A Comparative Study on Statistical Classification Methods in Relation Extraction

Xiaofeng Zhang[1], Zhiqiang Gao[2], Yaocheng Gui[3]

Abstract. This paper is a comparative study of statistical classification approaches in relation extraction and classification. The focus is on multiclass classification, not on sequence labeling. Five methods are evaluated, including naive Bayes (NB), decision tree (DT), k-nearest neighbor (kNN), support vector machine (SVM) and sparse network of Winnow (SNoW). Using DT on Roth and Yih data set, the best precision and recall are achieved on both tasks of named entity recognition (NER) and relation extraction (RE). SNoW is not so good as DT, but it performs better than the other approaches. SVM performs better on precision and worse on recall. In contrast, the simplest methods NB and kNN has relative poor performance but they are not sensitive to learning tasks and classes.

Keywords: Relation Extraction, Named Entity Recognition, Statistical Classification

1 Introduction

Information extraction (IE) is the task of extracting structured information from text. The two most common sub-tasks of IE are extracting named entities (NER) (like PER (person), LOC (location) and ORG (organization)) and extracting relations (RE) between them (like WORK_FOR which relates a person and an organization, ORGBASED_IN which relates an organization and a location etc.).

[1] Xiaofeng Zhang
School of Computer Science & Engineering, Southeast University
Jiangsu Administration Institute
e-mail: coldrain521@gmail.com
[2] Zhiqiang Gao (✉)
School of Computer Science & Engineering, Southeast University
e-mail: zqgao@seu.edu.cn
[3] Yaocheng Gui
School of Computer Science & Engineering, Southeast University
e-mail: yaochgui@seu.edu.cn

The focus in this paper is the evaluation and comparison of statistical classification methods in RE. We select five existing classifiers which have already been widely used in machine learning and data mining communities, which are naive bayes (NB), decision tree (DT), k-nearest neighbor (kNN), support vector machine (SVM), and sparse network of Winnow (SNoW). We seek answers to the following questions with empirical evidence: 1) What are the strengths and weaknesses of existing statistical classification methods applied in RE? 2) What causes the strengths and weaknesses of statistical classification methods applied to RE?

For clarity, in this paper we make the following assumptions: 1) Advanced features are not used, such as syntactic features of parsing tree and dependency relation, NP chunking, semantic features such as WordNet hypernym [Fellbaum, 1998], as well as various knowledge bases [Zhou et al., 2005] such as Wikipedia[4]. Additionally, kernels are not employed to combine different information sources [Alicante & Corazza, 2011]. 2) We focus on the task of classification of semantic relations, which is of assigning to each semantic relation a label taken from a finite set, and we don't assume that pairs of related entities are given together with their labels. 3) Although not all relations labels are compatible with every pair of entity types, such constraints are not used in the considered data set [Roth & Yih, 2004]. 4) We choose statistical classification approaches, and we do not consider sequence labeling approaches.

Major contributions of this paper include: 1) Up to our knowledge this is the first time to compare various statistical classification methods applied in RE. 2) It is found that DT performs best in both tasks of NER and RE, SNoW is not so good as DT but it performs better than the other approaches.

The paper is organized as follows. Sec. 2 presents the related works, especially those on the Roth and Yih data set. Details on the problem definition and various statistical classification approaches applied in RE are given in Sec. 3. Experiment settings, feature selection and experimental results as well as analysis are discussed in Sec. 4, while conclusions are presented in Sec. 5.

2 Related Works

Lots of works have been devoted to RE and classification. We are giving a preference here to systems which have been assessed on the Roth and Yih data set [Roth & Yih, 2004]. Systems assessed on other data sets include [Beamer et al., 2007] [Davidov & Rappoport, 2008] for SemEval 2007, [Rink & Harabagiu, 2010] for SemEval 2010 and [Kambhatla, 2004] [Zhou et al., 2006] [Qian et al., 2008] for the Automatic Content Extraction (ACE), which is not freely available.

Kate and Mooney presented an approach for mapping natural language sentences to their formal meaning representations using string kernel based classifiers

[4] http://www.wikipedia.org/

[Kate & Mooney, 2006]. Miller et al. adapted a probabilistic context-free parser for information extraction [Miller et al., 2000]. Giuliano et al. thoroughly evaluated the effect of NER on RE [Giuliano et al., 2007]. Alicante and Corazza proposed barrier features, such as N-Grams of POS (Part of Speech), word prefixes and suffixes, hypernyms from WordNet etc. [Alicante & Corazza, 2011]. Carlson et al. presented a method to simultaneously do semi-supervised training of entity and relation classifiers [Carlson et al., 2009].

Most systems usually first extract entities and afterwards relations. An important exception to this two pass approach is represented by [Roth & Yih, 2004, 2007]. Their methods first identified the possible entities and relations in a sentence using separate classifiers which were applied independently and then computed a most probable consistent global set of entities and relations using integer linear programming. Kate and Mooney presented a method for joint entity and RE using a graph called a "card pyramid" [Kate & Mooney, 2010]. They also gave an efficient algorithm that is analogous to parsing using dynamic programming.

In summary, existing RE researches are focused on selecting advanced syntactic or semantic features, or integrating NER with RE by integer linear programming or card pyramid parsing. No researches have been published on comparing different statistical classification approaches applied in RE.

3 Statistical Classification Approaches Applied in Relation Extraction

In this section, we first define the task of RE, and then we introduce the five statistical classification approaches compared in this paper.

3.1 Problem Definition

A sentence of length N is a string of N words or entities $\{E_1, E_2, ...E_N\}$, each corresponding to a sequence of consecutive tokens, that is a substring of S. The entity indexes follow their order in the sentence, and each entity is labeled by an entity-type in a finite set E of labels. A subset of all ordered entity pairs corresponds to relations: $R_{i,j} = (E_i, E_j)$; E_i is called agent and E_j target, where the entities E_i and E_j can be composed by one or more tokens of the sentence and E_i can either precede or follow E_j. A label taken from a finite set R of possible labels is associated to each relation. We are considering the task of associating the correct label to each relation, which is the classification task of semantic relations.

3.2 Statistical Classification Approaches

In this section, we introduce the five statistical classification approaches, as well as their experiment settings.

Decision Tree

Decision tree (DT) learning is one of the most widely used and practical methods for inductive inference. Our experiments for decision tree are based on J48, the open source Java implementation of the C4.5 decision tree learning algorithm in the Weka data mining tool[5] with its default settings.

Support Vector Machine

Support vector machines (SVMs) are supervised learning models. An SVM model can efficiently perform a non-linear classification using kernel trick. Our experiments for SVM are based on WLSVM[6]. We compared four basic kernels, which are: linear, polynomial, radial basis function (RBF), and sigmoid. The linear kernel achieves the best performance in the NER task, while in the RE task the RBF kernel is the best. In addition, we normalize the data for both tasks and use probability estimate for the RE task and leave other settings as default.

Naive Bayes

Naïve Bayes (NB) is a simple generative method based on applying Bayes' theorem with a conditional independence assumption. Our experiments for naive Bayes are based on the open source Java implementation of naive Bayes classifier in the Weka data mining tool. The best performance is achieved by using supervised discretization to convert numeric attributes to nominal ones.

k-Nearest Neighbor

k-nearest neighbor (kNN) is a non-parametric method for classifying objects based on closest training examples in the feature space. Our experiments for kNN are based on IBk, the open source Java implementation of kNN algorithm in the Weka data mining tool. The best performance is archived by choosing five neighbors for each testing example. Besides, the Euclidean similarity function is used to measure the distances, and the brute force search algorithm is used.

Sparse Network of Winnow

Sparse network of winnows (SNoW) is a learning architecture framework that is specifically tailored for learning in the presence of a very large number of features and can be used as a general purpose multi-class classifier. The learning frame work is a sparse network of sparse linear functions over a predefined or incrementally acquired feature space. In our experiment, we use the implementation and default setting of SNoW by Dan Roth[7].

---

[5] http://www.cs.waikato.ac.nz/ml/weka/

[6] http://www.cs.iastate.edu/~yasser/wlsvm/

[7] http://cogcomp.cs.illinois.edu/page/software_view/1

4 Experiments

In this section, we first introduce the data set used in the comparison, and then the features are given. At last, experimental results are demonstrated and analyzed.

4.1 Data Set

For experimental assessment we use the data set used by Roth and Yih [Roth & Yih, 2004], derived from TREC corpus, which is freely available[8]. The sentences in this data were annotated with entities and relations. It includes three types of entities, namely PER (person), LOC (location) and ORG (organization), and five types of binary relations, namely WORK_FOR (work for), KILL (kill), LIVE_IN (live in), LOCATED_IN (located in) and ORGBASED_IN (orgbased in), in addition there is an extra type, OTHER, indicates that the entity is of none of the given types. Similarly, there is an extra relation type, NR, indicates that its two entity arguments are not related. The Roth and Yih data set is not divided in training and test set. Therefore assessment is performed by following the 5-fold cross validation protocol, as in [Giuliano et al., 2007] [Roth & Yih, 2007] [Kate & Mooney, 2010].

As in the previous work with this data set, in order to observe the interaction between entities and relations, our experiments used only the 1441 sentences that include at least one relation. Note, the total number of sentences is 5516. The boundaries of the entities are already supplied by this data set. The number of three types of entities is: PER (1691), LOC (1968) and ORG (984), in addition there is a fourth type OTHER (706). The number of five types of relations is: LOCATED_IN (406), WORK_FOR (401), ORGBASED_IN (452), LIVE_IN (529) and KILL (268). There are 17032 pairs of entities that are not related by any of the five relations and hence have the NR relation between them which thus significantly outnumbers other relations.

4.2 Feature Selection

We use the following standard entity extraction features: the word form and part-of-speech (POS) tag sequence of the candidate entity words, two words before and after the candidate entity and their POS tags, whether any or all candidate entity words are capitalized, whether any word has suffix “ment” or “ing”. We used totally 30 features for NER, which are listed in **Table 1**.

[8] http://l2r.cs.uiuc.edu/?cogcomp/Data/ER/conll04.corp

Table 1: Features used in NER. The numbers indicates the offset between the current word and the one being classified

ID	Feature	ID	Feature	ID	Feature
1	word (-2)	11	POS(+1)	21	POS(-2)+POS(-1) + POS (0)
2	POS(-2)	12	word (+1) + POS (+1)	22	word (0)+ word (+1) + word (+2)
3	word (-2)+ POS (-2)	13	word (+2)	23	POS(0)+POS(+1) + POS (+2)
4	word (-1)	14	POS(+2)	24	whether the initial letter is upper case in word (0)
5	POS(-1)	15	word (+2) + POS (+2)	25	whether some letters are upper case in word (0)
6	word (-1) + POS (-1)	16	word (-1) + word (0)	26	whether all letters are upper case in word (0)
7	word (0)	17	POS (-1) + POS (0)	27	whether "ing" is the suffix of word (0)
8	POS(0)	18	word (0) + word (+1)	28	whether "ment" is the suffix of word (0)
9	word (0)+ POS (0)	19	POS (0) + POS (+1)	29	the length of word (0)
10	word (+1)	20	word (-2)+ word (-1) + word (0)	30	whether "Lt." or "Gov. " is contained in word (0)

Nearly most relation extraction systems consider some form of parsing. The complete parse tree of the input sentence was considered by [Miller et al., 2000], [Kambhatla, 2004] and [Reichartz et al., 2009]. Systems only considered some form of shallow parsing include [Giuliano et al., 2007] and [Zhang et al., 2005]. However, we focus on the caparison of various classification approaches. So we do not use these parsing features in RE. We use the features of e1 and e2, as well as the features of e1 + e2. In addition, we also use some patterns of specific words contained before, between and after e1 and e2. So, totally 100 features are used in RE, as listed in **Table 2**.

Table 2: Features used in RE

ID	Feature	ID	Feature
1-30	features of e1	95	whether "in" is before e1 and "at" is between e1 and e2
31-60	features of e2	96	whether "at" is before e1 and "in" is between e1 and e2
61-90	features of e1 + features of e2	97	whether "at" is before e1 and "at" is between e1 and e2
91	the number of words between e1 and e2	98	whether "native of" is between e1 and e2
92	whether the word of e1 is the same as the word of e2	99	whether "based in" is between e1 and e2
93	whether e1 is at the beginning of a sentence	100	whether "based at" is between e1 and e2
94	whether "in" is before e1 and "in" is between e1 and e2		

4.3 Results and Analysis

Table 3 and **Table 4** show the results of NER and RE. The statistical significance is shown for precision, recall and F1-measures. We first note that all the results of our statistical classification approaches are not better than the approach of [Roth &

Yih, 2007], for both entities and relations. This is due to a lot of advanced features are not considered in our experiments as given in Sec. 1. Experimental result shows that among the entity and relation classifiers DT has the best performance. SNoW is comparable to the performance of DT. Although SVM has the best precision, however, due to its worst recall, its performance is not so good as expected. NB and kNN performs better than SVM in NER than RE. The average F1 on NER and RE of these approaches are: NB (47.3), DT (62.0), kNN (48.2), SVM (40.5), SNoW (57.2).

Table 3: Comparison of performance of various classifiers in NER

Entity	PER			LOC			ORG		
Approach	R	P	F1	R	P	F1	R	P	F1
NB	76.9	81.1	78.9	74.9	79.7	77.2	73.4	58.5	65.1
DT	82.1	82.3	**82.2**	81.7	78.8	80.2	69.0	71.4	**70.2**
kNN	73.0	67.2	70.0	76.9	67.7	72.0	51.7	68.9	59.1
SVM	57.9	65.7	61.6	75.9	53.7	62.9	41.0	67.4	51.0
SNoW	87.8	76.4	81.7	80.8	82.2	**81.5**	37.4	79.2	50.8

Table 4: Comparison of performance of various classifiers in RE

Relation	LOCATED_IN			WORK_FOR			ORGBASED_IN			LIVE_IN			KILL		
Approach	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1
NB	76.7	13.5	23.0	82.0	13.9	23.8	64.7	12.2	20.5	45.7	14.6	22.1	99.6	8.1	15.0
DT	44.8	61.6	**51.9**	35.1	40.2	**37.5**	43.1	53.4	47.7	35.0	43.6	**38.8**	56.2	56.6	**56.4**
kNN	29.2	62.1	39.7	10.2	37.1	16.0	33.8	75.7	46.7	21.3	48.5	29.6	9.1	39.5	14.8
SVM	11.7	73.4	20.2	4.2	61.0	7.9	26.0	92.4	40.6	16.8	76.6	27.6	9.1	51.7	15.5
SNoW	37.2	56.8	45.0	21.6	49.4	30.1	36.4	74.2	**48.8**	28.0	55.0	37.1	43.6	71.0	54.0

5 Conclusions

This work is an evaluation of statistical classification methods in RE and classification, from using the simplest NB and kNN, to advanced DT, SVM and SNoW. We found DT most effective in both tasks of NER and RE. SNoW is found comparable to the performance of DT. SVM has better precision compared to other methods due to a bias favoring precision, however its recall is too low. NB and kNN are not sensitive to learning tasks and data folds as the baseline approaches.

6 Acknowledgements

This paper is funded partly by the National Science Foundation of China under grant 61170165.

References

[Alicante & Corazza, 2011] Anita Alicante, Anna Corazza: Barrier Features for Classification of Semantic Relations. Proceedings of RANLP11, pages 509–514. 2011

[Beamer et al., 2007] Brandon Beamer, Suma Bhat, Brant Chee, Andrew Fister, Alla Rozovskaya, Roxana Girju: A Knowledge-rich Approach to Identifying Semantic Relations between Nominals. Proceedings of SemEval07. pages 386–389, 2007

[Carlson et al., 2009] Andrew Carlson, Justin Betteridge, Estevam R. Hruschka, Tom M. Mitchell: Coupling Semi-supervised Learning of Categories and Relations. Proceedings of the NAACL HLT09 Workshop on SSLNLP, pages 1–9, 2009

[Davidov & Rappoport, 2008] Dmitry Davidov, Ari Rappoport: Classification of Semantic Relationships between Nominals Using Pattern Clusters. Proceedings of ACL08, pages 227–235, 2008

[Fellbaum, 1998] Christiane Fellbaum, editor: WordNet: An Electronic Lexical Database. MIT Press, 1998

[Giuliano et al., 2007] Claudio Giuliano, Alberto Lavelli, Lorenza Romano: Relation Extraction and the Influence of Automatic Named-Entity Recognition. ACM Trans. Speech Lang. Process., 5(1), pages 1–26. 2007

[Kambhatla, 2004] Nanda Kambhatla: Combining Lexical, Syntactic,and Semantic features with Maximum Entropy Models for Information Extraction. Proceedings of ACL04, pages 178–181, 2004

[Kate & Mooney, 2006] Rohit J. Kate, Raymond J. Mooney: Using String-Kernels for Learning Semantic Parsers. Proceedings COLING/ACL06, pages 913–920, 2006

[Kate & Mooney, 2010] Rohit J. Kate, Raymond J. Mooney: Joint Entity and Relation Extraction using Card-Pyramid Parsing. Proceedings of CoNLL10, pages 203–212. 2010

[Miller et al., 2000] Scott Miller, Heidi Fox, Lance A. Ramshaw, Ralph M. Weischedel: A Novel Use of Statistical Parsing to Extract Information from Text. Proceedings of NAACL00, pages 226–233, 2000

[Qian et al., 2008] Longhua Qian, Guodong Zhou, Fang Kong, Qiaoming Zhu, Peide Qian: Exploiting Constituent Dependencies for Tree Kernel-based Semantic Relation extraction. Proceedings of the 22nd International Conference on Computational Linguistics – Vol. 1, COLING08, pages 697–704, 2008

[Reichartz et al., 2009] Frank Reichartz, Hannes Korte, Gerhard Paass: Dependency Tree Kernels for Relation Extraction from Natural Language Text. Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, vol. 5782, chapter 18, pages 270–285. 2009

[Rink & Harabagiu, 2010] Bryan Rink, Sanda Harabagiu: Classifying Semantic Relations by Combining Lexical and Semantic Resources. Proceedings of SemEval10, pages 256–259, 2010

[Roth & Yih, 2004] D. Roth, W. Yih: A Linear Programming Formulation for Global Inference in Natural Language Tasks. Proceedings of CoNLL04, pages 1–8. 2004

[Roth & Yih, 2007] D. Roth, W. Yih: Global Inference for Entity and Relation Identification via a Linear Programming Formulation. L. Getoor and B. Taskar, editors, Introduction to Statistical Relational Learning. The MIT Press. pages 553–580, 2007

[Zhang et al., 2005] Min Zhang, Jian Su, Danmei Wang, Guodong Zhou, Chew Lim Tan: Discovering Relations between Named Entities from a Large Raw Corpus Using Tree Similarity-based Clustering. Proceedings of IJCNLP05, pages 378–389. 2005

[Zhou et al., 2005] Guodong Zhou, Jian Su, Jie Zhang, Min Zhang: Exploring Various Knowledge in Relation Extraction. Proceedings of ACL05, pages 427–434, 2005

[Zhou et al., 2006] Guodong Zhou, Jian Su, Min Zhang: Modeling Commonality among Related Classes in Relation Extraction. Proceedings of ACL06, pages 121–128, 2006