make an effort to find the appropriate use of syntactic information, including parent nodes, syntactic paths and term ratios under this machine learning framework. The experiment results show that CRF model can achieve good precision in term recognition if trained with known term list. However, with regard to discovering potential novel terms for terminology lexicon editors, CRF model fails to show good performance, if trained with known term list only to predict novel terms in testing corpus. Therefore, this result suggests that more semantic information may be needed to determine a word to be a novel term during a specific period.

Keywords: term recognition, novel term recognition, conditional random fields

1 Introduction

This paper explores the use of Conditional Random Fields (CRF) in novel term recognition. It investigates the recognition of medical terms using CRF model. A variety of methods have been used in term recognition, some are linguistics focused, some are statistically motivated, and a large part of approaches combine these two together. Recently, with the development of machine learning models, a lot of work has attempted to extract terms using machine learning methods.

The usual practice of machine learning depends on training data and a set of discriminating features. Then, machine learning systems use training data to "learn" features useful for term recognition. The widely used standard features for machine learning include orthographic features, POS tags, prefix, and suffix information (Krauthammer and Nenadic, 2004). However, few studies have tried to make use of dynamic linguistic features in respect of term usage in real text. As Zhang and Fang (2009) found out, syntactic functions can be used effectively in selecting and ranking term candidates, which means termhood can be captured by computing term ratios in syntactic paths.

* The work reported in this paper was supported in part by research grants from the City University of Hong Kong (Project No. 9610126, 7008002, 7002387 and 7002190).

Furthermore, as studies concerning novel terms are not so common, this paper considers how syntactic information integrated under a machine learning framework can be helpful in discovering novel terms. As early as in 1995, Justeson and Katz defined novel terminology as terms that are newly introduced and not yet widely established, or terms that are current only in more advanced or specialized literature than that with which the intended audience can be presumed to be familiar. Utsuro et al. (2006) specifically define novel terms to be technical terms that are not included in any of existing lexicons of technical terms of the domain. In MeSH1, there will be annual changes to its descriptors (terms). As quoted from their website, 'In biomedicine and related areas, new concepts are constantly emerging, old concepts are in a state of flux and terminology and usage are modified accordingly.' And 'in selecting the expressions to be used for a new MeSH descriptor, it is the usual practice to adopt the expression most commonly used by the authors writing in the English language.' Therefore, novel terms may not only refer to those new words that are newly created and specifically for some meaning in a certain domain, but also could be some known word whose meaning is changed from the common understanding to be specialized in some domain. In this paper, a group of systematic experiments are performed to explore how well syntactic functions, including parent nodes, syntactic paths and term ratios, can be used as features to recognize terms under CRF model.

2 Related Works

Different machine learning methods have been used in term identification and term recognition. Zheng et al. (2009) uses a CRF tool to train a template for term extraction. Six kinds of features are adopted in their template, including POS, semantic information, left information entropy, right information entropy, mutual information and TF/IDF. Most of these features are probability obtained by statistical formula. Takeuchi and Collier (2004) use Support Vector Machines to study the effects of training set size, feature sets, boundary identification and window size on biomedical entity extraction. The features they choose include surface word forms, POS tags, orthographic features and head-noun features.

Tsai et al. (2005) also adopt some linguistic features, orthographical features, context features, POS features, word shape features, prefix and suffix features, and dictionary features to CRF framework. On the GENIA 3.02 corpus, their system achieves an F-score of 78.4% for protein names, which is 2.8% higher than the next-best system.

As for relation between syntactic functions and term extraction, Zhang and Fang (2009) prove that there are certain kinds of syntactic behaviors of terms that indicate termhood. Their work employs term ratios in different syntactic paths to select and rank term candidates successfully. In their work, syntactic path refers to the path of syntactic functions of one NP. Specifically, it is defined as concatenation of elementary syntactic functions tagged by Survey Parser. Term ratios are then defined as the frequencies of term occurrences in each syntactic path over all term occurrence frequencies in all syntactic paths.

Utsuro et al. (2006) applied the technique of estimating domain specificity of a term to the task of discovering novel technical terms that are not included in any of existing lexicons of technical terms of the domain. They compare the number of candidates of novel technical terms collected from the Web, with those after excluding terms which do not share constituent nouns against the sample terms of the given set. Then on each domain 1,000 of those remaining candidates are randomly selected and their domain specificity is estimated by their proposed method. After manually judging the domain specificity of those 1,000 terms, they measure the precision of their method to be 75% and recall to be 80%.

With the objective to study how syntactic features can contribute novel term recognition, the current research will convert syntactic information into usable syntactic features for CRF framework at first. Exhaustive experiments were conducted to examine performances of such

[1] http://www.nlm.nih.gov/pubs/factsheets/mesh.html

features through 10 folds cross validation. Afterwards, a list of potential novel terms will be generated and evaluated against gold standard.

3 CRF Model

Conditional Random Field is an effective undirected graph learning framework first introduced in (Lafferty, et al., 2001), which has been successfully applied in many natural language processing tasks (Song et al, 2009a, 2009b, Zhao et al, 2006). The learning task for CRF is to maximize the formula

$$p(y \mid x) = \frac{1}{Z(x)} \prod_{(x,y)} \Phi(y, x; \lambda) \quad (1)$$

where *x* denotes the input samples of the training or testing data, *y* refers to the corresponding outputs, and λ is the parameter vector for weighting attached to the feature function Φ. *Z(x)* is the normalization factor over all output values. We use the linear chain form of CRF, in which x and y are the sequence texts and output labels respectively.

Similar to many other machine learning tasks, feature selection strongly affects the learning and prediction performance. In CRF learning, it is rather a hard job to keep the learning process effective and computationally feasible. We use the following feature templates as shown in Table 1. Moreover, the complexity of CRF learning is also affected by the tag set used for labeling. For the term extraction task we conduct in this paper, we use a binary tag set, {0, 1}, to identify whether a word is a term.

Table 1: Feature templates used in CRF based term tagging.

Description	Template
Word Unigrams	W_{-3}, W_{-2}, W_{-1}, W_0, W_{+1}, W_{+2}, W_{+3}
Word Bigrams	$W_{-2}W_{-1}$, $W_{-1}W_0$, W_0W_{+1}, $W_{+1}W_{+2}$
Word Jump Bigram	$W_{-1}W_{+1}$
POS tag	*O*
Syntactic Functions	*S*
Parent node	*F*
Single or Compound	*C*
Scale of Term Ratio	*L*
Syntactic Paths	*P*

Our implementation of sequence tagging process for term extraction uses the CRF++ package by Taku Kudo[2].

4 Experiments and Evaluation

4.1 Corpora Acquisition and Processing

For the purpose of studying behavior of novel terms in a special period, i.e. the year of 2009, the corpora used in this study are built up from MEDLINE[3] abstracts limited to the year of 2009 only. Among these abstracts, 20,020 abstracts are parsed by The Survey Parser (Fang, 1996) first. The Survey parser can produce detail syntactic information of each constituent. We

[2] http://chasen.org/~taku/software/CRF++/

[3] MEDLINE: (Medical Literature Analysis and Retrieval System Online) is the U.S. National Library of Medicine's (NLM) premier bibliographic database that contains over 16 million references to journal articles in life sciences with a concentration on biomedicine.

partitioned this corpus into 10 subsets, each consisting of 2002 abstracts. After parsing, each sentence will be changed into a parsing tree with tags indicating syntactic functions in both phrasal level and clausal level.

Three lists of medical terms were created from Medical Subject Headings beforehand as gold standard. The first term list collects terms from 1954 to 2009 and this list is released in 2010 (MeSH 2010). This MeSH term list consists of 602,436 terms and is referred as general term list in this study. The second term list used in this study is novel term list in 2009. This novel term list collects this special group of terms which are included into MeSH in the year of 2009 and are considered as novel terms of 2009 in current study. This novel term list contains 422 terms (New MeSH term in 2009). The third term list collects terms from 1954 to 2008 and contains 594,854 terms. This MeSH list was released in 2009, therefore referred as MeSH 2009.

For different experiments in this study, different training and testing sets are constructed. As for experiments on effects of different training sizes, different training sets are constructed. We are going to examine how the system performs when trained on corpora of different sizes while tested on the same corpus. We set the training corpora to testing corpora ratio to be ranged from 1 to 1 to 8 to 1. Basic statistics for training and testing corpora are shown in following table (Table 2). Training set 1 contains one subset of corpus, training set 2 contains two subsets of corpora and training set 3 contains subsets of corpora. The other training sets are constructed in the same fashion. Also, the testing set contains one subset of corpus. Therefore, the training to testing ratio of training set 1 over testing set is 1 to 1 and of training set 2 is 2 to 1, and of training set 3 is 3 to 1. Likewise, the other training sets have corresponding training and testing ratios. As a result, the training to testing corpora ratio of training set 1 to training set 8 is increased from 1 to 1 to 8 to 1.

Table 2: Basic statistics for training and testing sets for effects of training size.

Training set	# of Sentences	# of Features
1	13,587	2,431,756
2	27,636	4,225,600
3	41,614	5,806,370
4	55,661	7,248,040
5	69,710	8,614,474
6	83,888	9,926,254
7	97,731	11,144,854
8	111,923	12,350,598

However, for 10 folds cross validation test, there are ten sets of training and testing subsets. Each set has a different training subset and a different testing subset. Whereas, the ratio of training to testing corpora of each set is set as 9 to 1. The following table shows basic statistics of these ten sets (TT stands for Training and Testing set).

Table 3: Basic statistics of 10 subsets for cross validation.

	# of Sentences	# of Features
TT1	122,527	20,034,513
TT2	122,357	20,023,374
TT3	122,383	20,017,890
TT4	122,719	20,063,700
TT5	122,413	20,017,188
TT6	122,527	20,034,513
TT7	122,531	20,052,465
TT8	122,627	13,365,422
TT9	122,623	20,052,285
TT10	122,994	20,084,649

4.2 Conversion from Parsing Sentences to Matrixes

Feature selection is very important in machine learning systems. In this study, we use syntactic information, including POS tags, parent nodes, single or compound, syntactic paths and term ratios. After parsing by The Survey Parser, MEDLINE abstracts will be tagged and presented in syntactic trees. The next work is to convert each syntactic tree into a matrix that can be processed by CRF model, and each matrix is composed of 7 features (see Table 1).

For the feature 'Scale of Term Ratio', it means a scale from 1 to 3 to indicate three categories of term ratios. More specifically, term ratios of syntactic paths are obtained from training corpus under a separate system (Zhang and Fang, 2010). At first, the training corpora will be matched against the MeSH term list and term occurrence frequencies in each syntactic path will be calculated, and then the proportion of term occurrence frequencies in this syntactic path over all term occurrence frequencies is defined as term ratios.

Then in order to convert term ratios into an index that can be recognized by CRF, a scale from 1 to 3 is adopted instead. This index is based on the span between minimum term ratio and maximum term ratio calculated from all the sentences among training corpora. This span is divided into 3 scales, numbered as 1 (the weakest level), 2 (the middle level) and 3 (the strongest level). 1 means the term ratio of this syntactic path falls within the first scales, 2 means the middle scale and 3 the last.

In addition, the feature 'Single or Compound' indicates whether a token is one single word or part of a compound. More precisely, if it is part of a compound, which part it occurs in? Is the beginning of this compound (B), or the ending of it (E), or in the middle (I). As there are compounds consisting of more than 3 words, we tag the first word to be B, the last to be E, and the rest in the middle to be I. The reason to split compounds into single words is because that CRF framework requires the matrix for it handles single tokens in sequence.

4.3 Evaluation

For performance of the CRF framework, the evaluation is based on how well it automatically determines the term status of a word. When testing, the CRF framework will tag 1 to a word in a matrix if it determines the word to be a MeSH term, and 0 otherwise. Therefore, the precision of it is the ratio of the number of correctly determined terms to the number of terms it tags as MeSH terms, and Recall is the ratio of the number of correctly determined terms to the number of true MeSH terms in this testing corpora.

The evaluation will use the standard formula F-score, which is defined as F = (2PR)/ (P + R), where P denotes the precision and R denotes the recall.

4.3.1 Experiments on Training Size

The experiments in this round are conducted 8 times to look at effects from increasing the training corpora. The ratio of training corpora to testing corpora will be controlled from 1 to 1 to 8 to 1. The performance on each set is shown in Table 4. And in training corpora, tokens matched with MeSH 2010 will all be tagged terms; furthermore, for these matched with new MeSH term list (new MeSH term list in 2009), they are tagged as novel terms specifically.

Table 4: Performance on novel terms.

Training and Testing Ratio	Precision	Recall	F-score
1 to 1	0.917	0.048	0.091
2 to 1	0.911	0.286	0.435
3 to 1	0.908	0.286	0.435
4 to 1	0.906	0.286	0.434
5 to 1	0.905	0.429	0.582
6 to 1	0.904	0.476	0.624
7 to 1	0.903	0.524	0.663
8 to 1	0.903	0.524	0.663

From above tables, it can be noted that under this CRF framework, precisions for novel term extraction are all above 90%. These results are much better than aforementioned precision of 75% (Utsuro et al., 2006). Thus, it is fair to conclude that syntactic features as parent nodes, syntactic paths and term ratios can be used as effective features under CRF framework in respects of novel term recognition.

Moreover, we can find some general trends in following figure.

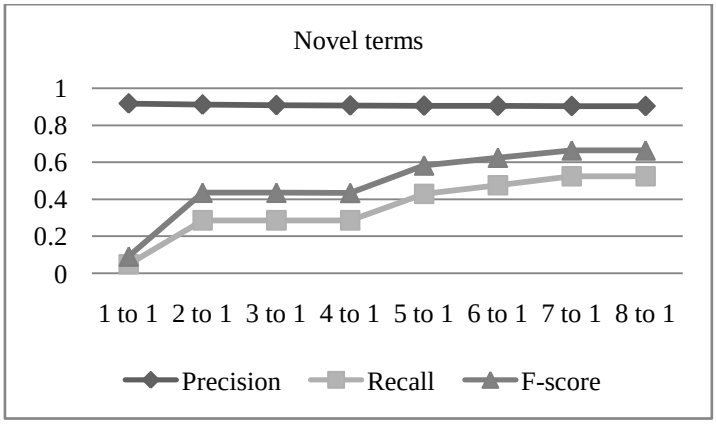

Figure 1: Performance on novel terms.

In Figure 1, which describes performance on novel terms, line describing precision is opposite to the other two lines of recall and F-score, which are in accordance with the graph of logarithmic function. It shows when training and testing ratio is set as 1 to 1, the precision is highest, while the recall and F-score are lowest. On the contrary, when training and testing ratio is increased to 8 to 1, the recall and F-score are highest while precision is lowest. Generally, the

precision of novel terms extraction is decreased consistently from the training and testing ratio of 1 to 1 to 8 to 1, while recall and F-score are increased consistently.

4.3.2 Experiments on 10-folds Cross Validation

The experiment for 10-folds cross validation is conducted 10 times on 10 sets of corpora. Each set has the same training to testing ratio of 9 to 1. The performances on each set are shown separately in following table (Table 5) and descriptive statistics of these performances are shown in table 6.

Table 5: Performance on novel terms of 10 subsets.

Novel terms	Precision	Recall	F-score
TT1	0.902	0.538	0.674
TT2	0.902	0.500	0.643
TT3	0.902	0.459	0.609
TT4	0.900	0.472	0.620
TT5	0.900	0.724	0.803
TT6	0.902	0.538	0.674
TT7	0.900	0.765	0.827
TT8	0.901	0.767	0.828
TT9	0.901	0.615	0.731
TT10	0.900	0.591	0.713
Average	0.901	0.598	0.712

Table 6: Descriptive statistics for novel terms.

Descriptive Statistics	Precision	Recall	F-score
Mean	0.901	0.597	0.712
SD	0.001	0.111	0.079
Minimum	0.900	0.459	0.609
Maximum	0.902	0.767	0.828

For the average performance of these 10 subsets (see Table 6), the highest recall among these 10 subsets only reaches 0.767 and the highest F-score is 0.828. However, we can see precisions across the 10 subsets are around 0.901 on average, which is a satisfying result of applying CRF model into recognizing novel terms, especially considering that only syntactic knowledge is used in this experiment.

However, in Figure 2 we can find greater fluctuations in lines of either recall or F-score for novel terms. This may indicate novel term recognition is greatly affected by sampling. The possible reason may be that novel terms are far sparser than general terms. They are scattered more broadly across testing corpora.

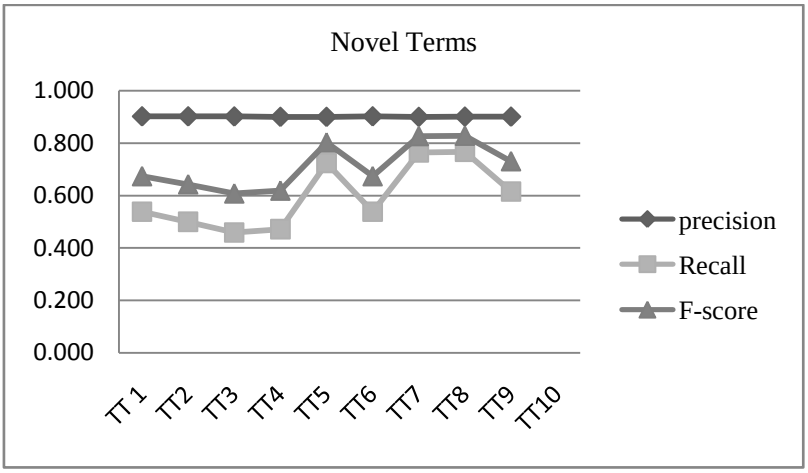

Figure 2: Performance on novel terms of 10 subsets.

4.3.3 Generation of Novel Term Candidates List

As introduced earlier, the final objective of this study is to generate a novel term candidate list for human experts to discover novel terms. Therefore, a novel term candidate list is generated. To evaluate this candidate list, the novel term MeSH term list in 2009 (New MeSH term in 2009) is used as gold standard in this regard. While for training, term list 'MeSH 2009' is used, which has no overlap with term list 'New MeSH term in 2009'.

After testing, CRF framework will have tagged 1 or 0 to a token in a matrix if it judges this token be MeSH term or not. Therefore, all these tokens tagged as 1 are terms recognized by CRF. However, for tokens labeled as B, I, or E, the situation is more complicated as these labels indicate there is originally a compound. For this reason, for a token labeled as either B, or I or E, and if it is also labeled as 1, this token will be combined with tokens immediately close to it to generate term candidates. Specifically, only tokens between the same group of B and E should be taken into consideration for combination. And furthermore, tokens within this group will only be combined in sequence. To illustrate this process, the combining principles are listed as following:

Rule 1: for token labeled as S, if it is labeled 1, add this token to term candidate list.

Rule 2 and Rule 3 are applied for token that is part of a compound.

Rule 2: for token labeled as B, or I, or E, if it is labeled 1, combine B+I, B+I+E and I+E, and then add these combinations to term candidate list.

Rule 3: for token labeled as B, or I, or E, if it is labeled 1, and if there are more token labeled as I, e.g. I_1 and I_2, combine I_1+I_2, $B+I_1+I_2$, $B+ I_1+I_2+E$, I_1+I_2+E, and add these combinations to term candidate list.

In addition, all these candidates will be matched against MeSH 2009 at first to check if this term is already included into MeSH 2009. If it is not, this term will be considered as novel term candidates in 2009 in a comparative sense

As we have testified that when ratio of training to testing corpora is set as 9 to 1, the recall of novel terms will be higher than other ratios. In this case, we choose training to testing corpora of the ratio 9 to 1 to conduct following experiment to generate novel candidate list. And the results are shown in Table 7.

The first column is the number of candidates that generated by above combining rules together. And the second column is the number of terms already included in MeSH 2009 among these term candidates. The third column is the number of terms that appear in new MeSH term list in 2009 and are considered as the novel terms predicted by current experiment.

As we can see from Table 7, there are only a small number of novel terms predicted by CRF model. Moreover, we have no confidence to say that this number could be increased if we increase the training corpora to be big enough, as currently the training corpus has already

around 122,000 sentences. The ability of CRF model to discover novel terms from large amount of corpora is not successful in this study.

Table 7: Novel terms discovered.

	# of Term Candidates	# of Terms of MeSH 2009	# of Novel Terms in 2009
TT1	33,821	1,989	2
TT2	40,258	2,134	3
TT3	39,568	2,054	2
TT4	42,105	2,510	2
TT5	38,005	1,878	1
TT6	39,456	1,989	2
TT7	41,243	2,136	2
TT8	41,903	2,445	1
TT9	40,537	2,100	3
TT10	39,879	1,943	2

5 Conclusion

Therefore, through above data analysis of the first and second experiment, we find increasing training size will help retrieve sparse novel terms, but the precision of novel term recognition will be impacted. This proves performance is influenced by sampling more greatly. Different sampling will have quite different recalls. The reason maybe as novel terms scattered sparsely in corpora, if some novel terms never appeared in training corpora, there is no chance that CRF model could learn its features and label it correspondingly. In such case, it would not be tagged as true term in testing corpora; therefore, this term would not be retrieved.

The third experiment explores to generate a novel term candidate list for human experts to judge before they decide which will be included into existing terminology lexicon. The results show that only few of novel terms are discovered by CRF model, though training corpora are quite large. This restriction may be caused by the fact that novel terms are quite sparse compared to existing terminology lexicon.

All in all, this research studies the performance of CRF framework on term recognition with the use of two kinds of unique syntactic information: syntactic paths and term ratios. On the basis of results from systematic experiments, the conclusion can be drawn that syntactic functions and syntactic paths can be used as fairly effective features under the CRF framework to recognize novel terms if trained with new term list beforehand. However, such syntactic features and the nature of CRF, fail to perform well to discover novel terms if only trained with known term list. Furthermore, the precision will be damaged when increasing the training corpora and the recall of novel terms remains unsatisfactory, which means that more distinguishing features are needed to improve the performance, like semantic features of potential novel terms. This finding proves there is room for future work of this study, such as, how to integrate semantic features into CRF model to help novel term extraction. Moreover, this work proves helpful for other machine learning based term extraction system in respect of exploiting effective syntactic features.

References

Chandra, Ashok K., Dexter C. Kozen, and Larry J.Stockmeyer. (1981). Alternation. *Journal of the Association for Computing Machinery*, 28(1):114-133.

Justeson, J. S., and Katz, S.M. (1995). Technical Terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1(1):9--27.

Krauthammer, M. and Nenadic, G. (2004). Term identification in the biomedical literature. *J Biomed Inform*, 37(6):512-526.

Lafferty, J. D., McCallum, A. and Pereira, F. C. N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pp 282-289, San Francisco, CA, USA, 2001.

National Library of Medicine. Fact sheet: medical subject headings (MeSH). [Web document]. Bethesda, MD: National Institutes of Health. 2010. [Last updated: 01 April 2010; cited 9 April 2000]. <http://www.nlm.nih.gov/pubs/factsheets/mesh.html>.

Song Y., Kit, C.Y., Xu, R.F., Zhao, H. How Unsupervised Learning Affects Character Tagging based Chinese Word Segmentation: A Quantitative Investigation, in *Proceedings of International Conference on Machine Learning and Cybernetics*, Jul, 2009.

Song Y., Kit, C.Y. PCFG parsing with CRF tagging for head recognition, in *Proceedings of CIPS-ParsEval-2009*, pp.133-137. Nov, 2009.

Takeuchi, K., and Collier, N. (2004). Bio-medical entity extraction using support vector machines, *Artificial Intelligence in Medicine*, Volume 33, Issue 2, Pages 125-137.

Tsai, T.H., Chou, W.C. and Wu, S.H.(2005). Integrating linguistic knowledge into a conditional random field framework to identify biomedical named entities. *Expert Systems Appl*. v30 i1. 117-128.

Utsuro, T., Kida, M., Tonoike, M., Sato, S. (2006). Collecting novel technical term from the Web by estimating domain specificity of a term. In: Matsumoto, Y., Sproat, R.W., Wong, K.-F., Zhang, M. (eds.) *ICCPOL 2006*. LNCS (LNAI), vol. 4285, pp. 173–180. Springer, Heidelberg.

Zheng, D, Zhao, T. and Yang, J. (2009). Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy, *Proceedings of 22nd International Conference*, ICCPOL 2009, Hong Kong, March 26-27.

Zhao H., Huang C.N., and Li M. An Improved Chinese Word Segmentation System with Conditional Random Field, *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing (SIGHAN-5)*, pp.162-165, Sydney, Australia, July 22-23, 2006.

Zhang, X. and Fang. A. C. (2010). An ATE System based on Probabilistic Relations between Terms and Syntactic Functions. In *10th International Conference on Statistical Analysis of Textual Data. Sapienza*, University of Rome (Italy), 9 to 11 June 2010.