FINANCIAL NAMED ENTITY RECOGNITION BASED ON CONDITIONAL RANDOM FIELDS AND INFORMATION ENTROPY

SHUWEI WANG, RUIFENG XU*, BIN LIU, LIN GUI, YU ZHOU

Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School, Harbin Institute of Technology,
Shenzhen 518055, China
E-MAIL: xuruifeng@hitsz.edu.cn

Abstract:

Named entity recognitionplays an important role in many natural language processingtasks, such as relation detection and information extraction. This paper presents a novel method to recognize named entities infinancial news textsin three steps. First,the domain dictionary is applied to recognizestock names.Second, the full form FNEs are identified by incorporatinginternal featuresin a classifier based on Conditional Random Fields. Third, the mutual information, boundary entropy and context features are employed to recognize the abbreviation FNE candidates. The experiments completed on a Chinese financial dataset show that the proposed approach achieves 91.02% precision and 92.77% recall.

Keywords:

Named entities recognition; Financial named entity; Conditional Random Fields; Information Entropy

1. INTRODUCTION

With the development of the Internet, the network information has shown an explosive growth poor wording .Large quantities of financial texts are delivered everyday.Thereis valuable information in these texts. Automatic recognition of the named entities in financial area, such as the name of organization and company, areimportant for data mining applications and knowledge mining.

Named Entity Recognition (NER) labels sequences of words in a text, which are the names of things, such as person and company names. It comes with well-engineered feature extractors for Named Entity Recognition, and many options for defining feature extractors. The existing works on NER are generally camped into rule-based, statistic-based and hybrid approach. These approaches achieved good performance in many general tasks. However, in the financial domain, many existing approaches have the difficulties to extract the financial named entity (FNE) at high performance, especially for the abbreviation FNEs.

This paper presents an approach to recognize named entities from financial news texts by using both the internal and context featuresof the FNEs. Firstly, the domain dictionary is employed to match the stock names and some specific terms. A classifier based on Conditional Random Fields (CRFs) and internal features such as suffix, place names is then applied to identify the full form FNEs with relative regular forms. Next, the mutual information and boundary information entropy are appliedto measure the internal association and left/right neighboringhomogeneityof the candidate abbreviation FNEs, respectively. Moreover, a similarity estimation method based onLevenshtein algorithm is applied to identify the abbreviation FNEs which are similar to full form FNEs.

The approach is evaluated on a financial text dataset. The achieved good performance shows that this approach recognizes the financial named entities effectively, especially for the abbreviation FNEs.

The paper is organized as follows. Section 2 briefly reviews the related work on FNEs recognition. Section 3presents our approach. Section 4 gives the evaluations and discussion. Finally, Section 5 concludes.

2. RELATED WORKS

Generally speaking, previous solutions toNER can be broadly categorized into rule-based[1-4], statistics-based [5-7]and mixture of both [8,9].

In early stage of NER, rule-based approach was common way to identify NEs, which are mostly summarized by linguists through study on large corpus and NE libraries [1]. However, this method is time-consuming, expensive and inflexible. With the rapid development of the Internet, thesituation is becoming more serious. Therefore, it is very difficult to summarize simple buteffective rules on NE compositionsand contexts. Secondly, the recognition process in this approach cannot be activated until some "indicator" tokens are scanned in. These methods need lots of prior knowledge, which should be manually summarized.

Many statistical methods are applied to NER task. This

978-1-4799-4216-9/14/$31.00 ©2014 IEEE

approach achieved better performances in many cases. Wu et al. presented a hybrid algorithm whichcombineda statistical model and various types of human knowledge [10]. It applies *TONG YI CI CI LIN* (同义词词林) not sure you should use Chines in the English text – could be strange – drop itand a back-off model to avoid the data sparseness problem. On the newswire dataset, this method achieves0.7622F-value. Many researches regard the NER as a sequence labeling problem. There arealso some approaches taking NER as a classification problem. Classifiers,such as SVM, are usually applied. Chen et al. introduced a SVM-based method by using active learning strategy to select new instances and appending the labeled new instances into the training set[11]. The methods based on Hidden Markov Model[12], external knowledge (Wikipedia) [13] and kernels [14]were proposed in recent years, and showedtheirown advantages.

For the cases of financial named entity, because of its unusual expression way, the recognition rules are difficult to obtain.Wong et al. proposed a company name identification model[15]. Through making use of financial field related features, they built six knowledgebases and came up with a twice-scan strategy for implementing the system. The systemachieved97.3% precisionand89.3% recall on close testing, 62.8% precision and 62.1% recall on open testing, respectively.Li et al. proposed a rule-based approach to recognize the Chinese organization names and corresponding abbreviations[16]. To recognize organization names, they built an exclusive key word library.The structure features and composition rules are extracted for FNE recognition. It achieved 67.18%precision and 74.14%recall for abbreviation recognition in open test.Generally speaking, most of the existing methods utilize the content feature, grammar of FNEs and composition rules as major features.

3. OUR APPROACH

Different from regular NER problem, financial named entity (FNE) recognition has shown its own characteristics. Especially, the abbreviation FNEs wording - unclear are widely exist. Generally speaking, they are hard to extract from text directly since no clear discriminative features. Considering the fact that in a news text the full form FNEsalways appear once or twice and then theyaresubstituted by its abbreviations,the context and the full form FNEsare helpful to recognize the abbreviation FNEs.

In this study, a corpus on the financial field is firstly constructed in which the FNEs are annotated. Using this corpus, a recognition model based on conditional random fields is developed for recognizing full form FNEs in the text. The mutual information, information entropy and word similarity measurement are combined to recognize the abbreviation FNEs. The implementation of this approachconsists of threesteps: text pre-processing, recognition of fullform FNEs, and recognition of abbreviation FNEs.The workflow is shown as Figure 1 below:

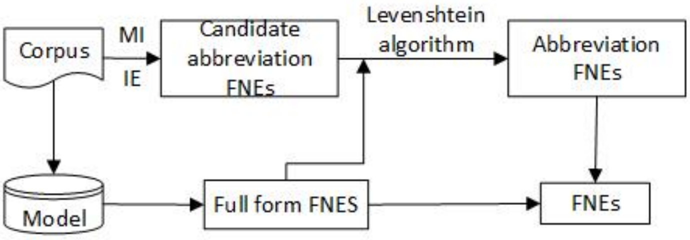

Figure 1. The frameworkof our approach

3.1. Text pre-processing

In financial news texts, the names of stock are always in great use. Using these stock names as clues, some abbreviation andfullFNEs may be found. Usingan examplestock name like "平安银行(*Ping An Bank*)" as clue, the other candidate FNEs having somekey words of the stock name tends to be a real FNEs, such as "平安集团(*Ping An Group*)" and "平安保险公司(*Ping An Insurance Company*)". We may split the texts into clause based on punctuations. Considering that the list of stock names can be easily accessed from the Internet, we may find out all the stock names exist in the texts. Using these stock key names as clues, the candidate FNEs are extracted from clauses. Do you need Chinese wording

3.2. Recognition of full form FNE

Generally speaking, for the full form FNEs, they are usually named in some certain ways. Severalkinds of context and content featuresare selected for determining whether a phrase is FNE,namely

1. Some company names end with organizationword or phrase, like "控股有限公司(*Holdings Ltd.*)","集团有限公司 (*Group Corporation*)", "公司子公司 (*subsidiary*)", or simply "公司 (*Corporation*)". These words or phrases as the suffix key words which indicate the right boundary of FNEs. The lexicon of suffixkey words may be extracted from knownFNE lexicon.
2. Many company names contain or start with location names, which present the place that the companies or organizations belong to. For example, "深圳(Shenzhen)" in "深圳华为公司 (Shenzhen Huawei Corporation)". The place names may be obtained from the dictionary.

Obviously, these place names are useful features for identifying FNEs, especially determining the left boundary of FNEs.

3. Precursor predicates. When a FNE plays as an object, the first predicate before the FNEs tends to be useful features for identifying the FNEs. For instance, “收购 (purchase)”, “参股 (*participate*)”, etc. The precursor predicates can be used as clues to determinethe left boundary of FNEs.
4. Subsequent words. The words frequently following the FNEs may be used as features for determining the right boundary of FNEs.

Based on the analysisabove, we obtain the following primitive features to be used for FNEs recognition.

Table 1. Primitive feature template forrecognizing FNEs

	Feature	Describe
1	Suffix key	Key features for organization type FNEs, especially the right bournary. In “杭州娃哈哈集团有限公司 (*Hangzhou Wahaha Group Co., Ltd.*)”, “有限公司 (*Co., Ltd.*)”is the suffix of the entity.
2	Place word	Place word as features. In “杭州娃哈哈集团有限公司(*Hangzhou Wahaha Group Co., Ltd.*)”, “杭州 (*Hangzhou*)”is the location word of the entity.
3	Title word	Title word is helpful to determine the right boundary. In “杭州娃哈哈集团有限公司董事长(*The chairman of Hangzhou Wahaha Group Co., Ltd.*)”, “董事长(*The chairman*)”is the title word of the entity.
4	Precursor predicate	The precursor predicate of a FNE is used for determining the left boundary of FNEs. In “收购京东 (*PurchaseJingdong*)”, “收购 (purchase)” is the precursor word.
5	Subsequent word	The frequently co-occurred subsequent word of a FNEs. In “娃哈哈集团日前发表声明 (*Wahaha Group recently issued a statement*)”, “日前(*recently*)” is the subsequent word.
6	POS	The Part-of-speech of current word

These features are incorporated in a linear Conditional Random Fields (CRFs) based classifier, which is trained by using the annotated FNE corpus. The obtained classifier is applied to identify full form FNEs.Conditional Random Fields (CRFs) is a graph-based model for calculating the conditional probability of values on designated output nodes given values assigned to other input nodes. Conditional random field is a discriminative probabilistic model for labeling the sequential text. In the study, a linear CRFs classifier is implemented based on the CRF++ package.

3.3. Recognition of abbreviation FNEs

The abbreviation form FNEs widely exist in financial text. Different from the full form FNEs with has relatively clear construction form and boundary clues, abbreviation form FNEs are constructed in various forms. It leads to quite different features for NER. Thus, the CRFs based sequence labeling algorithms have the difficulties to handle them.

The observation on abbreviation form FNEs showed that the stock names are good identification clues. Meanwhile, the recognized full form FNEs are helpful to identify its corresponding abbreviation forms. Therefore,this information is utilized for identifying abbreviation FNEs.

Mutual information (MI) is a widely used metric ininformation theory.It estimates the correlation between two events. highervalue of the mutual information indicates the stronger correlation between two events and vice versa. It may be used to estimate the internal association within of a word or phrase. Being Different from mutual information, the information entropy (IE) is the metric for measuring the average amount of information in an event. It is a good feature tomeasure the randomness of the left and right neighbor words to named entity. It means that information entropy may be used a feature to determine the boundary of FNEs.

The mutual information between two words, W_1 and W_2,is calculated as follows,

$$MI(W_1, W_2) = \log_2\left(\frac{P(W_1,W_2)}{P(W_1)\cdot P(W_2)}\right) \quad (1)$$

Where, $P(W_1)$ is the occurrence probability of word W_1, $P(W_1, W_2)$is the co-occurrence probability of W_1 andW_2.

The measurement of entropy has two parts, namely the left word entropy and right word entropy. The formulas are given below:

$$IE_l(C) = -\sum_{s\in l} P(sC|C)/\log_2 P(sC|C) \quad (2)$$

$$IE_r(C) = -\sum_{s\in r} P(Cr|C)/\log_2 P(Cr|C) \quad (3)$$

Where, $P(sC|C)$ isthe conditional probability of string $s$co-occurs with C at the left side. $P(Cr|C)$isthe conditional probability of string $r$co-occurs with C at the right side.

Considering that the FNEs always consist of more than two smaller components, (1)may be changed into the following form:

$$MI(W_1W_2 \dots W_{n-1}W_n) = \log_2\left(\frac{P(W_1W_2 \dots W_{n-1}W_n)}{AVP}\right) \quad (4)$$

Where, $AVP = \frac{1}{n-1}\sum_{i=1}^{n-1} P(W_1 \dots W_i)P(W_{i+1} \dots W_n)$(5)

The observation on the FNEs shows that most abbreviation FNEs consistsof no more than four words. Thus, the parameter of $n$in Formula 4 is set to less than 5. Furthermore, it is observed that few FNEs have auxiliary components. Thus, the candidate FNEs with auxiliary words like "的" and "了" will be filtered out before the MI and IE calculation.

The candidate FNEs are then ranked and identifiedby the value of multiplied results of MI, min $\{IE_l(C), IE_r(C)\}$ and their frequency.

The observation showsthose current candidates FNEs are not satisfactory. Many of them are commonly used meaningful phrases but not FNEs. Therefore, we further indentify the abbreviation FNEs by estimating the similarity between the candidate FNEs and the stock name/identified full form FNEs through applying the Levenshteinalgorithm. This algorithm refers to the minimum steps to convert a string to another. It contains steps ofreplacing a character by another character, inserting or removing a character at a time.The candidate FNEs with the highest similarity, which indicates most similar to known FNEs,are chosen as the final abbreviation FNEs.

4. EVALUATION AND DISCUSSION

The propose FNEs recognition approach is firstly evaluated on a financial text dataset. This dataset consists of 5,500 sentences, which wereextracted from 15,000 financial news texts. In these sentences, 2,500 ones have annotatedFNEs and 3,000 ones have no FNEs. The observation on this dataset shows that about 85% FNEs are in full form and 15% are in abbreviation form. In the experiments, 3,500 sentences are selected as training data and 2,000 ones are used as testing data. Ten folder cross-validations are performed.

For comparison, a baseline is constructed. It is a regular NER approach, which is based on CRFs. The features are extracted from the ORG entities in SIGHAN 2006 named entity recognition bakeoff MSRA dataset (in short, MSRA dataset). Generally speaking, the features cover the suffix key, place name, precursor predicate, subsequent word, and part-of-speech which is mentioned in Table 1, but without considering the feature of title. Furthermore, the baseline is a general NER approach without considering the abbreviation NER process.

Table 2 shows the performance achieved by the baseline and our approach, respectively.The performances are measured by precision (P), recall (R) and F-measure (F).

Table 2. Performances on financial dataset

	P	R	F
Baseline	79.21%	86.82%	0.8284
MI+context features	82.03%	87.04%	0.8446
IE+context features	82.38%	88.25%	0.8521
MI+IE+context features	85.53%	92.35%	0.8881

As shown in table 2, our approach has achieved a better performance onFNE recognition. F1 performance increases 5.97% from the baseline. Meanwhile, it is shown that the incorporation of mutual information (MI) and information entropy (IE) features improves the performance effectively. The further observation shows that the majority of performance improvement attributes to the recognition of abbreviation FNEs ,which is the shortcoming of baseline.

The observations on the cases poor wording – rewrite ofwrongly recognized show that many errors are attribute to the unsatisfactory coverage of training data. The financial text has many kinds of subcategories such as stock-specific research and stock comments. Our training dataset is relatively narrow. More training data should be helpful to further improve the performance.

In the second experiment, we expand the training dataset by appending 400 finance-related sentences selected from MSRA dataset. Experimental results achieved by the baseline and our approach with expanded training dataset are given in Table 3, respectively.

Table 3. Performancesforthe expanded training dataset (MSRA Dataset+Financial Dataset)

	P	R	F
Baseline	83.21%	88.41%	0.8573
MI+context feature	83.01%	89.11%	0.8595
IE + context feature	85.22%	90.35%	0.8771
MI+IE + context feature	92.02%	93.77%	0.9289

Table 2 and 3 show that using the expanded training dataset in the same domain, the performance of our approach increases. The further observation shows that if only the MSRA dataset is used for training, the achieved precision, recall and F1 are 70.98%, 73.01% and 0.7198, respectively. It is obviously lower than the performance achieved by using financial dataset. Moreover, the individual use of MI/IE features and context features has ledto few performance improvement compared to the baseline. While MI/IE features are incorporated with context features, the approach achieves the best result. It is because that MI and IE consideringdifferent characteristics aspects ofFNEs. MI focuses on the internal association, which emphasizes the co-occurrences of its components. IE focus on the flexibility of a FNE which emphasizes the homogeneity of its neighbors. When the organization name having different abbreviations, IEfeature cannot ensure a good performance. Once the above two features are incorporated, both the internal association and neighboring homogeneity are taken into consideration which leads to a better performance.

The experimental results also show that our approach improves precision effectively while the recall improvement is limited. The observation on wrongly recognized samples shows two reasons broken sentence – rewrite. Firstly, the types of abbreviation FNEs are limited. Thus, once a type of abbreviation FNE has beenrecognized, most of its occurrence,which is even more than ten times, can be correctly recognized. It leads to obviously precision improvement. In addition, though mutual information and information entropy features may be used recognize different styles of FNEs, the boundary detection is difficult to control. It decreases the performance of FNE recognition.

5. CONCLUSION

In this paper, we present thefinancial named entities recognition approach. This approach recognizes full form FNEs and abbreviation FNEs, respectively. The internal features such as suffix key, place word, title word and precursor predicate are adopted in a conditional random fields based classifier forrecognizing the full form FNEs. The mutual information, boundary information entropy and context features are incorporated to identify abbreviation FNE candidates. The similarity between these candidates and the stock name/identified full form FNEs are estimated to determine the final abbreviation FNEs. The experimental results show that the proposed approach achieves good performance, especially on the abbreviation FNE recognition.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61300112, 61370165), Natural Science Foundation of Guangdong Province (No. S2012040007390, S2013010014475), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, MOE Specialized Research Fund for the Doctoral Program of Higher Education 20122302120070, Open Projects Program of National Laboratory of Pattern Recognition, Shenzhen International Co-operation Research Funding GJHZ20120613110641217, and Shenzhen Foundationa –whoat is here Research FundingJCYJ20120613152557576 and JC201005260118A

References

[1] Luo Z. and Song R. Integrated and Fast Recognition of Proper Noun in Modern Chinese Word Segmentation. Proceedings of International Conference on Chinese Computing, Singapore, pp. 323-328, 2001.

[2] Luo H. andJi Z. Inverse Name Frequency Model and Rules Based on Chinese Name Identifying. In "Natural Language Understanding and Machine Translation", C. N. Huang & P. Zhang, ed., Tsinghua Univ. Press, Beijing, pp. 123-128, 2001.

[3] Song R. Person Name Recognition Method Based on Corpus and Rule. In “Computational Language Research and Development", L. W. Chen & Q. Yuan, ed., Beijing Institute of Linguistic Press.1993

[4] Tan H. Y. Chinese Place Automatic Recognition Research. Proceedings of Computational Language, Tsinghua Univ. Press, Beijing, China. 1999

[5] Zhang Hua-Ping, LiuQun,Model of Chinese Words Rough Segmentation Based on N-Shortest-Paths Method. Journal of Chinese Information Processing. pp.1-7, 2002.

[6] Sun J., Gao J. F., Zhang L., Zhou M Huang, C.N , Chinese Named Entity Identification Using Class-based Language Model,Proc. of the 19th International Conference on Computational Linguistics, 967-973, 2002

[7] Sun M.S. English Transliteration Automatic Recognition. In "Computational Language Research and Development", L. W. Chen & Q. Yuan, ed., Beijing Institute of Linguistic Press.1993.

[8] Ye S.R, Chua T.S., Liu J. M., An Agent-based Approach to Chinese Named Entity Recognition, Proc. of the 19th International Conference on Computational Linguistics, pp 1149-1155, 2002

[9] Lv Y.J., Zhao T. J. Levelled Unknown Chinese Words Resolution by Dynamic Programming. Journal of Chinese Information Processing., pp. 28-33, 2001.

[10] Y. Wu, J. Zhao, B. Xu. Chinese Named Entity Recognition by Combining a Statistical Model and Human Knowledge. In Proceedings of ACL 2003 workshop on Multilingual and Mixed-language Named Entity Recognition: 65-72.

[11] Xiao Chen, Hui Liu, Yu-quan Chen. Chinese Organization Names Recognition based on SVM. Application Research of Computers, 25(2), 2008.

[12] Hong-kuiYu, Hua-ping Zhang, et al. Chinese Named Entity Identification Using Cascaded Hidden Markov Model. Journal on Communications, 27(2), 2006.

[13] Jun Kazama, KentaroTorisawa, Exploiting Wikipdia as External Knowledge for Named Entity Recognition, In Proceeding of EMNLP-CoNLL2007:698-707, 2007

[14] Kebin L, Fang L, Lei L, et al. Implementation of a Kernel-Based Chinese Relation Extraction System. Journal of Computer Research and Development, 44(8): 1406-1411, 2007.

[15] N. Wang, Kam-Fai Wong et al. Recognition on Companies Names in Chinese Financial News. Journal of Chinese Information Processing, 16(2):1-6, 2002

[16] J. Shen, F. Li, F. Xu, et al. Recognition of Chinese Organization Names and Abbreviations. Journal of Chinese Information Processing, 21(6), 2007