Semantic Role Labeling of Chinese Using Transductive SVM and Semantic Heuristics

Yaodong Chen Ting Wang Huowang Chen Xishan Xu
Department of Computer Science and Technology, School of Computer,
National University of Defense Technology
No.137, Yanwachi Street, Changsha, Hunan 410073, P.R.China
{yaodongchen, tingwang, hwchen}@nudt.edu.cn xxs@hnmcc.com

Abstract

Semantic Role Labeling (SRL) as a Shallow Semantic Parsing causes more and more attention recently. The shortage of manually tagged data is one of main obstacles to supervised learning, which is even serious in SRL. Transductive SVM (TSVM) is a novel semi-supervised learning method special to small mount of tagged data. In this paper, we introduce an application of TSVM in Chinese SRL. To improve the performance of TSVM, some heuristics have been designed from the semantic perspective. The experiment results on Chinese Propbank showed that TSVM outperforms SVM in small tagged data, and after using heuristics, it performs further better.

1 Introduction

Semantic analysis is one of the fundamental and key problems for the research in computational linguistics. Traditional semantic research is mainly concerned with deep analysis, which provides a representation of the sentence in predicate logic or other formal specification. Recently, shallow semantic parsing is becoming a hotspot in semantic analysis research. Semantic Role Labeling is a shallow semantic parsing technology and defined as a shared task in CoNLL-04. It aims at recognizing semantic roles (i.e. arguments) for each target verb in sentence and labeling them to the corresponding syntactic constituents. Many SRL research utilizes machine learning methods (Park, 2005; Pradhan, 2005; Cohn, 2005), in which the high performance reported was attributed to large tagged dataset (Carreras, 2005). But one of the main obstacles to supervised learning is the shortage of manually labeled data, which is even serious in SRL. It could bring about one question: whether these methods perform well when large mount of tagged data are not available? In this paper, we investigate Transductive SVM (Joachims, 1999), a semi-supervised learning method, for this question. The proposed method uses large untagged data in training with the support of the linguistic knowledge of semantic roles.

Generally speaking, not all constituents in syntactic tree could act as argument candidates in SRL. Large redundant constituents lead to a high training cost and decrease the performance of statistical model especially when tagged data is small. In contrast to the pruning algorithms in Park (2005) and Xue (2004) which are based on syntax, some argument-specific heuristics, based on word semantic features of arguments, make semantic restrictions on constituent candidates to optimize dataset of statistical models. The experiment results on Chinese Propbank shows that TSVM outperforms regular statistical models in small tagged data, and after using argument-specific heuristics, it performs further better.

The rest of this paper is organized as follows. Section 2 gives the definition, method, and resources about SRL. Section 3 discusses how to apply TSVM for SRL. Some argument-specific heuristics are introduced in Section 4. And then, section 5 shows the experiment results of the proposed methods and compare it with SVM. Finally, we conclude our work in section 6.

2 Problem Definitions & Related Works

Comparing with full parsing, SRL acts on part of constituents in sentences in order to achieve high performance and robustness, as well as low complexity in practices. The SRL problem can be described as follows.

Definition Given a semantic role (or argument) collect *R* and a sentence *S*, for any substring *c* of *S*, SRL is a function: $c \rightarrow R \cup NONE$, where *NONE* is the value excluded in *R*.

Notice that *c* usually indicates phrases in a sentence. SRL can be classified to two steps:

- **Identification:** $c \rightarrow \{NONE, ARG\}$. It is a binary-value function where *ARG* is assigned to *c* when it should be labeled at some element of *R*, or *NONE* is assigned. Identification separates the argument substrings from the rest of sentence, in another words, finds the argument candidates.
- **Classification:** $c \rightarrow R$. It is a multi-value function which assigns a role value to *c*, that is, labels a role to some candidate.

Some typical systems, based on inductive learning, have been evaluated in CoNLL-05 (Carreras, 2005). It concluded that the performance of SRL depends on the combination of several factors including models, features, and results of syntactic parsing. The best result achieved F1=75.04 [1]. These systems have strong dependency on large tagged data. This paper evaluates the performance of a classical supervised learning method--SVM in small tagged data and introduces a novel semi-supervised method to handle this problem.

There are two tagged corpora available for SRL: one is Proposition Bank (Propbank); the other is FrameNet. The Propbank annotates the Penn Treebank with verb argument structure according as Levin class (Levin, 1993). It defines a general set of arguments for all types of predicates, and these arguments are divided into core and adjunct ones. FrameNet, as a linguistic ontology, describe the scenario related to each predicates. The scenario (i.e. frame) is filled with specific participants (i.e. role). In this paper, we use Chinese Propbank 1.0 provided by Linguistic Data Consortium (LDC), which is based on Chinese Treebank. It consists of 37,183 propositions indexed to the first 250k words in Chinese Treebank 5.1, including 4,865 verb types and 5,298 framesets.

[1] F1 measure computes the harmonic mean of precision and recall of SRL systems in CoNLL-2005

3 TSVM based SRL

3.1 TSVM

There are two kinds of learning modes that are applied in Artificial Intelligence, i.e. inductive inference and transductive inference. In classification problems, inductive inference trains a global model based on tagged instances from the whole problem space and classify new untagged instances by it. The classical statistical models such as SVM, ME have been developed in this way. Since large mount of tagged data are usually acquired difficultly in practice, and the global models are hard to get when tagged training data are not enough to find the target function in the hypothesis space. In addition, this global model may be unnecessary sometimes when we only care for specific data. Compared with inductive inference, transductive inference classifies untagged instances by a local model based on the clustering distribution of these untagged instances. The TSVM, a representative of transductive inference method, was introduced by Joachims (1999). TSVM is a good semi-supervised method special to some cases where the tagged data is difficult to acquire on a large scale while large untagged data is easily available. TSVM can be formulated as an optimization problem:

Minimize Over $(y_1^* .. y_n^*, w, b, \xi_1 .. \xi_n, \xi^* .. \xi_k^*)$ in

$$\frac{1}{2}\bar{w}^T \bar{w} + C \cdot \sum_{i=1}^{n} \xi_i + C^* \cdot \sum_{i=1}^{k} \xi_i^*, \text{ subject to:}$$

$$\forall_{i=1}^{n}\, y_i (\bar{w} \cdot \bar{x}_i + b) > 1 - \xi_i \quad \text{for } \forall_{i=1}^{n} \xi_i \geq 0$$

$$\forall_{i=1}^{k}\, y_i^* (\bar{w} \cdot \bar{x}_i^* + b) > 1 - \xi_i^* \quad \text{for } \forall_{i=1}^{k} \xi_i^* \geq 0$$

where $(x_1,y_1),\ldots,(x_n,y_n) \in S_{train}$, $y_1,\ldots,y_n \in \{-1,+1\}$, $x_1^*,\ldots,x_n^* \in S_{test}$, $y_1^*,\ldots,y_n^*$ is the labels of $x_1^*,\ldots, x_n^*$, C and C^*, specified by user, are the effect factor of the tagged and untagged examples respectively, $C^* \xi_i^*$ is the effect term of the *i*th untagged example in the above objective function. In addition, a cost-factor C_{temp}, which indicates the ratio of positive untagged examples, should be specified experientially by user before training.

Here we introduce the algorithm briefly, and the detail is referred to Joachims (1999). The algorithm starts with training regular SVM with the

tagged examples and then classifies the untagged examples by the trained model. Then several couples of examples (one is positive, the other is negative) are switched in class labels according to some rule, and the model is retrained to minimum the objective function. At the same time, C_{temp} will increase in consistent way. The iteration will end when C_{temp} goes beyond C^*. The algorithm is proved to converge in a finite number of steps.

3.2 Apply TSVM for SRL

The SRL using TSVM is related to following portions:

Dataset The principle of TSVM described in above section implicitly indicates the performance depends deeply on dataset (including tagged and untagged data). In particular, tagged data have an influence on original regular SVM in the first step of training, while the untagged data will affect the final performance through the iteration of training. It is obvious that the more even the data set distribution is, the better the learning classifier will perform. Similar to most practical classification task, a serious uneven problem (Li, 2003) exists in SRL. For instance, the number of constituents labeled to arguments (positive instances) is much less than the number of the rest (negative instances). To handle this problem, we design some heuristics for several kinds of arguments (that is, ARG0, ARGM-TMP, ARGM-LOC, ARGM-MNR, ARGM-DIR and ARGM-EXT) semantically. These heuristics filter out redundant constituents and raise the ratio of positive instances in the dataset. We will compare these argument-specific heuristics with Xue (2004), and some results are showed in Section 4.

Parameters The ratio of positive examples in dataset, *P*, is a key parameter in TSVM and should be assigned as one prior value in experiment. In this paper, *P* is dynamically assigned according to different argument since different heuristics could produce different proportion of positive and negative instances used to training data.

Features A wide range of features have been shown to be useful in previous work on SRL (Pradhan, 2005; Xue et al, 2004). This paper chooses 10 features in classification because of two reasons: at first, they are the core features considered to have significance on the performance of SRL (Carreras, 2005); secondly, these features provide a standard to evaluate different methods of Chinese SRL. These features are listed in Table 1, detail description referred in Xue (2005).

Feature	Description
Predicate	The predicate lemma
Subcat-Frame	The rule that expands the parent of verb
Path	The syntactic path through the parse tree from the parse constituent to the predicate being classified
Position	A binary feature identifying whether the phrase is before or after the predicate
Phrase Type	The syntactic category of the phrase corresponding to the argument
Phrase type of the sibling to the left	The syntactic category of the phrase is sibling to the argument in the left
Head Word and Part Of Speech	The syntactic head of the phrase
First and last word of the constituent in focus	First and last word of phrase corresponding to the argument
Syntactic Frame	The syntactic frame consists of the NPs that surround the predicate

Table 1. The features of Semantic Role Labeling

It should be mentioned that we have not considered the Combination features (Xue et al, 2005) because the above 10 features have already coded them. Verb class is also not be used here since we have no idea about the syntactic alternations used for verb classification in Xue (2005) and could not evaluate them equally. So, the experiment in this paper refers to the results without verb class in Xue (2005).

Classifiers Chinese Propbank has 22 argument types, in which 7 argument types appearing less than ten times or even having no appearance have not been considered, that is,ARGM-FRQ, ARGM-ASP, ARGM-PRD, ARGM-CRD, ARGM-T, and ARGM-DGR. So we have developed 15 binary classifiers for those 15 type of arguments and excluded the above 7 because they hardly provide useful information for classification, as well as have slightly influence on results (account for 0.02% in all arguments appeared in the corpus).

4 Heuristics

In this section, we discuss the principle of the designing of the argument-specific heuristics. To handle the uneven problem in SRL, six semantic heuristics have been designed for six types of arguments, such as ARG0, ARGM-TMP, ARGM-

LOC, ARGM-MNR, ARGM-DIR, and ARGM-EXT. The heuristic is actually some restrictive rules which can be viewed as pre-processing of identification. (Xue et al, 2004) introduced a primary algorithm for pruning argument non-candidates. The algorithm still remain large redundant unnecessary constituents yet (correct arguments account for 7.31% in all argument candidates extracted). (Park, 2005) used the clause boundary restriction and tree distance restriction for extracting candidates based on Government and Binding Theory. All of these restrictive rules, however, are on the syntax level. Here we consider several semantic features directly extracted by the head word of the argument in lexicon. This is based on facts that ARG0 contain mostly NPs whose head words are animate objects or entities. (Yi, 2007) shows *agent* and *experiencer* as ARG0 accounts for 93% in all ARG0s in Propbank. In addition, some head words of the constituents labeled by ARGM-TMP have temporal sense, which is the same as ARGM-LOC whose head words usually have spatial sense. The semantic information can be extracted from a Chinese-English bilingual semantic resource: HowNet (Dong, 2000). HowNet is an on-line common-sense knowledge base providing a universal lexical concept representation mechanism. Word sense representations are encoded by a set of approximately 2,000 primitive concepts, called sememes. A word sense is defined by its primary sememes. For example, *小孩*(*child*) is defined with sememes "human|人", "young|幼"; *目前 (at present)* has sememes "time|时间", "now|今"; *街* (*street*) contains sememes "location|位置", "route|路". We considered sememes as the basis of heuristics, and Table 2 shows these heuristics.

Table 2 shows the argument-specific heuristics on the semantics level, for example, only when the head word of a PP contains a sememe "time|时间", it could be a candidate of ARGM-TMP, such as *目前*, *当今*; only a sememe "location|位置" has a head word of one phrase, it may be labeled to ARGM-LOC. Furthermore, we make a comparison with Xue (2004) in whole argument types on Chinese Propbank (the extraction principle about argument_types which are not listed in Table 1 is the same as Xue (2004)). We find the argument-specific heuristics decrease in uneven problem more effectively than Xue (2004). The overall coverage [2] rises from 7.31% to 20.30%, that is, 65% constituents which have no possibility to labeling have been pruned based on six types of arguments. And the overall recall of arguments in corpus decline slightly from 99.36% to 97.28%.

Args	Def	Heuristic	Cover-age
ARG0	agent,ex-eriencer	the NP whose head word has sememe that is hyponymy with animate|生物 or whose head word is place or organization	38.90
ARGM-TMP	temporal	The NP and LCP whose head word has sememe time|时间 or the PP whose prep is from|从, from|自, to|到, in|於, or at|在	58.7
ARGM-LOC	location	The NP and LCP whose head word has sememe location|位置 or the PP whose prep is in|在 ,at|在 or from|于	44.4
ARGM-MNR	manner	The PP whose prep is "according to|根据, 按,据, 按照" or by|通过, as|随着	30.98
ARGM-DIR	directional	The PP whose prep is to|对 or from|从, to|向	20.56
ARGM-EXT	extent	The NP and QP whose head word is number	70.27

Table 2. The arguments-specific heuristics.

5 Experiment and discussion

This section will describe the experiment on the SRL in Chinese Treebank, compare TSVM with regular SVM, and evaluate the effect of the proposed argument-specific heuristics.

5.1 Experiment Setting

SVM-light[3] is used as a SVM classifier toolkit in the experiment, which includes some sub-tools for optimizing performance and reducing training time. It also provides an approximate implementation of transductive SVM. At first, about 80% propositions (1711891) has been extracted randomly from the corpus as the dataset, which had been divided into tagged set and untagged set according to 4:1. Then, for each type of arguments,

[2]The coverage means the ratio of arguments in all role candidates extracted from Chinese Propbank by given heuristic.

[3] http://svmlight.joachims.org/

numeric vectors are extracted from these two sets (one proposition could produce many instances) as the dataset for the following learning models through the heuristics in Table 2. When training the classifier, linear kernel function had used, setting the C to 2 experientially.

5.2 Results and Discussion

A baseline was developed with 10 features and 15 SVM classifiers (tagged set for training, untagged set for testing) as described in Section 3. We made a comparison between the baseline and the work in Xue (2005), and then used the argument-specific heuristics for baseline. Table 3 shows the performance of these methods. Baseline matches Xue approximately despite of the absence of combination features. We also find that the argument-specific heuristics improve the performance of baseline from 89.97% to 90.86% for F1 and beyond the Xue. It can be explained that when using heuristics, the proportion of positive and negative instances in dataset are adjusted reasonably to improve the model. About 1 percent improvement attributes to the effectivity of these six argument-specific heuristics.

Systems	Precision	Recall	F1
Baseline	89.70	90.24	89.97
Xue	90.40	90.30	90.30
Heuristics	91.45	90.28	90.86

Table 3. A comparison among baseline, Xue and heuristics through regular SVM

In order to investigating the learning performance of SVM, TSVM and TSVM using argument-specific heuristics in small tagged data, we extracted randomly different number of propositions in Propbank as tagged data and another 5000 propositions held out as untagged data. Both of them are used for training TSVM model. Table 4 shows the overall performance and the performances of two arguments--ARG0 and ARGM-TMP--along with the different training data size. As we can see in (a) of Table 4, the TSVM leads to an improved performance on overall argument types when tagged data less than 100 propositions (raising F1 about 10%). It indicates that transductive inference performs much better than inductive inference because it makes use of the additional information about the distribution of 5000 untagged propositions. More important, we find that TSVM using argument-specific heuristics, comparing to TSVM, has a distinctive improvement (raising about 3%). It confirmed that our heuristics have positive influences on transductive inference.

Number of tagged propositions	SVM	TSVM	TSVM + Heuristics
10	36.51	50.51	50.82
20	41.65	50.52	53.66
40	41.64	55.42	60.63
160	76.40	80.84	82.32
1000	82.00	83.87	84.00
5000	84.41	85.61	86.45

(a). The overall results on all argument types.

Number of tagged propositions	SVM	TSVM	TSVM + Heuristics
10	20.51	29.51	30.21
20	22.34	32.45	38.54
40	35.00	45.42	50.63
160	45.45	50.45	55.74
1000	52.43	55.43	57.40
5000	58.00	60.34	61.45

(b) The detail results on ARG0

Number of tagged propositions	SVM	TSVM	TSVM + Heuristics
10	15.98	20.45	19.98
20	25.34	29.45	35.43
40	30.32	32.80	39.43
160	38.31	40.00	45.09
1000	48.43	50.43	55.45
5000	60.34	62.34	63.90

(c) The detail results on ARGM-TMP

Table 4. A comparison with Regular SVM, TSVM and TSVM using argument-specific heuristics holding 5000 untagged propositions

Number of untagged propositions	SVM	TSVM	TSVM + Heuristics
500	69.03	68.50	69.44
1000	70.12	70.22	70.82
2000	68.64	71.30	73.01
4000	69.53	72.01	76.50
5000	68.95	72.54	77.21
10000	70.28	74.78	79.74

Table 5. A comparison with Regular SVM, TSVM and TSVM using argument-specific heuristics holding 100 tagged propositions

We then evaluate the six argument-specific heuristics introduced in Section 4 with the same 5000 untagged propositions. It is noticeable that the training time of TSVM doubles that of SVM approximately. The (b) and (c) of Table 4 give the detail results on ARG0 and ARGM-TMP. Com-

pared with (a), it is obvious that the improvement between TSVM using heuristics with TSVM for ARG0 and ARGM-TMP is larger than the overall improvement. That is to say, the more distinctive knowledge is embedded in heuristics, the better performance can be achieved for the corresponding argument. This observation encourages us to investigate more heuristics for more arguments.

Finally, the influence of untagged data on performance of TSVM has been investigated. We extract different size of untagged propositions and hold 100 tagged propositions for training TSVM. Table 5 shows the results. It should be mention that the result of SVM fluctuates slightly, which is due to different number of testing examples. On the other hand, TSVM and TSVM using argument-specific heuristics improve highly as the increase in untagged data size. The bigger the untagged data, the larger the performance gap between SVM and TSVM and the gap between TSVM and TSVM using argument-specific heuristics. It indicates that the argument-specific heuristics, optimizing the dataset, have substantial effectivity in the performance of TSVM when untagged data is large.

6 Conclusions

Most machine learning methods such as SVM, ME have a strong dependence on tagged data, which lead to a poor generalization when large tagged data are not available. This paper introduces a novel semi-supervised method--TSVM for this problem. TSVM can effectively use clustering information from untagged data for training the model. The experiment demonstrated the TSVM achieve better performance than regular SVM when only very few tagged examples are available. Aiming at serious uneven problem in SRL, argument-specific heuristics are proposed correspond to six kinds of arguments. These heuristics are developed by extracting semantic features of arguments from HowNet. The experiment proves that these heuristics have much effect not only in the inductive inference (regular SVM) but also in transductive inference (TSVM), especially when the untagged data is large. The high performance of six heuristics demonstrated that semantic characteristics are significant on SRL, which encourages us to develop more semantic characteristics of more arguments in the future.

Acknowledgement This research is supported by the National Natural Science Foundation of China (60403050), Program for New Century Excellent Talents in University (NCET-06-0926) and the National Grand Fundamental Research Program of China under Grant (2005CB321802).

References

Levin Beth. 1993. *English Verb Class and Alternations: A Preliminary Investigation.* Chicago: University of Chicago Press.

Xavier Carreras and Llu´ıs M`arquez, 2005. *Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling*. CoNLL-2005.

Trevor Cohn and Philip Blunsom. 2005. *Semantic role labeling with tree conditional random fields*. CoNLL-2005.

Zhendong Dong. 2000. http://www.keenage.com/.

Thorsten Joachims. 1999. *Transductive inference for text classification using support vector machines.* ICML-99, pages 200–209, Bled, Slovenia, Jun.

Yaoyong Li and John Shawe-Taylor. 2003. *The SVM with Uneven Margins and Chinese Document Categorization*. PACLIC-2003, Singapore.

Kyung-Mi Park and Hae-Chang Rim. 2005. *Maximum entropy based semantic role labeling.* CoNLL-2005.

Pradhan S, Hacioglu K, Krugler V, et al. 2005. *Support Vector Learning for Semantic Argument Classification.* Machine Learning journal. 60(1-3): 11-39.

Nianwen Xue and Martha Palmer. 2004. *Calibrating Features for Semantic Role Labeling.* EMNLP.

Nianwen Xue and Martha Palmer. 2005, *Automatic Semantic Role Labeling for Chinese Verbs.* The IJCAI-2005, Edinburgh, Scotland.

Szuting Yi, Edward Loper and Martha Palmer. 2007. *Can Semantic Roles Generalize Across Genres?* NAANL-HLT 07, Rochester, N Y.