Domain-Specific Term Extraction and Its Application in Text Classification

Tao Liu[1], Xiao-long Wang[1], Guan Yi[1], Zhi-ming Xu[1], Qiang Wang[1]

[1]Harbin Institute of Technology, Harbin, China

Abstract

A statistical method is proposed for domain-specific term extraction from domain comparative corpora. It takes distribution of a candidate word among domains and within a domain into account. Entropy impurity is used to measure distribution of a word among domains and within a domain. Normalization step is added into the extraction process to cope with unbalanced corpora. So it characterizes attributes of domain-specific term more precisely and more effectively than previous term extraction approaches. Domain-specific terms are applied in text classification as the feature space. Experiments show that it achieves better performance than traditional methods for feature selection.

Keywords: domain-specific term, entropy impurity, normalization, text classification, feature selection

1. Introduction

Domain-specific term extraction [1,2] is an important task in natural language processing since it can be applied to a variety of fields such as domain ontology construction [3,4], information retrieval [5], information extraction [6] and text classification [7]. Existing related resources include Wordnet [8] for English and Hownet [9] for Chinese, but these resources are knowledge databases for common sense knowledge of the world. The deficiency of domain knowledge limits their application greatly. So generating domain lexicon will contribute to all domain oriented applications [10]. Developing an automatic domain lexicon construction method has become a hot research issue in this field.

Previous approaches can be classified as rule-based and statistic-based methods. In rule-based methods [3] words which match the predefined rule templates are extracted as terms. The costly manual collection of related rule templates by domain experts limits their widely application. In addition rules set can never be exhaustive for all language phenomena and multiple rules interfere with each other. Statistical methods are normally based on domain comparative corpora which are categorized corpora. Each corpus is a collection of documents which belong to a same domain based on contents. The main idea for statistical methods [4] is to count the occurrence information of words in domain comparative corpora. Words occurrence frequency or TFIDF weighting formula are used as ranking criteria in these methods, but purely using words frequency is biased to words frequency and ignores the comparative distribution of words in different corpora [2]. Feiyu Xu [1] updated the TFIDF formula by KFIDF formula. More domain-specific information is injected to KFIDF method. Paola Velardi [2] proposes domain relevant (DR) and domain consensus (DC) ranking formula for terms. DC formula takes the distribution consensus of a term in all documents of a certain domain into account, but DR and DC don't discriminate categories sizes and lengths of different documents and can't produce promising results when corpora are unbalanced.

A statistical method is proposed in this paper and it is based on two evaluation criteria. Firstly, terms which are relevant to a certain domain occur frequently in this domain and rarely in others [1,2]. Secondly, words occur in most documents of a domain tend to be relevant to this domain [2]. Impurity measure [14] is introduced into measuring a word's distribution in corpora. The normalization step is added into impurity measurement to discriminate different lengths of documents and different sizes of domain corpora. Experiments show the impurity measure and the normalization step make this method more precisely to characterize terms. To evaluate the quality of terms collected, an evaluation method also an application area, text classification, is presented.

2. Domain-specific term extraction

2.1. Two extraction principles

A good statistical method for domain-specific term extraction is to make best use of domain comparative corpora and explore information the corpora indicate. From the analysis of domain comparative corpora, two principles are presented for a term to be domain-specific. Firstly, the term should have a skew occurrence distribution in all the domains. The less amounts of domains it occurs in, the more relevant it tends to be with the domains which it occurs in. Secondly, an ideal domain-specific term should have a uniform occurrence distribution in documents of its relevant domain. If a term occurs in only one document of a domain, it is unlikely to be a domain-specific term for its occurrence is probably haphazard or exceptional and can't represent the common sense of this domain. The second principle means distribution of domain-specific terms in documents of their relevant domains should be as uniform as possible.

2.2. Impurity measure

Impurity measure is popularly used in finding optimal branching of attribute in decision tree. Based on the two principles above, impurity measure [14] is also suitable to measure distribution of words in corpora. Suppose to decide the impurity of samples which are drawn randomly from *N* categories. If these samples are actually of one category, then the impurity gets minimum value. If all the *N* categories are equally likely in the samples, then the impurity gets maximum value.

The popular measure for impurity is entropy impurity (information impurity) [14]:

$$Entropy\ impurity = -\sum_j P(\omega_j)\log P(\omega_j) \quad (1)$$

Where ω_j denotes the category labeled with *j*. $P(\omega_j)$ is the probability of ω_j category in all categories. Another measure for impurity is variance impurity for two-category cases and the generalization form (Gini impurity) of variance impurity for two or more categories.

2.3. Normalized corpora impurity (NCI) of a word

The aim of the first principle mentioned in 2.1 is to measure words' distribution in the corpora of different domains. It is computed by NCI. Suppose there are m domains as $D_1, D_2, ..., D_m$. The number of documents for these domains are $n_1, n_2, ..., n_m$ respectively. The probability that a word *W* occurs in the corpus of D_i is $P(D_i \mid W)$:

$$P(D_i \mid W) = \frac{count(W, D_i)}{count(W)} \quad (2)$$

From statistical information of domain comparative corpora (Fig. 1), we discovered that lengths of different categories are different greatly. To discriminate sizes of different categories, $P(D_i \mid W)$ is normalized by L_i (length of D_i):

$$P^{'}(D_i \mid W) = \frac{P(D_i \mid W)/L_i}{\sum_{j=1}^{m}[P(D_j \mid W)/Lj]} \quad (3)$$

Where L_i is computed by the sum of documents lengths in D_i. The NCI is defined as:

$$NCI(W) = -\sum_{i=1}^{m} P^{'}(D_i \mid W)\log P^{'}(D_i \mid W) \quad (4)$$

From the attribute of entropy impurity, the larger NCI of a word is, the more domain-irrelevant it is.

2.4. Normalized domain impurity (NDI) of a word

The occurrence distribution of a word within a domain is computed by NDI. Suppose documents in domain D_i are $d_1, d_2, ..., d_{n_i}$ successively. The probability of a word *W* occurs in the document d_j of D_i is $P(d_j \mid W)$:

$$P(d_j \mid W) = \frac{count(d_j, W)}{count(W, D_i)} \quad (5)$$

From the statistical information of a certain domain corpus, documents lengths within a domain vary greatly. So normalization is also used:

$$P^{'}(d_j \mid W) = \frac{P(d_j \mid W)/l_{ij}}{\sum_{j=1}^{n_i} P(d_j \mid W)/l_{ij}} \quad (6)$$

Where l_{ij} is length of d_j in D_i. The NDI is defined as:

$$NDI(W, D_i) = -\sum_{j=1}^{n_i} P^{'}(d_j \mid W)\log P^{'}(d_j \mid W) \quad (7)$$

2.5. Trade off between NCI and NDI

Words whose NCI value is lower than NCI threshold and NDI is larger than NDI threshold are extracted. And probable related domains are assigned for every extracted word. A trade off ranking score for an extracted word to its related domain D_i is assigned by the following formula:

$$RS(W, D_i) = -\alpha NCI(W) + (1-\alpha) NDI(W, D_i) \quad (8)$$

From a repetitive optimizing process, preferable value for α (0.5) is obtained.

3. Application and evaluation in text classification

Domain-specific terms extracted are evaluated by text classification system [11]. This text classification system is a classifier based on k-nearest neighbor. It uses vector space model (VSM) to map documents to feature space. In order to compress the dimension of documents vectors in VSM and map documents into semantic feature space, the latent semantic indexing is used to VSM. Semi-discrete matrix decomposition (SDD) is used in matrix decomposition for LSI. SDD requires less storage space and less executive time comparing with popular used singular value decomposition (SVD) method. The performance of this system ranks top first in evaluation of the national 863 project in both 2003 and 2004.

To evaluate the domain-specific terms extracted in text classification is an automatic evaluation method. That is to regard terms extracted as feature space which represents documents in text classification. Traditional well known methods for feature selection include TFIDF, expected cross entropy (ECE), mutual information (MI), the weight of evidence for text (WE), etc. The TFIDF measure neglects the relation between terms and categories. Other measures neglect the second principle in the section 2.1 and lead to some mistakes. Though in [12] the second principle is thought to be important to feature selection and tested by several popular used classifiers, it only considers the amounts of documents that a word occurs within a domain and omits distribution information of a word within a domain.

4. Experiment and result

4.1. Experimental corpora

Chinese corpora for text classification in evaluation of the national 863 project are used for this experiment. They are Chinese library classification corpora. The training data of text classification includes 36 categories and 250 documents for one category averagely.

From statistical results, sizes of different categories differ greatly. Corpora sizes of all categories ranges from 48687 to 682508 words. Fig. 1 shows the size information of categories. For every category, documents lengths distribution is also unbalanced. For example, there are 96 documents in social science category. Documents lengths range from 41 to 10251 words. Fig. 2 presents the documents lengths distribution of the social science category.

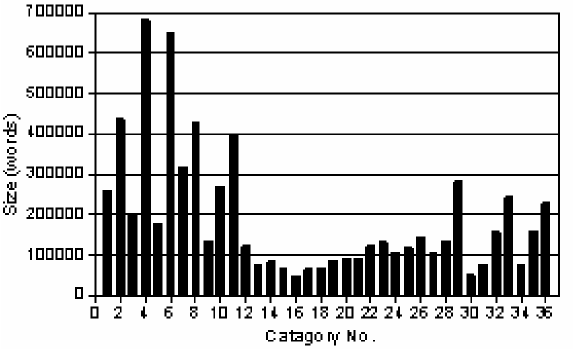

Fig. 1: Distribution of categories sizes

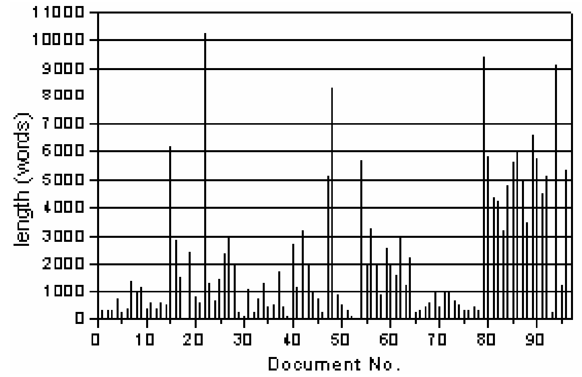

Fig. 2: Distribution of documents lengths

4.2. Domain-specific term extraction

Training corpora of text classification are used as domain comparative corpora for domain-specific terms extraction. Firstly, corpora are segmented by forward maximum match segment algorithm. Secondly, domain-specific terms are extracted from preprocessed domain comparative corpora. Different amounts of terms are extracted from different categories. Domain-specific terms extracted can be viewed at http://www.insun.hit.edu.cn/~tliu/.

From the result, we discover that some words which have no distinct domain feature and maybe their domain labels are ambiguous. But from a statistical point of view, they have domain-specific features and are extracted. Are they indeed domain-specific? Only the domain linguists can tell. Larger corpora can make the results more reliable.

4.3. Comparative experimental results in text classification

The proposed method is compared with both other feature selection methods (MI, WE, ECE and TFIDF) and other domain-specific term extraction methods (TFIDF, KFIDF and DR+DC). The comparative results are presented in Table 1. The proposed NCI+NDI method outperforms the best performed method TFIDF in traditional feature selection methods by 6.3 percent for F1measure. The proposed NCI+NDI method outperforms the best performed method DR+DC in domain-specific term extraction by 3.1 percent for F1measure. The proposed method outperforms other methods in precision, recall and F1 measure.

Table 1: Comparative results in open testing

Method	Precision	Recall	F1
MI	0.3541	0.3500	0.3520
WE	0.5575	0.5233	0.5399
ECE	0.5844	0.5686	0.5764
TFIDF	0.5998	0.5614	0.5800
KFIDF	0.6105	0.5980	0.6042
DR+DC	0.6209	0.6033	0.6120
NCI+NDI	0.6455	0.6411	0.6433

5. Conclusion and future work

An automatic method for domain-specific term extraction from domain comparative corpora is proposed. It takes two principles into account to characterize domain-specific terms. Impurity measure is used to measure words' distribution and normalization is added into it. Its evaluation in text classification indicates the proposed method is more effective than previous methods for extraction of domain-specific terms. Also the proposed method achieves better performance than traditional feature selection methods in text classification. This indicates replacing features by domain-specific terms is an effective way to text classification.

The proposed method only deals with existing words and can't recognize out of vocabulary (OOV) [13] words. Our future work will focus on how to discover the domain-specific OOV words.

6. Acknowledgement

This investigation was supported by the National Natural Science Foundation of China (60435020).

7. References

[1] Feiyu Xu, et al., "A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping," *Proc. Of the 3rd International Conference on Language Resources and Evaluation*, 2002.

[2] Paola Velardi, et al., "Identification of relevant terms to support the construction of Domain Ontologies," *Proc. Of ACL-01 workshop on Human language Technologies*, 2001.

[3] Wu, S.H., et al., "SOAT: A Semi-Automatic Domain Ontology Acquisition Tool from Chinese Corpus," *Proc. Of the 19 International Conference on Computational Linguistics*, 2002.

[4] Kietz, J.-U., et al., "Extracting a Domain-Specific Ontology Learning from a Corporate Intranet," *Proc. Of 2nd Learning Language in Logic (LLL) workshop*, Lisbon, 2000.

[5] L. F. Chien, "PAT-tree-based adaptive keyphrase extraction for intelligent Chinese Information retrieval," *ACM SIGIR'97*, Phliadelphia, USA, pp. 50-58, 1997.

[6] Yangarber, R., et al., "Automatic Acquisition of Domain Knowledge for Information Extraction," *Proc. Of the 18 International Conference on Computational Linguistics (COLING)*, 2000.

[7] Roberto. Basili, et al., "Empirical Investigation of fast text categorization over linguistic features," *Proc. Of the 15th European Conference on Artificial Intelligence (ECAI 2002)*, 2002.

[8] Miller G, WordNet: "An On-line Lexical Database," *International Journal of Lexicography*, 1990.

[9] HowNet, http://www.keenage.com

[10] Henri Avancini, et al., "Expanding Domain-Specific Lexicons by Term Categorization," *Proc. Of 18th ACM Symposium on Applied Computing*, pp. 793-797, 2003.

[11] Wang Qiang, et al., "A Study of Semi-Discrete Matrix Decomposition for LSI in Automated Text Categorization," *Proc. Of first International Joint Conference on Natural Language Processing (IJCNLP)*, 2004.

[12] Gongshen Liu, et al., "New feature Selection and weighting Methods Based on Category Information," *Proc. of the First National Conference on Information Retrieval and Content Security (NCIRC)*, 2004.

[13] Jun Zhao, et al., "Lexicon optimization for Chinese language modeling," *Proc. Of International Symposium Conference on Spoken Language Processing*, 2000.

[14] Richard O.Duda, et al., "Pattern classification (Second Edition)," China Machine Press, pp. 321-322, 2003.