Applying a multitask feature sparsity method for the classification of semantic relations between nominals

Guoqing Chao, Shiliang Sun

Department of Computer Science and Technology, East China Normal University
500 Dongchuan Road, Shanghai 200241, P.R. China
E-mail: guoqingchao10@gmail.com, slsun@cs.ecnu.edu.cn

Abstract:

This paper extracts seven effective feature sets and reduces them to same dimension by principle component analysis (PCA), such that it can utilize a multitask feature sparsity approach to the automatic identification of semantic relations between nominals in English sentences under maximum entropy discrimination (MED) framework. This method can make full use of related information between different semantic classifications to perform multitask discriminative learning and don't employ additional knowledge sources. At SemEval 2007, our system achieved a F-score of 69.15 % which is higher than that by independent SVM.

Keywords:

maximum entropy discrimination; support vector machine; semantic relation; multitask learning

1. Introduction

Following the fast development of Internet, the information obtained from Internet increases rapidly, which makes how to automatically get the useful information become more meaningful. Semantic relation classification is one fundamental work for that goal. The sentence "*list all x that causes cancer*" implies one semantic relation cause-effect(x,cancer). If we have identified such a relation, we can automatically search for the causes of cancer from Internet, which is appealing to many people.

In fact, automatic recognition of semantic relations has many applications such as information retrieval, information extraction, text summarization, question answering and so on. But there exist several challenges: extracting what features, how to effectively extract these features, what classification algorithm is the best. There is no doubt that addressing these problems is of great significance.

This paper will focus on two aspects: feature extraction and classification algorithm selection.

As to feature extraction, many researchers extracted different features which involved lexical, syntactic, and semantic knowledge. [1] selected 18 features that covered nearly every area of NLP, [2] used the hit counts from web search engines to obtain the lexical feature information. [3] utilized six features, which were also effective. In our system, we followed four feature sets that appeared in previous articles and presented feature sets 2, 3 and 7 that will present in Section 2. Especially the feature sets 7 changes this classification into one binary classification which is better to be solved with more sophisticated methods.

Speaking of classification algorithm, we can think of support vector machine (SVM), Bayesian optimal classifier, maximum entropy (MaxEnt) classifier, k-nearest neighbor algorithm (kNN), conditional random field and so on. But previous work rarely used multitask learning for the classification of semantic relations between nominals, We make some attempts in our paper. Maximum entropy discrimination (MED) is a general framework for discriminative estimation based on the maximum entropy principle, which was firstly presented by Jaakkola in [4]. Lately, Tony Jebara extended it to multitask learning in [5],[6]. Multitask learning is an effective machine learning method to take advantage of the information contained in related tasks to improve the generalization performance. Therefore, multitask learning [8][9][10] can bring better performance than single task learning.

In SemEval 2007 task 4[1], almost all the participants adopted SVM and MaxEnt classifiers. Experimental results also verified both methods are effective and useful. In our paper, we adopt the multitask feature sparsity method which makes full use of the advantages of multitask learning and MaxEnt principle. The experimental results on the SemEval

[1] http://nlp.cs.swarthmore.edu/semeval/tasks/

2007 task 4 dataset show that this method achieved better performance than independent SVM learning.

The rest of this paper is organized as follows. Section 2 firstly introduces SemEval 2007 task 4, and then describes feature extraction on classification of semantic relations. Section 3 reviews multitask feature sparsity via MED. Section 4 demonstrates how to apply multitask feature sparsity for semantic relation classification. Section 5 shows the experiment and its results. Section 6 concludes this work and points out the future work direction.

2. SemEval 2007 task 4 and feature extraction

Since this paper mainly directs at semantic relation classification of SemEval 2007 task 4, the following parts will introduce SemEval 2007 task 4, and then extracts seven effective feature sets.

2.1. SemEval 2007 task 4

The task 4 of SemEval 2007 defines seven sematic relations including *cause-effect, instrument-agency, product-producer, origin-entity, theme-tool, part-whole, content-container*. For each relation, the dataset contains 140 training examples and about 70 test examples. The following is one training example for cause-effect relation:

127 "I find it hard to bend and reach and I cannot use the $< e1 >$*cupboards*$< /e1 >$ *in my* $< e2 >$*kitchen*$< /e2 >$*."*
WordNet(e1) = "cupboard$\%1 : 06 : 00 ::$*",*
WordNet(e2) = "kitchen$\%1 : 06 : 00 ::$*",*
Content-Container(e1, e2) = "false",
*Query = "the * in my kitchen",*
Comment: Located-Location or, better, Part-Whole.

The first two lines include the sentence itself, preceded by a numerical identifier. The two nominals, "cupboards" and "kitchen", are marked by $< e1 >$ and $< e2 >$ tags. The third, fourth and fifth lines give the WordNet[2] sense keys for the two nominals and indicate whether the semantic relation between the nominals is a positive ("true") or negative ("false") example of the content-container relation. The sixth line gives the query that is used to find the sentence (mostly by searching on Google). The queries are manually generated heuristic patterns that are intended to find sentences that are examples of the given relation. The last line is an optional comment line to explain their labeling decisions.

The following is a testing example:

127 "I find it hard to bend and reach and I cannot use the $< e1 >$*cupboards*$< /e1 >$ *in my* $< e2 >$*kitchen*$< /e2 >$*."*
WordNet(e1) = "cupboard$\%1 : 06 : 00 ::$*",*
WordNet(e2) = "kitchen$\%1 : 06 : 00 ::$*",*
Content-Container(e1, e2) = "?",
*Query = "the * in my kitchen".*

In comparison with the training example, note that the relation, content-container$(e1, e2)$, is labeled "?", instead of "true" or "false". For all testing examples, the relations are labeled "?". Also, the comment lines have been removed for all testing examples.

The challenge of SemEval 2007 task 4 is to learn how to automatically distinguish the positive and negative examples, and this needs to extract effective features and choose appropriate classification algorithm.

2.2. Feature extraction

In order to accomplish SemEval 2007 task 4, we extract seven feature sets based on lexical-syntactic and semantic information from the sentences in which the two nominals located. Let e1 and e2 be two nominals appeared sequentially in the sentence. In addition, some terminologies related to WordNet can refer to WordNet documentation[3].

Feature set 1: Lemma of e1 and e2.
The lemma of the two nominals, which carry much information to help classify their relation, actually indicate two entity forms after stemming the two nominals .

Feature set 2: Stem words with specified part of speech (POS) between e1 and e2.
Firstly make POS of each word between two nominals, and then choose the preposition and verb to stem.

Feature set 3: Stem words with specified POS out of e1 and e2.
Similar with feature set 2, this feature set extracts the preposition and verb out of e1 and e2 but still inside the sentence.

Feature set 4: WordNet semantic class of e1 and e2.
Nominals are classified into 26 classes by their semantics in WordNet. We preprocess the nominals e1 and e2, and then obtain their semantic classes from the WordNet. Here, our system simply used the first noun senses of the nominals, that's because of the high cost of word sense disambiguation.

Feature set 5: Meronym-holonym relation between e1 and e2.
WordNet3.0 provides meronym and holonym information for some nouns. These information are quite important for part-whole relations. If there is a same word between the holonym set of e1 and the synonym set & hepernym set of e2, this will make the binary feature be "1". After that, we exchange the position of e1 and e2 and perform the same processing.

[2] http://wordnet.princeton.edu/wordnet/download/current-version/

[3] http://wordnet.princeton.edu/wordnet/documentation/

Feature set 6: Hyponym-hypernym relation between nominal and the word of "container".
This feature is designed for content-container relation. For each nominal, WordNet returns its hypernym set. Then the system examines whether the hypernym set contains the word "container". The result leads to a binary feature.

Feature set 7: Position arguments of e1 and e2.
Position arguments indicate the order they appeared in the sentence. Note that the order of both nominals is very important. For example, cause-effect(e1,e2) is different from cause-effect(e2,e1). The former indicates e1 is the cause and e2 is the effect but the latter means e2 is the cause and e1 is the effect. Therefore we use the position arguments indicate the positions of the arguments in the semantic relation.

3. Multitask feature sparsity via MED

Multitask feature sparsity via MED is an effective multitask learning method which is presented by Jebara in [6]. Now let's review MED and its extended multitask feature sparsity. Section 3.1 explains what the MED framework is. Section 3.2 makes further assumption on the form of the likelihood function to generate multitask feature sparsity.

3.1. MED

Maximum entropy principle has been successfully used in natural language processing area, and SVM also shows its powerful advantages to address classification problems. MED not only obeys the maximum entropy principle, but also yields both accurate classification and large margins as SVM do when to predict the label of a new query.

In essence, MED works similar with Bayesian method to some extent. The standard Bayesian approach to inference begins with a prior $p(\Theta)$ over a model class Θ. Given the data, the posterior is obtained by Bayesian rule $p(\Theta|D) \propto p(D|\Theta)p(\Theta)$. Subsequently, the posterior is used to predict for new observations.

But different from Bayesian method, MED constructs a posterior which produces predictions with large margin and accurate classification, which are key considerations in SVM. In order to achieve this goal, it forces the marginal likelihood of the correct label $y_{m,t}$ to larger than that of incorrect labels for each observation $t = 1, ..., T_m$ in all $m = 1, ..., M$ data sets with a margin.

MED finds a posterior as close as possible to the prior in terms of Kullback-Leibler divergence, and also subjects to the above constraints. MED formulates as follows:

$$\begin{cases} \min_{p(\Theta|D)} \mathrm{KL}(p(\Theta|D) \parallel p(\Theta) \\ s.t. \int \log\Big(\dfrac{p(y_{m,t}|x_{m,t},\Theta_m)}{p(y|x_{m,t},\Theta_m)}\Big) p(\Theta|D) d\Theta \geqslant \gamma \\ \quad \forall y \neq y_{m,t}, m, t. \end{cases} \tag{1}$$

According to the theorem in [4], the following posterior is obtained:

$$p(\Theta|D) = \frac{1}{Z(\lambda)} P(\Theta) \prod_{m=1}^{M} \prod_{t=1}^{T_m} \prod_{y \neq y_{m,t}} \Big(\frac{p(y_{m,t}|x_{m,t},\Theta_m)}{p(y|x_{m,t},\Theta_m)}\Big)^{\lambda_{m,t}} \exp(-\gamma\lambda_{m,t}). \tag{2}$$

Here, λ is a collection of non-negative Lagrange multipliers $\{\lambda_{m,t}\}$ for $m = 1, ..., M$ and $t = 1, ..., T_m$ that are used to enforce the inequality constraints. $Z(\lambda)$ is the normalizer for the above posterior. Lagrange multipliers are obtained by maximizing $J(\lambda) = -\log Z(\lambda)$.

MED makes predictions for a new query point as follows:

$$\hat{y} = \operatorname*{argmax}_y \mathrm{E}_{p(\Theta|D)}[\log p(y|x,\Theta_m)]. \tag{3}$$

Here, for computational convenience, it uses log-likelihood rather than likelihood.

3.2 Multitask feature sparsity

In order to couple multiple tasks, the likelihood function is needed to be modified to rely on a shared variable s (reference to [6]) as follows:

$$p(y|x,\Theta_m,s) \propto \exp\Big(\frac{y}{2}\Big(\sum_{d=1}^{D} s(d)x(d)\theta_m(d) + b_m\Big)\Big). \tag{4}$$

Here, s indicates whether it would choose its according entry of x. Rewrite the posterior $p(\Theta|D)$ more specially as:

$$p(\Theta|D) = \frac{1}{Z(\lambda)} p(\Theta) \prod_{m=1}^{M} \prod_{t=1}^{T_m} \exp\Big(\lambda_{m,t} y_{m,t} \Big(\sum_{d=1}^{D} s(d)x_{m,t}(d)\theta_m(d) + b_m\Big) - \gamma\lambda_{m,t}\Big). \tag{5}$$

Make some assumptions to obtain the following formula:

$$\begin{cases} \max_\lambda \sum_{m=1}^{M} \sum_{t=1}^{T_m} \gamma\lambda_{m,t} - \sum_{d=1}^{D} \log \\ \quad (\alpha + e^{\frac{1}{2}\sum_{m=1}^{M}(\sum_{t=1}^{T_m} \lambda_{m,t} y_{m,t} x_{m,t}(d))^2}) \\ \quad + D\log(\alpha+1) \\ s.t.\ 0 \leq \lambda_{m,t} \leq C\ \forall m,t \\ \quad \sum_{t=1}^{T_m} y_{m,t}\lambda_{m,t} = 0\ \forall m. \end{cases} \tag{6}$$

Obviously, the objective function is no longer additive across $m = 1...M$ which means that learning is coupled across tasks.

When the λ setting has been obtained, the formula (3) is used to predict the label of a new query.

4. Applying multitask feature sparsity for semantic relation classification

For seven semantic relations, we extract similar feature sets to make the seven tasks possible to benefit from multitask learning. As we all know that there are many multitask learning methods, why do we select this one to classify the semantic relations? Because multitask feature sparsity method posses two merits of large margin and accuracy classification, which are key considerations in SVM. And the experimental results obtained by SVM is comparably good in all classification methods.

In order to apply multitask feature sparsity method, we need to run dimensionality reduction by principle component anylysis (PCA) to make all the features to have the same dimension. Note that, running PCA is to utilize the multitask feature sparsity method not to speed up the classification. Since running PCA may cause performance reduction, we run PCA to do dimensionality reduction with a premise to keep performance as much as possible.

5. Experiment

We used the dataset from SemEval 2007 task 4 for experiment. The performance measures will be P (precision), R (recall), and F (the harmonic mean of precision and recall). F is calculated according to the following formula: $F = \frac{2*P*R}{P+R}$.

In the extreme case, if P is large, R can be very small, and vice versa, which means sometimes both measures are unbalanced. Therefore we primarily use F to measure the classification performance.

In order to use the multitask feature sparsity method via MED, we firstly reduce the features of seven relations to 200 dimension by PCA. Furthermore, to make the performance comparison, we also do the experiments with independent SVM in the same experimental environment including 200 dimensional features obtained by PCA, the experimental results are provided in Table 1 and Table 2.

From the Table 1 and Table 2, we can find the primary performance measure F-score of multitask feature sparsity via MED is higher 0.85% than that of independent SVM. The other measures precision and recall also favors multitask feature sparsity via MED. Considering the classification accuracy of almost all the seven semantic relations improved, and when the performance amounts some extent, the classification accuracy will be hard to improve, our work is important and valuable.

Table 1. The experimental results with the independent SVM.

Sem-Relation	P	R	F
Cause-Effect	0.6078	0.756	0.6739
Instrument-Agency	0.6364	0.7368	0.6829
Product-Producer	0.6912	0.7581	0.7231
Origin-Entity	0.6047	0.7222	0.6582
Theme-Tool	0.6111	0.7586	0.6769
Part-Whole	0.5385	0.8077	0.6462
Content-Container	0.5807	0.9474	0.72
Average	0.6100	0.7838	0.6830

Table 2. The experimental results with the multitask feature sparsity via MED

Semantic-Relation	P	R	F
Cause-Effect	0.6188	0.756	0.6806
Instrument-Agency	0.6230	0.7605	0.6849
Product-Producer	0.6948	0.7710	0.7309
Origin-Entity	0.6162	0.7222	0.6650
Theme-Tool	0.6416	0.7586	0.6951
Part-Whole	0.5385	0.8077	0.6462
Content-Container	0.6042	0.9474	0.7378
Average	0.6196	0.7891	0.6915

Without the help of the unlabeled examples, this method cannot achieve the state of the art. But it extracts simple features and makes full use of the useful information between all related tasks, producing large margin and accuracy classification. More important is that it produces the higher performance than independent SVM.

6. Conclusion and future work

In our paper, we only extract a few simple and effective features and don't use other knowledge sources, but we obtained considerable good result. In addition, we take full advantage of the similar relation of the seven semantic relations of nominals to achieve a better performance than independent SVM.

In future, we will determine more special feature to express every example to distinguish whether it belongs to cor-

responding relations. Moreover, owning to the high cost to get labeled examples, we can consider semi-supervised learning to improve the classification performance with the help of large number of unlabeled examples.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Project 61075005, and the Fundamental Research Funds for the Central Universities.

References

[1] B. Beamer, S. Bhat, B. Chee, A. Fister, A. Rozovskaya, and R. Girju. "UIUC: A knowledge-rich approach to identifying semantic relations between nominals". Proceedings of the 4th International Workshop on Semantic Evaluations, Prague, Czech Republic, pp. 386-389, 2007.

[2] P. Nulty. "UCD-PN: Classification of semantic relations between nominals using WordNet and web counts". Proceedings of the 4th International Workshop on Semantic Evaluations, Prague, Czech Republic, pp. 374-377, 2007.

[3] Y. Chen, M. Lan, J. Su, Z.M. Zhou, and Y. Xu. "ECNU: Effective semantic relations classification without complicated features or multiple external corpora". Proceedings of the 5th International Workshop on Semantic Evaluations, Uppsala, Sweden, pp. 226-229, 2010.

[4] T. Jaakkola, M. Meila, and T. Jebara. "Maximum entropy discrimination". Advances in Neural Information Processing Systems, 1999.

[5] T. Jebara. "Multi-task feature and kernel selection for SVMs". Proceedings of the 21th International Conference on Machine Learning, pp. 55, 2004.

[6] T. Jebara. "Multitask sparsity via maximum entropy discrimination". Journal of Machine Learning Research, pp. 75-110, 2011.

[7] F. Costello. "UCD-FC: Deducing semantic relations using WordNet senses that occur frequently in a database of noun-noun compounds". Proceedings of the 4th International Workshop on Semantic Evaluations, Prague, Czech Republic, pp. 370-373, 2007.

[8] S. Sun. "Multitask learning for EEG-based biometrics". Proceedings of the 19th International Conference on Pattern Recognition, Florida, USA, pp. 1-4, 2008.

[9] Y. Ji and S. Sun. "Multitask multiclass support vector machines". Proceedings of the 11th International Conference on Data Mining Workshops, Vancouver, Canada, pp. 512-518, 2011.

[10] R. Caruanna. "Multitask learning". Machine Learning, pp. 41-75, 1997.