E-015

A Study on Enhancing Vietnamese Text Classification by Using an Ontology

Nguyen GIANG SON   Shigeru OYANAGI

Graduate School of Science and Engineering
Ritsumeikan University, Biwako-Kusatsu Campus Noji Higashi
1 chome, 1-1 Kusatsu, 525-8577 Shiga-ken, JAPAN

1 Introduction

Text classification (TC) has many useful applications in organizing textual information. TC is resolved mainly by using machine learning techniques based on bag of words of document presentation. This technique uses only some techniques for building feature vectors like stemming, stop-word removing without semantically related . In this manuscript, to enhance text classification performance, an ontology based technique is used for generalizing or smoothing feature vectors semantically with pseudo hypernym words got from an raw ontology created automatically from a corpus. Therefore document feature vectors are closer semantically in case they have relationship. For example, word meat and vegetable has hypernym food vectors which contain meat and vegetable have relationship in terms of food. To apply this, it requires to have an ontology. In Vietnamese, there is no ontology. Hence techniques for building ontology automatically are necessary for Vietnamese language.

The manuscript is organized as following. Section 2 is for describing framework of TC with an ontology based. Section 3 describes issues when doing the experiments. Section 4 is for conclusion.

2 TC Framework with Ontology Based

Figure 1 describes the framework for TC with an ontology based. This idea is mainly based from [2] and is applied for Vietnamese appropriately.

Firstly, corpus must be segmented by a Vietnamese segmenter because Vietnamese words can not be separated by spaces. Then this segmented corpus is tagged shallowly by a Vietnamese tagger for building similar feature vectors of terms. This tagged corpus will be used for building the raw ontology automatically by the clustering algorithms such as Bisection Kmeans or agglomerative clustering.

After the raw ontology have been built, this ontology is used for generalizing feature vectors of documents by additional hypernym words in learning phase and classifying phase of TC.

Then document vectors are smoothed semantically by the raw ontology through adding hypernym in which documents have hyponym words. And they are used for the input of TC in testing an classification steps.

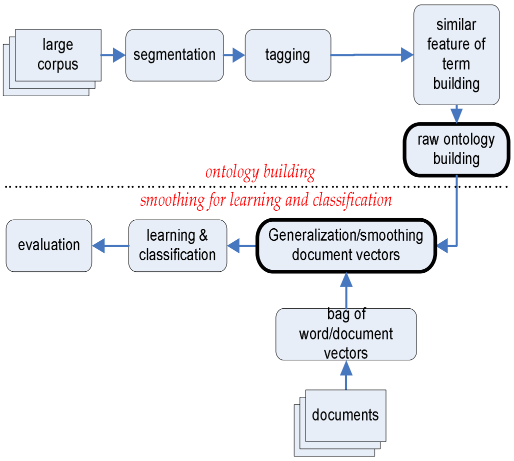
large corpus
segmentation
tagging
similar feature of term building
raw ontology building
ontology building
smoothing for learning and classification
evaluation
learning & classification
Generalization/smoothing document vectors
bag of word/document vectors
documents

Figure 1. TC framework with ontology based

2.1 Building the Raw Ontology

There are two steps in this process. One is for building similarity between terms and the other is for clustering term to have hierarchical concepts.

The idea of building Vietnamese ontology is based on distributional hypothesis that terms are semantically similar to the extent to which they share similar syntactic context. From that point of view, the similarity of terms are calculated by feature vectors built by around other terms syntactically related. This dependencies can be extracted by a large tagged corpus. This structure is mainly the same between English and Vietnamese. So we apply some dependent rules from [2] as following:

- adjective modifiers: *nhà(house) đẹp (beautiful)* → *nhà: đẹp*++ (weight of feature *đẹp* of term *nhà* is increased by 1)
- possessive modifiers: *cửa sổ (window) của(of) ngôi nhà(house)* → *ngôi nhà: has cửa sổ ++*
- prepositional phrase modifiers: *một lỗi (an error) trong(in) hệ thống (system)* → *lỗi: trong hệ thống++; lỗi: hệ thống trong++*
- noun phrase in subject or object position: *giá cả(price) cao(high) ảnh hưởng(affect) đến(to) đời sống(living)* → *giá cả: subj ảnh hưởng; đời sống: obj ảnh hưởng*
- copular constructs: *thịt lợn(pig meat) là(is) loại(type) thực phẩm(food) thiết yếu(essential)* → *thịt lợn: thực phẩm++*

- verb phrases with the verb have (có): *bữa ăn(meal) có(has) nhiều(many) món(dish) → bữa ăn: has món++*

More specifically, considering the following example of a paragraph:

Nhiều thành phố đã tích cực đẩy nhanh việc xây dựng các trung tâm chữa bệnh và thực hiện tốt chính sách cai nghiện và dạy nghề cho hàng trăm đối tượng nghiện ở địa phương mình.

So concept vectors are extracted*: thành phố: subj đẩy nhanh(1), subj thực hiện(1); trung tâm: chữa bệnh(1); chính sách: obj thực hiện(1), cai nghiện(1), dạy nghề(1); dạy nghề: cho đối tượng nghiện(1); đối tượng nghiện: obj dạy nghề(1), obj cai nghiện(1); ở địa phương(1); địa phương: đối tượng nghiện ở(1)*

After having these feature vectors of terms, similar measures of terms are calculated by using cosine measure between them. This similarity is used for building raw ontology by clustering terms from supervised bi-section Kmeans technique or hierarchical agglomerative technique. So we have a raw ontology which is a conceptual tree . This ontology will be used for generalizing feature vectors of documents in text classification process.

The ontology which is built in this step has no label on internal nodes. Leaf nodes are terms to be clustered. Here internal nodes are likely hypernym words of leaf term nodes. We can label these nodes like pseudo hypernym and use them for next smoothing step.

2.2 Generalization of Document Vectors

This step is called by semantically smoothing step for document vectors. The result of this step is that we have a conceptual like document vectors which are used for both testing steps and classifying steps. There are ambiguities between concepts and words in Vietnamese because Vietnamese word are multi-syllabus words. Vietnamese words can be one, two, three, or more syllabuses. It means words can be concepts and contrary. So the strategy for generalization of document vectors is simplified by searching hypernym words with a certain depth k from the ontology and add them to the document vectors. Then weight of these hybrid feature can be computed with standard TFxIDF term weighting method and used for classification algorithm.

2.3 Classification Algorithms Used for TC

Naïve Bayes classifiers and Linear Support Vector Machines which are machine learning techniques applied in TC are proposed for applying this approach. Naïve Bayes classifiers are mainly used as baseline for TC and it has quite a competitive results with Linear SVM in Vietnamese. Linear SVM are well-known in the field to have highest performance. To evaluate and also to compare their performance, micro-F1 measures, and macro F1 measures are employed.

3 Discussion

The applying an ontology for TC in English improves performance of classifiers as seen in [2]. However, there are difficult factors in processing Vietnamese language texts when considering to apply this technique to have higher performance in the TC:

- Domain of the corpus: Very large domain corpus like general news articles may have bigger ambiguity in concepts in comparison with specific domain like economic articles only.
- Tagger's quality: Actually, up to now, there have been little research on natural language processing for Vietnamese in general and specifically pos tagging problems. At present there is only one Vietnamese pos tagger which have been said to have performance of accuracy about 85 percent. Tagging is one of crucial factors to build a good ontology for other purposes.

Conclusion

In this manuscript, we propose an approach to apply ontology for Vietnamese TC. This approach has been applied in English and has promising results. The reason why it is considered to apply it for Vietnamese is due to its ability to build ontology automatically from a large corpus. This is reasonable because of abundant textual documents. The ontology is adaptively built with the domain of the corpus. So the semantic relationships of words are better than that of universal ontology like wordnet. Therefore the performance will be improved.

References

[1] Fabrizio Sebastiani. Machine learning in automated text categorization. In *ACM Computing Surveys*. 2002

[2] Stephan Bloehdorn, Andreas Hotho. Learning Ontologies to Improve Text Clustering and Classification. In *Proceedings of the 29th Annual Conference of the German Classification Society*, 2006.

[3] Hotho, A.; Staab, S.; Stumme, G. Ontologies improve text document clustering. In *ICDM 2003. Third IEEE International Conference on Volume*

[4] CIMIANO, P.; HOTHO, A. and STAAB, S. (2004): Comparing Conceptual, Partitional and Agglomerative Clustering for Learning Taxonomies from Text. In Proceedings of ECAI'04. IOS Press.