Postprint: Domain-Specific Concept Discovery Based on Conditional Random Fields and Information Entropy

Authors: Fu Yao, Wan Jing, Xing Lidong

Date: 2019-01-03T00:00:00+00:00

Abstract

To address the problem of automatically identifying existing concepts and discovering new concepts in specific domains, we propose an extraction method based on conditional random fields and information entropy. By using conditional random fields to predict the boundaries of concept words in text and comparing them with concepts in a dictionary, we filter candidate new concepts and identify their approximate locations. Mutual information and left-right entropy are then used respectively to evaluate the internal cohesion within the concept window and the freedom of concept boundaries, thereby discovering new domain-specific concepts. Experiments demonstrate that using this method for concept discovery achieves better results than using conditional random fields alone, with the accuracy of concept discovery improving by 20.06% and 46.54% for character-based and word-based models, respectively.

Full Text

Preamble

Vol. 37 No. 3
Application Research of Computers
ChinaXiv Partner Journal

Crf and Information Entropy Based Method for New Words Discovery in Specific Domain

Fu Yao[1], Wan Jing[1]†, Xing Lidong[2]
(1. College of Information Science & Technology, Beijing University of Chemical Technology, Beijing 100029, China;
2. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: To address the problem of automatically identifying existing concepts and discovering new concepts in specific domains, this paper proposes an extraction method based on conditional random fields and information entropy. The method uses conditional random fields to predict the boundaries of conceptual words in text. By comparing these predictions with existing concepts in a dictionary, candidate new concepts are selected and their approximate locations identified. Mutual information and left-right entropy are then employed to evaluate the internal cohesion within the concept window and the boundary freedom of concepts, thereby discovering new professional concepts. Experiments demonstrate that this approach achieves better results than using conditional random fields alone, improving concept discovery accuracy by 20.06% for character-based models and 46.54% for word-based models.

Key words: concept recognition; new concept discovery; conditional random field; information entropy; specific field

0 Introduction

The rapid development of scientific research in recent years has led to an endless emergence of specialized vocabulary across various domains. Typically, domain-specific terminology appears frequently in knowledge dissemination media for the corresponding field. These terms exhibit considerable particularity and specialization, and only domain experts possess adequate familiarity with them. The pace of new domain concept emergence has even outstripped the cognitive capacity of scholars in some fields. Consequently, efficient, accurate, and comprehensive identification and discovery of domain-specific new words hold significant importance.

Existing research on domain-specific new word discovery primarily falls into two categories: rule-based methods and statistical methods [**?**]. Rule-based approaches require constructing a rule library, where domain experts formulate universal and specific word-formation rules based on professional knowledge development and linguistic principles. Li Ming [**?**] utilized an improved Apriori algorithm to process corpora and generate association rules for extracting new domain vocabulary. Sasano et al. [**?**] addressed Japanese neologisms by employing derivation rules and onomatopoeia patterns, discovering optimal paths by adding new nodes to sentence structure frameworks, achieving favorable recognition results for certain new word categories. Zheng Jiaheng et al. [**?**] established a rule library based on Chinese word-formation principles, discovering new words through "mutually exclusive string" filtering and morphological rules. While rule-based methods achieve high accuracy, they suffer from severe domain limitations. The rapid update rate of specialized vocabulary necessitates continuous rule updates, resulting in high construction costs.

Statistical methods identify domain neologisms by calculating statistical features such as word frequency and co-occurrence probabilities from large-scale

corpora. Li et al. [?] proposed a method based on word internal cohesion and boundary freedom for discovering new words from fragmented strings generated during word segmentation. Yao Rongpeng et al. [?] employed the N-Gram algorithm to obtain candidate new words, then applied improved mutual information and adjacent entropy for candidate expansion and filtering, combining dictionary screening to obtain new words. Lei et al. [?] introduced a hierarchical clustering approach that grouped Weibo corpora into topic-specific clusters to enhance statistical features and improve new word extraction accuracy. Although statistical methods are not domain-restricted, they typically suffer from low recognition accuracy due to data sparsity.

To address the limitations of each approach, many researchers have proposed hybrid statistical and rule-based methods. Du Liping et al. [?] combined an improved mutual information algorithm with a small set of basic rules to automatically identify internet neologisms from corpora, demonstrating effectiveness in large-scale new word discovery using Baidu Tieba data. Lei Yiming et al. [?] employed a mutual information statistical model with right-neighbor iteration for candidate acquisition and introduced external statistical measures to filter low-frequency words. Zhou Shuangshuang [?] classified and summarized Weibo neologisms using heuristic rules, then improved new word boundary recognition accuracy and low-frequency neologism identification precision by fusing CRF and SVM models with a modified C/NC-value algorithm. Hybrid approaches combine the advantages of both paradigms, often yielding superior results in new word identification.

Currently, various machine learning methods have been applied to new word discovery with promising results, including CRF (Conditional Random Fields), HMM (Hidden Markov Models), SVM (Support Vector Machines), and DT (Decision Trees). Chen Fei et al. [?] summarized numerous statistical features for distinguishing new word boundaries and utilized CRF for experiments. Ding Xiangwu et al. [?] employed word2vec to convert words into vectors during medical domain text structuring, using vector similarity to represent internal word cohesion, combined with left-right entropy and word frequency statistics to discover new medical terms. Xu Yuanfang et al. [?] vectorized candidate new words and word features to construct candidate vectors, building matrices with trained support vectors for SVM-based new word classification.

This paper proposes a novel domain-specific new word discovery method combining CRF with mutual information and left-right entropy. First, existing professional vocabulary is used to annotate corpora for CRF training. The trained model then identifies candidate strings, which consist of complete professional terms, incomplete terms, and non-professional terms. After filtering out existing concepts from the lexicon, mutual information is applied to concatenate remaining strings, and left-right entropy is used for screening to discover new concepts.

1.1 Conditional Random Fields

Conditional Random Fields (CRF) [**?**] represent a discriminative probabilistic model and a type of Markov random field. They enable training and inference using complex, overlapping, and non-independent features, fully leveraging contextual information while accommodating additional external features. This model captures rich feature information while addressing label bias issues present in maximum entropy models.

The CRF model structure is illustrated in [Figure 1: see original paper]. Connections between vertices represent dependencies among random variables. In CRF, random variables satisfy conditional probability distributions, with given observation sequences as random variables.

Let the observation sequence be $X = \{x_1, x_2, ..., x_n\}$, where input data can be characters or words from text. The corresponding state sequence is $Y = \{y_1, y_2, ..., y_n\}$. When the observation sequence X takes value x, the conditional probability of state sequence Y taking value y is calculated as:

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{k=1}^{K}\sum_{t=1}^{n} \theta_k f_k(y_t, y_{t-1}, x_t)\right)$$

where θ_k represents the weight parameter for feature function f_k; y_t and y_{t-1} denote the current and previous output states respectively; x_t is the current input state; and $Z(x)$ is the normalization factor calculated as:

$$Z(x) = \sum_{y'} \exp\left(\sum_{k=1}^{K}\sum_{t=1}^{n} \theta_k f_k(y'_t, y'_{t-1}, x_t)\right)$$

In CRF applications, feature selection directly impacts feature function effectiveness. There is no fixed format for feature selection; it requires comprehensive consideration of target domain, text language, and expression characteristics. Typically, input state sequence features are combined through superposition.

1.2 Mutual Information

Mutual Information (MI) [**?**] measures the mutual dependence between two random variables. Larger mutual information indicates greater likelihood that a bigram constitutes a new word or part thereof, reflecting stronger internal word cohesion. By comparing with a preset threshold, when internal cohesion exceeds the threshold, the two components are considered to form a word. The mutual information between random variables X and Y is defined as:

$$MI(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

where $p(x,y)$ is the joint probability distribution function (co-occurrence probability in corpus); $p(x)$ and $p(y)$ are marginal probability distribution functions (individual occurrence probabilities). When $MI(x,y) \gg 0$, X and Y are highly correlated, frequently co-occurring, indicating higher probability that string xy forms a new word. When $MI(x,y) = 0$, X and Y are independent. When $MI(x,y) < 0$, X and Y are negatively correlated.

1.3 Left-Right Entropy

Left-right entropy, or branch entropy (BE) [**?**], measures the uncertainty of characters adjacent to a word. Meaningful words in corpora typically appear with high frequency across different documents, demonstrating high flexibility in collocating with various external conditions. More diverse collocations indicate greater word flexibility and higher boundary freedom. This paper employs left-right entropy of candidate new words to quantify boundary freedom. The left and right entropy for candidate word w are defined as:

$$H_l(w) = -\sum_{w_l \in S_l} p(w_l|w) \log p(w_l|w)$$

$$H_r(w) = -\sum_{w_r \in S_r} p(w_r|w) \log p(w_r|w)$$

where S_l is the left-neighbor character set of candidate word w; w_l is an element in S_l; S_r is the right-neighbor character set; and w_r is an element in S_r. If both left and right entropy values are large, the candidate word has many adjacent word strings with relatively uniform frequency distribution, indicating low probability of forming new words with neighbors. If either left or right entropy is small, the frequency distribution of adjacent word strings is non-uniform, suggesting higher probability of forming new words with high-frequency adjacent strings.

---

2 Domain-Specific Concept Discovery Based on CRF and Information Entropy

This paper frames domain-specific concept discovery as a boundary prediction problem for text sequences. Concept discovery is integrated into the word segmentation process, identifying new concepts by comparing with existing lexicons.

The method comprises three main components: corpus annotation, CRF model training and candidate concept identification, and mutual information concatenation with left-right entropy filtering, as illustrated in [Figure 2: see original paper].

2.1 Corpus Annotation

The annotation scheme used in this paper is shown in . We adopt the commonly used BEMSN tagging set for word segmentation, which identifies word beginnings (B), middles (M), ends (E), single-character words (S), and irrelevant words (N). Additionally, part-of-speech tagging is required for segmentation, with details referring to the *HanLP Part-of-Speech Tagging Set*.

2.2 CRF Model Training and Candidate Concept Identification

The data processing pipeline is shown in [Figure 3: see original paper]. Concept identification is treated as a sequence labeling problem. The raw corpus is segmented using a word segmentation tool, leveraging contextual information and external features. Segmentation results are annotated with part-of-speech, word beginning, middle, and end tags according to the concept lexicon. The input consists of word sequences with part-of-speech markers, as shown in Equation (6):

$$W = \{w_1/t_1, w_2/t_2, ..., w_n/t_n\}$$

where n represents the number of words in the sentence; w_i denotes a word; and t_i represents its part-of-speech tag.

Using the annotated corpus, CRF model parameters are trained to obtain a domain concept recognition model through the learning process. As shown in [Figure 4: see original paper], the trained model combined with CRF decoding algorithms identifies concept word boundaries in new corpus texts, performing word splitting and merging to output an optimal "word form/part-of-speech" sequence:

$$S^* = \arg\max_S P(S|W)$$

where $S = \{S_1, S_2, ..., S_n\}$ represents the state sequence.

2.3 Mutual Information Concatenation and Left-Right Entropy Filtering

The CRF model recognition yields candidate concept words, many of which are incorrect due to incomplete strings. The proposed algorithm edits these incorrect concepts based on the following principle: 统计 incorrect concept words' left and right candidate words, calculate corresponding mutual information, select

candidates with larger MI values for concatenation to form new words, then compute the left-right entropy of these new words, taking the smaller entropy value as the word' s information entropy. Through iterative recursion, the new word with maximum information entropy under constraints is identified as the discovered concept.

For example, CRF might identify the incomplete candidate "施工过程仿真" (construction process simulation). By concatenating the right neighbor via mutual information, we obtain "施工过程仿真分析" (construction process simulation analysis). After calculating its left-right entropy and taking the minimum value, comparison with other candidates' entropy values reveals that "施工过程仿真分析" has the maximum information entropy, thus constituting a discovered new concept.

The algorithm is described as follows:

Input: Candidate concept word set S
Output: Concept word with maximum information entropy

1. For each S_i in S where $i \in \{1, 2, 3, ..., n\}$:
   - If $|S_i| = 1$, use mutual information method
   - If $|S_i| > 1$, use part-of-speech distribution concatenation method
   - Let l_w and r_w be left and right adjacent words respectively
   - Store the value of S_i when $H(S_i)$ reaches maximum
2. End for

---

3 Experimental Setup and Results Analysis

3.1 Experimental Dataset

This paper uses books and journals from the construction engineering domain as the corpus for new word discovery experiments. A total of 70,962 construction engineering concept terms were extracted and used to annotate 245 construction engineering books. [Figure 5: see original paper] shows a sample of the raw data, where all text from the books serves as experimental material.

To investigate information extraction from different segmentation granularities, the experimental corpus is divided into two groups: one segmented using the HanLP tool with part-of-speech information from the *HanLP Part-of-Speech Tagging Set*, and the other using characters directly as experimental units.

Experimental data is formatted as shown in [Figure 6: see original paper], where each annotation line contains: the word to be recognized, its part-of-speech feature, and the correct tag.

3.2 Experimental Scheme

Feature templates are used to generate feature functions, as illustrated in [Figure 7: see original paper]. Each line in the template file represents a template specifying each unit in the input data via `%x[Row, Column]`, where Row indicates line offset and Column indicates column position.

Experiments employ CRF++ 0.58 as the CRF implementation tool. For convenience, the aforementioned feature sets are labeled with letters as shown in .

Four cross-validation experiments were conducted to verify algorithm effectiveness:

Experiment 1: Character vs. word-based CRF comparison. Investigates CRF model performance with different annotation granularities (character or word) for information extraction.

Experiment 2: Part-of-speech feature effectiveness test. Evaluates the impact of adding part-of-speech features in word-based experiments.

Experiment 3: Mutual information and left-right entropy addition comparison. Compares performance with and without MI and entropy processing in character- and word-based experiments.

Experiment 4: Fusion experiment. Uses the recognition model from Experiment 2 to annotate new texts for concept discovery, applying MI and left-right entropy for concatenation and filtering, comparing with other groups to assess the impact of part-of-speech features and post-processing.

3.3 Experimental Results

Results are shown in [Figure 8: see original paper], where WP denotes new word discovery precision, WR denotes recall, and WF denotes F-score; NP, NR, and NF represent precision, recall, and F-score for construction engineering concept recognition respectively.

Experiment 1 results ([Figure 8: see original paper]) show that character-based recognition achieves higher precision and recall than word-based approaches. Using CRF with characters as basic units yields recall rates 3.56% higher for new word discovery and 5.7% higher for concept recognition compared to the best word-based performance.

Experiment 2 results ([Figure 9: see original paper]) demonstrate that adding part-of-speech features to word-based experiments improves information extraction recall.

Experiment 3 results ([Figure 10: see original paper]) reveal significant improvement after adding MI and left-right entropy processing. Character-based model accuracy improved by 20.06%, while word-based model accuracy improved by 46.54%. The smaller improvement for character-based models is

attributed to character concatenation producing less complete results than direct word concatenation.

The fusion experiment, implementing the complete proposed method, achieved excellent results by using part-of-speech-enhanced CRF to identify approximate concept locations, then applying MI and entropy for concatenation and filtering. Results are shown in [Figure 11: see original paper].

Summary: Post-processing CRF recognition with mutual information and left-right entropy effectively improves concept discovery precision and recall. CRF identifies approximate concept locations, and information-theoretic methods extract complete, accurate construction engineering concepts from these positions. Comparison between character- and word-based methods shows that while CRF identifies similar concept locations, the word-based approach with part-of-speech features and post-processing yields the best results.

---

4 Conclusion

This paper proposes a domain-specific concept identification method combining conditional random fields with left-right entropy. The approach involves corpus annotation, CRF model training for candidate word recognition, lexicon-based filtering of existing concepts, and concatenation/screening of remaining candidates using mutual information and left-right entropy to discover new concepts. Experiments on construction engineering corpora validate the method' s effectiveness from perspectives of segmentation granularity, feature selection, and MI/entropy processing. Results show that the method improves information extraction precision and recall without additional manual annotation while enhancing recognition efficiency. Note that annotation set quality and machine performance significantly impact model training and consequently concept identification effectiveness.

References

[1] Zhang Haijun, Shi Shumin, Zhu Chaoyong, et al. Survey of Chinese new words identification [J]. Computer Science, 2010, 37(3): 6-10.

[2] Li Ming. New words discovery research for specific areas [D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2012.

[3] Sasano R, Kurohashi S, Okumura M. A simple approach to unknown word processing in Japanese morphological analysis [C]//Proc of International Joint Conference on Natural Language Processing. 2013.

[4] Zheng Jiaheng, Li Wenhua. A study on automatic identification for internet new words according to word-building rule [J]. Journal of Shanxi University: Nat. Sci. Ed., 2002, 25(2): 115-119.

[5] Li Wenkun, Zhang Yangsen, Chen Ruoyu. New word detection based on inner combination degree and boundary freedom degree of word [J]. Application Research of Computers, 2015, 32(8): 2302-2304.

[6] Yao Rongpeng, Xu Guoyan, Song Jian. Micro-blog new word discovery method based on improved mutual information and branch entropy [J]. Journal of Computer Applications, 2016, 36(10): 2772-2776.

[7] Lei K, Zhang W Y, Zhang K, et al. Extracting unknown words from Sina Weibo via data clustering [C]//Proc of IEEE International Conference on Communications. 2016: 1182-1187.

[8] Du Liping, Li Xiaoge, Yu Gen, et al. New word detection based on an improved pmi algorithm for enhancing segmentation system [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52(1): 35-40.

[9] Lei Yiming, Liu Yong, Huo Hua. New word discovery based on microblog corpus for network language [J]. Computer Engineering and Design, 2017, 38(3): 789-794.

[10] Zhou Shuangshuang. Combinating of rules and statistics for new word detection of microblog text [D]. Beijing: Beijing Jiaotong University, 2017.

[11] Chen Fei, Liu Yiqun, Wei Chao, et al. Open domain new word detection using condition random field method [J]. Journal of Software, 2013, 24(5): 1051-1060.

[12] Ding Xiangwu, Zhang Xihua. Text structuralization in medical field [J]. Computer Engineering and Design, 2017, 38(10): 2873-2878.

[13] Xu Yuanfang, Li Chengcheng. Research on new word identification based on svm and word characteristics [J]. Computer Technology and Development, 2012, 22(5): 134-136.

[14] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]//Proc of International Conference on Machine Learning, Massachusetts. 2001: 282-289.

[15] Ong T H, Chen H. Updateable PAT-tree approach to Chinese key phrase extraction using mutual information: a linguistic foundation for knowledge management [C]//Proc of the 2nd Asian Digital Library Conference. 1999: 63-84.

[16] Liu Weitong, Liu Peiyu, Liu Wenfeng, et al. New word discovery algorithm based on mutual information and branch entropy [J/OL]. Application Research of Computers, 2019, 36(5). (http://www.arocmag.com/article/02-2019-05-017.html)

Note: Figure translations are in progress. See original paper for figures.