3F1-01

Set Expansion of Low-Frequency Semantic Relations

André Kenji Horie Mitsuru Ishizuka

The University of Tokyo

When concerning the task of classification of semantic relations, a small training dataset contains only a small portion of the possible morphological, syntactic and lexical-semantic features, rendering useless all the unknown features from the unlabeled data for the classifier. The present work thus presents a method for expanding the training dataset, providing a framework for the qualitative and quantitative analysis of the seed relation instances, extracting new relation candidates from the Web, and classifying them using spectral clustering.

1. Introduction

Extraction and classification of semantic relations is a task of utmost importance in the field of Semantic Computing, in which entities and links among these are defined so as to model a computer-comprehensible graph-like structure of the semantics of natural language texts. One initiative that defines these relations is CDL [1].

The most intuitive approach for this situation would be using fully supervised methods for the classification, such as proposed in [2] for frequent CDL relations. However, in the cases when dealing with relation classes that have few instances in the training data, whether it be because it may be too costly to produce this kind of data or because the relation class is itself rare, the classifier will overlook features that are only present in the relations to be classified.

Some works that aim to address this problem in a semi-supervised fashion have been proposed, but in different scopes. [3] uses feature vector extension to deal with discourse relations, and [4] uses set expansion and graph-based methods to deal with relations between entities that are bonded by an implicit lexical structure, which are as a result independent of their role in the sentence.

The scope of this work is relations whose semantics are strongly attached to their role within a given context, such as the ones in CDL. We first provide a framework for qualitatively and quantitatively analyzing the training dataset, and then propose a method for extracting candidate relations from the Web, and classifying them using spectral clustering, in order to expand the training data for a semi-supervised classification task.

2. Modeling Semantic Relations

A semantic relation, in the scope as described previously, may be characterized by morphological, syntactic and lexical-semantic features. This work uses as features part-of-speech tags, named entity tags and WordNet sense information for both head and tail entities, and shortest paths of the phrase structure and dependency tree of a sentence.

Moreover, in order to assess unknown features, the distance between any two features of the same type are defined as follows:

Contact: University of Tokyo, Hongo 7-3-1, Bunkyo, Tokyo, 113-8656 Japan. E-mail: andre@mi.ci.i.u-tokyo.ac.jp

- Part-of-speech tag: Levenshtein distance
- Phrase structure and dependency tree: Modified Levenshtein distance for arrays
- Named entity tag: Binary (0 if equal, 1 otherwise)
- WordNet sense: $\frac{min(\delta(s_1,cp),\delta(s_2,cp))}{\delta(cp,root)+min(\delta(s_1,cp),\delta(s_2,cp))}$

The distance between two relations instances R_i and R_j, $i,j \in [1,n]$, which is inversely proportional to the probability of them belonging to the same class, is a function d_{ij} of the distances among all features f_i^k and f_j^k of type k. Assuming a linear model, the following is observed:

$$d_{ij} = \beta_0 + \sum_{i=1}^{k} \beta_k \cdot \delta(f_i^k, f_j^k) \quad (1)$$

Considering binary classification, i.e. the problem will be broken down into several one-vs-all classification problems, we define Y as an $n \times n$ matrix of training relation instances whose element y_{ij} equals to 0 if both R_i and R_j belong to the given relation class, or 1 otherwise. Let then Y' be an m-dimensional array where each y' corresponds to one $y \in Y$. Let also $\beta = \{\beta_k\}$ be a k-dimension array, and $D = \{d_{ij}^k\}$ an $m \times k$ matrix. The values of β can be easily obtained using least squares multiple regression:

$$\beta = (D^T D)^{-1} D^T Y' = D^+ Y' \quad (2)$$

An $n \times n$ matrix $\mathcal{D}$ is defined as the distance matrix, and each $\mathcal{D}_{ij}$ corresponds to a d_{ij} from equation 1. It represents a unified view of all feature distance values, and is ideally block diagonal. From this matrix, several information on the quality of the feature set selection, comprehensiveness of the dataset, and response of the dataset to the chosen feature set can be obtained.

(a) BEN (b) PER

Figure 1: Distance matrix for two CDL relations

This framework provides, for instance, a graphical representation of the distance matrix (figure 1), and measures of prediction of clustering quality.

3. Semantic Relation Set Expansion

3.1 Relation Extraction

The initial step of the semantic relation set expansion is the extraction of candidate relations from the Web. First, search engine queries are made from a set of initial seed relations, preserving only the essential lexical structure between head and tail entites, and assigning a wildcard to each of them. For example, the relation flew→Japan in "he flew to Japan" generates the queries "flew to *" and "* to Japan".

Next, the candidate relations are identified by pattern-matching, using patterns generated from the initial seed relations. The patterns used in this work are of syntactic nature, generated from the phrase structure of the sentence and from the POS tags of both head and tail entities.

3.2 Clustering and Classification

The final step is the classification of the candidate relations generated from the previous step. This is accomplished by clustering both positive and negative relations from the training dataset for a given relation class, and classifying based on the results of this clustering.

For the proposed method, spectral clustering [5] is carried upon the distance matrix presented in the previous section, and two different classification methods are carried. The first is based on the distance to the closest cluster, and whose confidence measure is expressed in eq. 3, and the second uses a one-vs-all SVM, using [6] as confidence. The baseline method [2], which uses feature vectors and SVM, is also used for comparison purposes.

$$Confidence = 1 - \frac{d(R_C, C_{closest})}{\sum d(R_C, C)} \quad (3)$$

The confidently classified relations may be used as seeds of a new iteration of a bootstrapped set expansion process.

4. Experimental Results

The experimental setting is composed of a dataset which contains at most 10 relations for 28 CDL relation classes randomly selected from annotated articles from Wikipedia. The negative relations used in the clustering step will be relations from classes similar to the given class, information which is available in the CDL specification.

The measures used herein are macro-average accuracy and precision, which are unweighted averages of accuracy and precision values for all relation classes. Recall will not be considered in this work, since candidate relations marked as negative will not be used for the set expansion. Results are displayed in figure 2. The observed average growth of the positive relations set is 687% from the initial set.

5. Conclusion

By generating the distance matrix, analyzing the quality of the feature set and training data becomes more feasible, which is desirable when dealing with small training datasets, especially since these may be used as seed of a semi-supervised bootstrapped classification. The experiment shows that the proposed methods have better results than the baseline, with spectral clustering being the best.

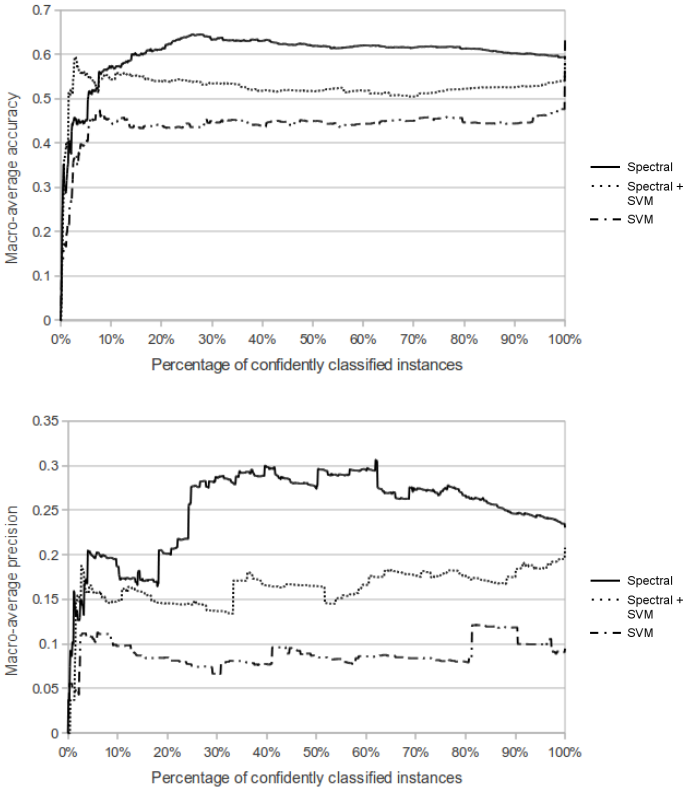

Figure 2: Macro-average accuracy and precision values

References

[1] Yokoi, T., Yasuhara, H., et al. *CDL (Concept Description Language): A common language for semantic computing*. In: WWW2005 Workshop on the Semantic Computing Initiative (2005).

[2] Yan, Y., Matsuo, Y., et al. *Annotating an Extension Layer of Semantic Structure for Natural Language Text*. In: Proc. of IEEE ICSC 2008.

[3] Hernault, H., Bollegala, D. and Ishizuka, M. *A Semi-Supervised Approach to Improve Classification of Infrequent Discourse Relations using Feature Vector Extension*. In: Proc. of EMNLP 2010.

[4] Li, H., Bollegala, D., et al. *Using Graph Based Method to Improve Bootstrapping Relation Extraction*. In: Proc. of CICLing 2011.

[5] Kannan, R., Vempala, S. and Vetta, A. *On Clusterings: Good, bad and spectral*. Technical report, Yale University (2000).

[6] Platt, J. C. *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*. In: Advances in Large Margin Classifiers, pp. 61-74 (1999).