

A Knowledge-Based Semantic Role Labeling System for Chinese

First Author

Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author

Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Third Author

Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

We present a knowledge-based semantic role labeling system for Chinese parsed sentences. Our proposed system outperformed an existing one from two aspects: knowledge utilization and model design. As to the former, the semantic knowledge obtained from E-HowNet were utilized to solve the data sparseness issue; as to the latter, a combination of back-off models was proposed for semantic role classification. In the classification part, the proposed strategy defeated a number of well-established classification approaches (i.e. Naive-Bayes, Decision Trees, Maximum Entropy, Linear Interpolation) resulting in an overall accuracy improvement from 92.71% to 94.73% of the existing system.

1 Introduction

Over the last few years, syntactic parsing paradigms have enjoyed an admirable success, and this has paved the way for semantic parsing. Semantic role labeling (SRL), also known as shallow semantic parsing, is the task of semantically annotating natural language text. Conventionally, a syntactically parsed sentence is taken as input, and semantic arguments associated with predicate of the sentence are identified and classified to a particular semantic class.

The first automatic semantic role labeling systems was reported by (Gildea and Jurafsky, 2002), and since then, their ideas have been dominating the field. In their approach, they emphasized on selection of appropriate lexical and syntactical features for SRL, use of statistical classifiers and their combinations, and ways to handle the data sparseness issue. Researchers have tried to build on that by augmenting and/or altering the

feature set (Chen and Rambow, 2003; Xue, 2004), by experimenting with various classification approaches (Park and Rim, 2005; Tan et al., 2009), and by attempting different ways to handle data sparseness (Zapirain et al., 2007; Lin et al., 2010). In recent years, many other similar attempts were made, and loosely speaking, it can be concluded that enhancing the performance of a SRL system by calibrating an effective feature-set has reached to a saturation level. One needs to find other ways to take the field to the next level of excellence.

In this study, we propose the use of semantic knowledge, and a novel classification approach to enhance the performance of an already reported SRL system. A semantic knowledge-base is exploited to address the data sparseness concern, and the proposed classification method, which is based on a combination of weighted probabilistic models, is used for better semantic role classification. To show the effectiveness of our approach, we build a number of systems that are based on other well established classification approaches (e.g. Naive-Bayes, Decision Trees, Maximum Entropy, Linear Interpolation), and compare the outcomes. The experimental results show that our system outperformed all others, and this lead to a considerable improvement in the accuracy of the previously reported system.

2 Experimental Material

The following two subsections briefly describe the training and testing data we used, and our semantic knowledge source respectively.

2.1 Sinica Treebank

Sinica Treebank (Chen et al., 2003) is a semantically annotated Chinese tree-bank. Its version 3.0 contains 61,087 syntactic tree structures, and 361,834 words. In addition to conventional lexical and syntactic annotations, each tree is marked for semantic relations of a verbal predicate. It

used 74 abstract semantic roles including primary thematic roles (e.g. 'agent', 'theme', etc.), secondary roles (e.g. 'location', 'time', etc.), and noun-modification roles (e.g. 'quantifier', 'possessor' etc.). Fig 1 shows an example parse tree from Sinica Treebank.

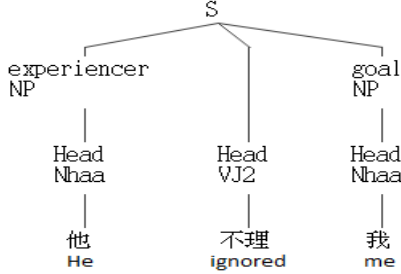


Figure 1: An example tree from Sinica Treebank

2.2 E-HowNet

E-HowNet (Chen et al., 2010) is a semantic lexical knowledge-base for Chinese. It is a frame based entity-relation model that defines relationships between 2600 well-defined primitive concepts ("entities"). Around 90,000 Chinese words are categorized into different groups based on their sense similarity, and then those groups are tagged with the primitive concept they represent. For example, the entry `speaker|說話者` [人家, 余, 吾, 我, 咱, 咱家, 俺, 俺, 咱] from E-HowNet shows that the nine Chinese words (enclosed in square brackets) share the same sense, and they belong to the primitive concept 'speaker'.

In this study, we use this grouping to generalize lexical statistics, and hence, to abate the effects of data sparseness on SRL. Throughout this paper, we have used the term "semantic type (`semType`)" to refer to the primitive concept a given Chinese word represent.

3 Previous SRL System and Issues

A semantic role labeling (SRL) system for Chinese was reported by (You and Chen, 2004). In that system, Sinica Treebank was used to train probabilistic models, which, together with a back-off strategy, were used for semantic role classification. As far feature set, the system relied on a number of conventional lexical features: target word (`t`), target word pos (`t_pos`), head word (`h`), head word pos (`h_pos`), and syntactic features: phrase type (`pt`) and `position` (i.e. position of the target word with respect to the head

word). The overall accuracy of the system was reported to be 92.71%. Even though this accuracy can be considered on the higher side, there was room for improvement, especially in two sub-tasks: data sparseness handling and classification approach.

Data sparseness handling first, the previous system used a back-off strategy to tackle this problem – see (You and Chen, 2004) for more details. This considerably improved the baseline performance, but data sparseness still remained an issue. A major reason for this was dependency of the system on non-generalized lexical features (e.g. `h_word`, `t_word`). As mentioned previously, in the proposed system, we use E-HowNet to address this problem (Section 5).

Classification methodology next. In the existing system, six probabilistic models, together with a back-off strategy, were used for semantic role assignment. As shown below, a particular feature combination was used for each model, and the backing-off was based on number of examples seen in the training data.

```

if#of(h,h_pos,t,t_pos,pt,position)
>threshold
P(r|constituent)=
P(r|h,h_pos,t,t_pos,pt,position)
else
if#of(h_pos,t,t_pos,pt,position)
>threshold
P(r|constituent)=
P(r|h_pos,t,t_pos,pt,position)
else
.....

```

A critical observation is that backing-off based on a constant threshold value does not allow to exploit the statistical knowledge in the best possible way. The reason is that if enough examples, with a more constrained feature combination, have been seen in the training data, the less constrained statistical information becomes irrelevant, which actually may suggest a more probable role. We propose a different strategy that is based on a combination of weighted probabilistic models. It utilizes the statistical knowledge in a better way, and hence, provide improvement in performance (see Section 4.3 for more details).

4 Proposed SRL System

The proposed SRL system is statistical. Its feature set, probabilistic models, and classification

approach are given in the following subsections.

4.1 Feature Set

Lexical Features: Part of speech tag of the target word (t_pos), part of speech tag of the head word (h_pos), part of speech tags of the immediate left and right siblings of a test node ($l_r_ch_pos$), a set of part of speech tags of all nodes under a test node including the test node itself (all_pos).

Syntactic Features: phrase type (pt), a sentence-level boolean feature indicating whether the sentence containing the test node is passive or not ($passive$), position of the target node w.r.t phrasal head ($position$).

Semantic Features: Semantic type of the head word extracted from E-HowNet ($h_w_semType$), semantic type of the target word ($t_w_semType$), a set of semantic types of all nodes of the tree the test node is a child of ($all_semType$).

Table 1 shows the values of these features for each constituent of the parse tree shown in Figure 1.

Semantic Role	experiencer	goal
$t_w_semType$	3rdPerson	speaker
t_pos	NP	NP
$h_w_semType$	ShowInterest	ShowInterest
h_pos	VJ2	VJ2
pt	S	S
$position$	before	after
$passive$	false	false
$l_r_ch_pos$	empty-VJ2	VJ2-empty
all_pos	NP-Nhaa	NP-Nhaa
$all_semType$	3rdPerson-ShowInterest-speaker	3rdPerson-ShowInterest-speaker

Table 1: Features-values of each constituent of Figure 1 parse tree.

4.2 Probabilistic Models

Ten probabilistic models were built using the Sinica Treebank labeled data. The probabilities were estimated using the following formula:

$$\begin{aligned}
P(r|constituent) \\
&= P(r|f_c) \\
&= \#(r, f_c) / \#f_c
\end{aligned}$$

Where f_c represents a particular feature combination. A number of combinations were tested, and for the final system we used the set of ten combinations given in Table 2.

#	Feature Combination
1	$semType_h_w, h_pos, semType_t_w, t_pos, pt, position, all_pos, passive, all_semType, l_r_ch_pos$
2	$h_pos, t_w_semType, t_pos, pt, position, passive$
3	$semType_h_w, h_pos, t_pos, pt, position, passive$
4	$semType_t_w, t_pos, pt, position, passive$
5	$semType_h_w, semType_t_w, pt, position, passive$
6	$semType_t_word, t_pos, pt, passive$
7	$semType_t_word, t_pos, pt, passive$
8	$semType_t_word, t_pos, passive$
9	$t_pos, pt, position, passive$
10	$semType_t_w$

Table 2: The set of feature combinations

4.3 Classification Method

Notations

- Let F be the set of ten feature combinations (given in Table 2), and f_c be an element of F
- P be the set of ten corresponding probabilistic models, and P_{f_c} be an element of P that is based on feature combination f_c
- D_{f_c} be the probability distribution computed by P_{f_c}
- $M_{((p,r)|f_c)}$ be the (*probability, role*) pair in D_{f_c} with highest probability
- W be a set of optimal weights, w be an element of W , and wp be a weighted probability (i.e. rank)

Algorithm

The proposed classification algorithm is:

1. For a test candidate, extract values of all features (mentioned in section 4.1) using the parse tree and E-HowNet.
2. initialize $potential_roles \leftarrow empty$
3. For each $f_c \in F$ do:
 - Find probability distribution D_{f_c} using the corresponding P_{f_c} model
 - Select $M_{((p,r)|f_c)}$ from D_{f_c}
 - Rank $M_{((p,r)|f_c)}$ by multiplying p with the corresponding w from W
 - Append (wp, r) to $potential_roles$
4. Return the top ranked r from $potential_roles$

The major idea is to utilize both more and less constrained statistical knowledge in determining the semantic role. For this purpose ten probabilistic models are built. Each model is based on particular feature constraints, which are relaxed as we move from feature combination 1 to 10 in Table 2. The output of each model is ranked by learned weights, and the top ranked role is finally assigned to each test constituent. The weights, which encode the worth of a particular feature combination in determining the role, are learned using genetic algorithms and a held out development data set. The learned set of optimal weights is:

{ (w1:0.9), (w2:1.0), (w3:0.9),
(w4:0.7), (w5:0.8), (w6:0.4),
(w7:0.5), (w8:0.4), (w9:0.5),
(w10:0.4) }

5 Data Sparseness Handling

As pointed by (Gildea and Jurafsky, 2002), lexical statistics, though very useful for semantic role labeling, often becomes a source of data sparseness. For the same reason, we replaced two lexical features (i.e. `h_word`, and `t_word`) used by the previous system with more general features (i.e. `semType_h_word` and `semType_t_word`) respectively. Table 3 gives coverage and the corresponding accuracy statistics before and after adding these generalizations for each of the ten probabilistic models (given in Section 4.2). As can be seen, after generalization, the coverage improved considerably for each feature combination, which ultimately resulted in better performance.

P_{fc}	With Words		With Semantic Types	
	Coverage	Accuracy	Coverage	Accuracy
1	7.10%	23.61%	12.68%	28.00%
2	52.06%	59.14%	73.05%	74.98%
3	67.43%	69.87%	82.68%	81.15%
4	74.56%	68.92%	93.35%	79.16%
5	97.97%	92.25%	97.97%	92.27%
6	28.60%	40.13%	64.20%	62.95%
7	76.30%	68.07%	94.23%	76.87%
8	82.97%	63.50%	96.85%	67.40%
9	99.78%	79.17%	99.78%	79.41%
10	90.09%	58.18%	99.00%	53.31%

Table 3: Coverage & Accuracy Statistics

6 Experiments and Evaluation

To evaluate the performance and to show the usefulness of our approach, we built five semantic role

labeling systems: (1) **DT**: Based on Decision Tree classifier, (2) **NB**: Based on Naive Bays classifier, (3) **ME**: Based on Maximum Entropy classifier, (4) **LI**: Based on probabilistic models together with linear interpolation, and (5) **WPM**: Based on the proposed weighted probabilistic models (WPM) classification approach. All of these systems were tested using the same feature set (given in Section 4.1), and the same training and testing data (i.e. Sinica Treebank). Results of a 10-fold cross validation scheme are given in Table 4. From

System	Accuracy	Precision
DT	91.03%	91.04%
NB	92.58%	92.57%
ME	92.67%	92.68%
LI	94.26%	94.27%
WPM	94.49%	94.49%

Table 4: Evaluation results

the resulting numbers, we can see that our system outperformed all other systems.

To further enhance the performance of our system, we added a post-processing component to fix some of the obvious mistakes made by the probabilistic models. One such case is disambiguation between possessor/property roles. We used a heuristic rule that if the semantic type of a target word is human, it is more likely to be a possessor than a property. With few other such rules, the scores were improved from 94.49% to 94.73%.

7 Conclusion and Future Work

In the scenario, where improving the performance of a SRL system by better engineering the feature-set looks more challenging, we have proposed an alternative way to achieve the same goal. We have presented a distinct classification approach, and the use of semantic knowledge to enhance the performance of an already reported SRL system. In conclusion, the proposed system outperformed a number of other systems, which were based on well-known classification approaches. With its better performance, the overall accuracy of the existing SRL system was improved from 92.71% to 94.73%.

Since the proposed approach is language independent, in future, we have plans to experiment using the English and Chinese prop-bank data. Our hope is that we can improve state-of-the-art SRL scores.

References

- John Chen and Owen Rambow. 2003. Use of deep linguistic features for the recognition and labeling of semantic arguments. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 41–48, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Keh-Jiann Chen, Chi-Ching Luo, Ming-Chung Chang, Feng-Yi Chen, ChaoJan Chen, and Chu-Ren Huang. 2003. Sinica treebank: Design criteria, representational issues and implementation. *Anne Abeille (Ed.) Treebanks Building and Using Parsed Corpora. Language and Speech series. Dordrecht:Kluwer*, pages pp231–248.
- Wei-Te Chen, Su-Chu Lin, Shu-Ling Huang, You-Shan Chung, and Keh-Jiann Chen. 2010. E-hownet and automatic construction of a lexical ontology. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING '10, pages 45–48, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Comput. Linguist.*, 28(3):245–288, September.
- Xiaojun Lin, Meng Zhang, and Xihong Wu. 2010. Chinese semantic role labeling with hierarchical semantic knowledge. In *Proceedings of the 2010 International Conference on Electrical and Control Engineering*, ICECE '10, pages 583–586, Washington, DC, USA. IEEE Computer Society.
- Kyung-Mi Park and Hae-Chang Rim. 2005. Maximum entropy based semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 209–212, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Yongmei Tan, Xu Wang, and Yong Chen. 2009. Chinese semantic role labeling using crfs and svms. In *Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on*, pages 1–5.
- Nianwen Xue. 2004. Calibrating features for semantic role labeling. In *In Proceedings of EMNLP 2004*, pages 88–94.
- Jia-Ming You and Keh-Jiann Chen. 2004. Automatic semantic role assignment for a tree structure. In Oliver Streiter and Qin Lu, editors, *ACL SIGHAN Workshop 2004*, pages 109–115, Barcelona, Spain, July. Association for Computational Linguistics.
- Beñat Zepirain, Eneko Agirre, and Lluís Màrquez. 2007. Ubc-upc: Sequential srl using selectional preferences: An aproach with maximum entropy markov models. In *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval '07*, pages 354–357, Stroudsburg, PA, USA. Association for Computational Linguistics.