CONDITIONAL RANDOM FIELDS FOR TERM EXTRACTION

Xing Zhang, Yan Song
Dialogue Systems Group, Department of Chinese, Translation and Linguistics, City University of Hong Kong
TatChee Avenue, Kowloon, Hong Kong SAR

Alex Chengyu Fang
Dialogue Systems Group, Department of Chinese, Translation and Linguistics, City University of Hong Kong
TatChee Avenue, Kowloon, Hong Kong SAR

Keywords: Term extraction, Conditional Random Fields, Syntactic function, Term ratio.

Abstract: In this paper, we describe how to construct a machine learning framework that utilizes syntactic information in extraction of biomedical terms. Conditional random fields (CRF), is used as the basis of this framework. We make an effort to find the appropriate use for syntactic information, including parent nodes, syntactic paths and term ratios under the machine learning framework. The experiment results show that syntactic paths and term ratios can improve precision of term extraction, including old terms and novel terms. However, the recall rate of novel terms still needs to be increased. This research serves as an example for constructing machine learning based term extraction systems that utilizes linguistic information.

1 INTRODUCTION

This paper investigates the use of Conditional Random Fields (CRF) in novel medical term extraction. A variety of methods have been used in term extraction, some are linguistics focused, some are base on statistically motivated, and recent approaches try to 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 widely used standard features for machine learning include orthographic features, POS tags, prefix, and suffix information. However, few studies have tried to make use of dynamic linguistic features in respect of term usage in real text. As Zhang and Fang (2010) 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.

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. In MeSH, 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.’ Therefore, novel terms may not only refer to those new words that are newly and specifically created for some meaning in a certain domain, but also could be some known word whose meaning is changed from common to special. In this study, systematic experiments are performed to explore how syntactic information can be used as effective features to extract 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 six kinds of features in their CRF template, including POS, semantic information, left information entropy, right information entropy,

Zhang X., Song Y. and Fang A..
CONDITIONAL RANDOM FIELDS FOR TERM EXTRACTION.
DOI: 10.5220/0003077304140417
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 414-417
ISBN: 978-989-8425-28-7

mutual information and TF/IDF. They test their method on military materials, the precision achieved was 79.63 %, recall was 73.54%, and F-measure was 76.46%. 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 GENIA 3.02 corpus, their system achieves an F-score of 78.4% for protein names.

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, 2009; Zhao et al, 2006). The learning task for CRF is to maximize the formula

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

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. Feature templates used are shown in Table 1.

Table 1: Feature templates in CRF based term tagging.

Description	Template
Word Unigrams	W-3, W-2, W-1, W0, W+1, W+2, W+3
Word Bigrams	W-2W-1, W-1W0, W0W+1, W+1W+2
Word Jump Bigram	W-1W+1
POS tag	O
Syntactic Functions	S
Parent node	F
Singe 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 (http://chasen.org/˜taku/software/CRF++/).

4 EXPERIMENTS

4.1 Resources Construction

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 abstracts (Medical Literature Analysis and Retrieval System Online by U.S. National Library of Medicine) limited to the year of 2009 only. Among these abstracts, 20,020 abstracts are sampled and parsed by Survey Parser (Fang, 1996) first. The Survey parser can produce detail syntactic information of each constituent. We partitioned this corpus into 10 subsets, each consisting of 2002 abstracts.

Two lists of medical terms were created from Medical Subject Headings beforehand as gold standard. The first 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. The second 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.

4.2 Feature Selection

Feature selection is very important in machine learning systems. In this study, 7 kinds of features are used, including POS tags, syntactic functions, parent nodes, single or compound, scale of term ratio, syntactic paths and term ratios. For the feature ‘Scale of Term Ratio’, it means a scale from 1 to 3 to indicate three categories of term ratio. And term ratios of syntactic paths are obtained from training corpus under a separate system (Zhang and Fang, 2010). At first, training corpora will be matched against the MeSH term list and term occurrence frequencies in each syntactic path will be calculated, and proportion of term occurrence frequencies in this syntactic path over all term occurrence frequencies is computed as term ratios.

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).

4.3 10-folds Cross Validation

The experiment uses 10-folds cross validation to test system performance. And the evaluation of CRF framework is based on how well it automatically determines the term status of a word. 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. Results of 10-folds cross validation, with respect to old terms and novel terms, are shown separately in following tables.

Table 2: Performance on general terms.

Old terms	Precision	Recall	F-score
1	0.992	0.936	0.963
2	0.993	0.943	0.967
3	0.992	0.939	0.965
4	0.993	0.945	0.968
5	0.993	0.945	0.968
6	0.992	0.936	0.963
7	0.993	0.943	0.967
8	0.992	0.941	0.966
9	0.992	0.941	0.966
10	0.993	0.943	0.967
Average	0.992	0.941	0.966

From Table 2, we can see average precision on general terms of the ten subsets is 0.992, which is quite good. On average recall is 0.941, which is also very good. The F-score for them is as good as 0.966. On the whole, this table shows performance on old terms across these ten subsets is consistently good as standard deviation of precision is 0.000, and of recall is 0.003 and of F-score is 0.002.

Table 3: Performance on novel terms.

Novel terms	Precision	Recall	F-score
1	0.902	0.538	0.674
2	0.902	0.500	0.643
3	0.902	0.459	0.609
4	0.900	0.472	0.620
5	0.900	0.724	0.803
6	0.902	0.538	0.674
7	0.900	0.765	0.827
8	0.901	0.767	0.828
9	0.901	0.615	0.731
10	0.900	0.591	0.713
Average	0.901	0.598	0.712

For 422 novel terms in 2009 (see Table 2), the performances are not as good as on old terms, but the highest recall among these 10 subsets still reaches 0.767 and the highest F-score is 0.828, which are fairly competitive compared to other reported results mentioned earlier.

However, we can see precisions across the 10 subsets are around 0.901 on average (Table 3), which is a satisfying result of applying CRF model into recognizing novel terms, especially considering that only syntactic knowledge is used in this experiment.

In following figure (Figure 1), as far as general terms are concerned, we can see the line for precision is rather smooth, and there are some changes among recalls and F-scores. Still, these changes are within a small range.

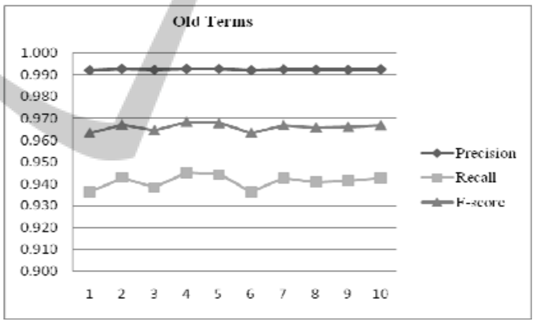

Figure 1: Performance on old terms.

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 much sparser than general terms. They are scattered more broadly across testing corpora.

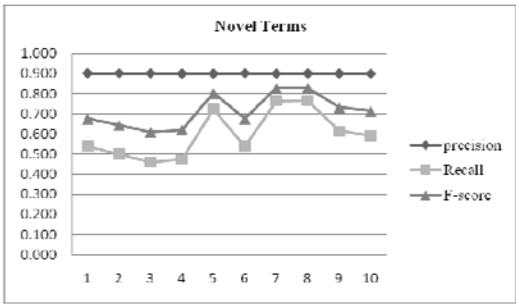

Figure 2: Performance on novel terms.

4 CONCLUSIONS

Through above experiments, we find performance on old term recognition is quite consistent across ten subsets of testing corpora. While for novel terms, the performance is influenced by sampling more greatly. Different sampling will have quite different recalls. The reason may be novel terms are scattered sparsely in corpora; and 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.

All in all, this research studies the performance of CRF framework on term extraction with the use of two kinds of unique syntactic information: syntactic paths and term ratios. The conclusion can be drawn that syntactic functions and syntactic paths can be used as effective features under the CRF framework. And the system performs quite well with respect to old term extraction. For novel term extraction, the precisions are also promising, though the recalls are quite low compared to old term extraction. This limitation indicates that more distinguishing features are needed to improve the performance, like semantic features of potential novel terms to help novel term extraction. And also, this work will be helpful for other machine learning based term extraction system in respect of exploiting effective syntactic features.

ACKNOWLEDGEMENTS

Research described in this article was supported in part by research grants from City University of Hong Kong (Project No. 7002387, 7008002, and 9610126).

REFERENCES

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.

Fang, A. C. (1996). The Survey Parser: Design and Development. In S. Greenbaum (Ed.), Comparing English World Wide: The International Corpus of English (pp. 142-160). *Oxford: Oxford University Press.*

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 200]. <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.

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.

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

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.