INTERSPEECH
2009 BRIGHTON

Semantic Role Labeling with Discriminative Feature Selection for Spoken Language Understanding

Chao-Hong Liu and Chung-Hsien Wu

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, TAIWAN
chl@csie.ncku.edu.tw, chwu@csie.ncku.edu.tw

10.21437/Interspeech.2009-321

Abstract

In the task of Spoken Language Understanding (SLU), Intent Classification techniques have been applied to different domains of Spoken Dialog Systems (SDS). Recently it was shown that intent classification performance can be improved with Semantic Role (SR) information. However, using SR information for SDS encounters two difficulties: 1) the state-of-the-art Automatic Speech Recognition (ASR) systems provide less than 80% recognition rate, 2) speech always exhibits ungrammatical expressions. This study presents an approach to Semantic Role Labeling (SRL) with discriminative feature selection to improve the performance of SDS. Bernoulli event features on word and part-of-speech sequences are introduced for better representation of the ASR recognized text. SRL and SLU experiments conducted using CoNLL-2005 SRL corpus and ATIS spoken corpus show that the proposed feature selection method with Bernoulli event features can improve intent classification by 3.4% and the performance of SRL.

Index Terms: discriminative feature selection, semantic role labeling, intent classification, spoken language understanding (SLU)

1. Introduction

Spoken Language Understanding is a crucial task in Natural Language Processing (NLP) dedicated to the understanding of semantic meaning as exhibited in speech. To fulfill the goal of SLU, techniques for intent classification (or Speech-Act/Call-Type Classification for different domains) are developed for spoken dialog systems. Recently as the development of semantic role labeling has progressed, the Semantic Role information was shown to be an important information to facilitate spoken language understanding [1, 2]. However, to derive SR information from speech is a very difficult task. The state-of-the-art ASR systems only provide less than 80% recognition rate for most spoken dialog environments. Furthermore, even if the recognition rate is 100%, the day-to-day speech utterances tend to include substantial ungrammatical fragments from which no parser and Semantic Role Labeler can provide useful results. These difficulties have made the task of SRL tasks using speech input a very challenge problem for SLU.

The task of Semantic Role Labeling is to identify the predefined semantic roles such as subject and object of the predicate structure in a sentence. The labeling results can help automatic understanding of the information about “who” does “what” to “whom,” “when” and “where” in sentences. With respect to a predicate there can be several constituents with different semantic roles across one sentence. Furthermore, in a single sentence, more than one predicate structure can co-exist with boundaries intersected.

An interesting characteristic of the task of Semantic Role Labeling is that it usually involves millions of samples and features for classifier training. This has made Semantic Role Labeling a challenging machine learning task since it has been shown that the performance of many machine learning methods will deteriorate when facing a huge number of features. In order to reduce the number of features and to possibly use only the discriminative subset of features, feature selection methods can be applied to deal with the huge data problem. Table 1, as an example, shows the effect of using different numbers of features to train the semantic role A0 (refers to the subject of the sentence) using CoNLL-2005 shared task corpus [3]. It can be seen that increasing the number of features does not necessarily increase the classification performance, whereas the performance reaches its apex when an appropriate number of features are selected. Furthermore, considering the huge number of features encountered in the task of Semantic Role Labeling, feature selection becomes a necessary step to be used to improve the efficiency and performance of this task.

There are two broad categories of stand-alone feature selection techniques: filters and wrappers [4]. Filters define scores for each feature using several criteria such as correlation coefficient, mutual information and entropy. Then the scores are used to select the features using an empirically determined threshold. Different from the filters method, wrappers depend on the results of classifiers trained on the subsets to determine which subset of feature space is to be selected. The difference between these two techniques is that filters do not use the performance of the classifiers trained on the subsets of features for consideration [5].

While filters are usually more computationally efficient than wrappers, both techniques face the same computational complexity problem because the space of all possible combinations of subsets of a feature space is usually too large to be exhaustively considered [5]. Therefore search strategies are often used in feature selection. On the other hand, clustering techniques, already employed as a means for fast feature selection itself [6], can also help other feature selection methods to reduce the space of feature subsets. In the field of Natural Language Processing, Incremental Feature Selection (IFS) was shown to be a useful feature selection method [7], while the RELIEF algorithm is extensively used for general purpose feature selection [8].

In this study, an approach to Semantic Role Labeling using discriminative feature selection is proposed. Based on the criterion of minimum misclassification error, the discriminative feature selection approach can improve the classification performance by selecting the distinguishing features from a huge number features. To fulfill the requirement of SLU, Bernoulli event features is introduced to help address the problem of ASR errors and ungrammaticality commonly occurred in spoken language. The experiments of

Table 1. *The Effect of Number of Features Selected on the Performance of Semantic Role Labeling on A0.*

Testset	# features	Precision	Recall	F_1
Brown	2,000	51.51%	86.69%	64.62
	10,000	55.91%	83.76%	67.06
	40,000	57.27%	83.27%	67.86*
	200,000	56.18%	82.68%	66.90
	ALL	54.13%	82.09%	65.24
WSJ	2,000	44.54%	88.13%	59.18
	10,000	53.64%	83.62%	65.36
	40,000	56.87%	81.23%	66.90*
	200,000	57.01%	80.83%	66.86
	ALL	56.80%	80.29%	66.53

The values with an asterisk indicate the highest F-score.

evaluation on the performance SRL and SLU using CoNLL-2005 SRL corpus and ATIS corpus [9, 10] were conducted and confirmed the effectiveness of the proposed approach.

The organization of this paper is as follows. Section 2 gives an overview of the system architecture and presents the features used in our system. Section 3 describes the proposed discriminative feature selection method. Section 4 provides the experiments on performance evaluation of the proposed methods. The last section presents the conclusions.

2. System Architecture

The Intent Classification system employed in this study follows the iterative approach as described in [1]. The training steps of the intent classification system are as follows:

1. Automatically deriving the intent-related SRL results from the general domain corpus and the specific domain corpus using predicate/argument pairs.
2. Grouping these primitive predicate/argument pairs into semantically equivalent groups with "mapping rules" written by experts to form the "Intents" for the domain in question.
3. Automatically tagging the corpus with these mapping rules for initialization.
4. Training the classifiers using the tagged portion of the corpus.
5. Classifying the corpus using the trained classifiers.
6. Evaluating if the performance of the classifiers is better than the previous results. If yes, then go to Step 4; else stop.

The objective of involving experts in grouping semantically equivalent pairs is to ease the job of the design of spoken dialog systems. It was shown that this iterative approach can greatly eliminate the under-performance of SRL resulting from ASR errors, with which experiments on the transcribed speech presents only about 4% edge over experiments on the text obtained from ASR [1].

For SRL performance evaluation on the ATIS corpus the intentions are grouped manually into five semantic classes as listed below. These intents are not necessarily of strict predicate/argument format because in the ATIS corpus many utterances do not even contain a predicate.

- Class 1: list/flight; need/flight; find/flight; show/flight; know/flight
- Class 2: book/flight; cancel/
- Class 3: cost/fare; show/fare; cheapest/; better/; lowest/

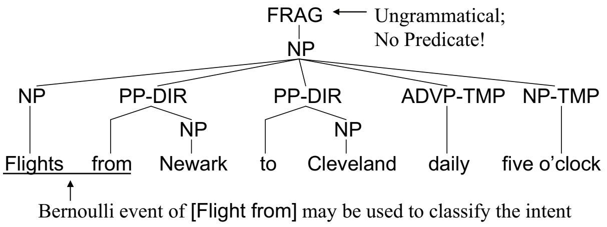
FRAG
Ungrammatical; No Predicate!
NP
NP
PP-DIR
NP
PP-DIR
NP
ADVP-TMP
NP-TMP
Flights
from
Newark
to
Cleveland
daily
five o'clock
Bernoulli event of [Flight from] may be used to classify the intent

Figure 1: *A Common Ungrammatical Utterance from ATIS Corpus Exhibits the Difficulty for SLU.*

- Class 4: arrive/; depart/; stop/; leave/; return/; visit/; travel/; /between; /from; /to
- Class 5: serve/dinner; serve/breakfast; /transportation; /airline; /airport

2.1. Semantic Role Labeling

In a common architecture an SRL system is divided into three consecutive stages: Sample Pruning, Argument Identification, and Argument Classification.

In the Sample Pruning stage, each constituent is usually grouped into a possible argument (Non-NULL) or a null argument (NULL) using heuristic rules or classifiers. The constituents determined as unlikely to be an argument are then pruned out in both training and test procedures.

In the Argument Identification stage, a classifier is trained using the remaining samples to actually detect if a constituent is an argument of a semantic role or not, pretending that no constituents are pruned out. It was shown that a classifier trained on the pruned sample subset derived from the Sample Pruning stage has a comparable performance compared to the one trained on the whole sample space.

In the Argument Identification stage, using the same pruned sample space of constituents, a classifier (or classifiers) is trained to determine the semantic role of a constituent. In this stage, the results of the semantic role classifier can be inconsistent with each other across a sentence with respect to a predicate. Therefore in most systems a simple conflict resolution or a sophisticated global inference method is always applied to yield the final labeled results [11]. This conflict resolution or inference method can be regarded as a part of the Argument Identification stage, and is usually used as a means to introduce the linguistic knowledge into to the Semantic Role Labeling task.

2.2. Features for SRL and Intent Classification

The baseline features employed in this research consist of the commonly used features for Semantic Role Labeling and some additional features used in highly ranked systems reported in CoNLL-2005 SRL shared task. We will also introduce new type of features, which incurs tens of millions of features for the problem, to show how the new features and feature selection can be applied to improve the labeling performance. Following most systems reported in CoNLL-2005, the entities considered in this paper are the constituents of a parse tree of the sentence to be labeled, by which the structural features can be extracted accordingly. These commonly used feature types are listed as follows.

- Predicate and its POS; Word/Stem/POS of Head
- Sub-categorization; Governing Category; Phrase Type

- Voice; Named Entities
- First Word/Stem/POS; Last Word/Stem/POS
- Chunk; Chunk Pattern Length
- Path; Path to Apex Upward and Downward (ApexPath)
- Position; Clause Relative Position (CRPosition)

2.2.1. *Bernoulli Event Features*

To exploit the benefit provided by feature selection, new types of features in addition to the baseline features can be introduced without prudent consideration of whether they are linguistically sound to the task of Semantic Role Labeling. Inspired by the *N*-gram based metrics such as BLEU [12] and NIST [13] used in machine translation evaluation, a reasonable representation of an argument or a chunk is defined to comprise the Bernoulli events of continuous word sequences of lengths one to three and their corresponding POS sequences. Since the number of all POS tags (including punctuation marks) used in Penn Treebank [10] is 53, the total number of features for Bernoulli events of POS sequences is therefore $\sum_{i=1}^{3} 53^i = 151{,}739$, and the number of word sequences alone easily exceeds millions.

Utterances of spoken language are usually ungrammatical and impose great challenges on parsing and other NLP tasks such as Semantic Role Labeling. Since the results of SRL are crucial for spoken language understanding, it is especially beneficial if the performance of SRL can be improved even if the input utterance is ungrammatical. Fig. 1 shows an ungrammatical utterance that contains no predicate, which makes the task of SRL unable to proceed. However, with the introduction of the “Flight from” Bernoulli event feature, it is possible for the task of intent classification to correctly identify the intent, thus improves the performance of spoken language understanding.

3. Discriminative Feature Selection

In this paper the Discriminative Feature Selection method is adopted by considering the likelihood [7] and loss function in data classification to provide discriminative property such that the yielded feature space is not only very highly representative to the original feature space but also minimizing the misclassification rate. The proposed feature selection method is described as follows.

With an initial feature space $S = \varnothing$ (an empty set), the optimal classification model, implemented using SVMs in this research, can be obtained as

$$\Lambda^*_{S\cup f} = \arg\max_{\Lambda\in\Gamma_{S\cup f}} \left(DL(\Lambda_{S\cup f}) - DL(\Lambda_S) \right) \qquad (1)$$

where $\Gamma_{S\cup f}$ is the set of all possible models with feature space $S\cup f$ which is the union of the original feature set *S* and the newly included feature *f*, and Λ_S is a classification model trained using *S*. $\hat{f}$ is then iteratively adjoined into *S* satisfying the following condition.

$$\hat{f} = \arg\max_{f} \left(DL(\Lambda_{S\cup f}) - DL(\Lambda_S) \right) \quad (2)$$

The discriminative likelihood for model Λ_S is defined as

$$DL(\Lambda_S) = (1-\kappa)\cdot L_{likeli}(\Lambda_S) - \kappa\cdot L_{loss}(\Lambda_S) \qquad (3)$$

where $L_{likeli}(\Lambda_S)$ is the likelihood of model Λ_S and $L_{loss}(\Lambda_S)$ is the loss of model Λ_S. The likelihood function is defined as

Table 2. *Contributions of Structural Features for SRL.*

Increase in F_1	Path	ApexPath	Position	Bernoulli
A0	05.07	03.24	05.10	05.87
A1	02.51	04.93	06.37	05.57
A2	20.27	19.37	18.95	11.20
A3*	-19.19	-20.83	-25.75	-22.58
AM-ADV	12.99	08.44	09.81	09.24
AM-DIR*	-05.64	-06.89	-10.66	-07.82
AM-DIS	37.86	36.80	37.25	17.79
AM-LOC*	-10.91	-14.29	-13.33	-10.81
AM-MNR	11.25	08.08	10.12	05.73
AM-MOD	49.36	28.16	37.90	36.74
AM-NEG*	-30.00	-23.08	-09.52	-08.70
AM-TMP	42.81	39.83	42.79	37.91
R-A0	30.15	43.35	04.50	03.63
R-A1	30.48	14.92	29.08	11.63
R-A2*	-06.65	-13.92	-04.52	-08.78
R-AM-LOC*	-26.97	-30.14	-19.89	-20.52
R-AM-MNR	22.27	29.40	25.58	24.44
R-AM-TMP	24.62	06.80	10.18	05.82
V	11.55	04.66	07.49	08.42

Roles with an asterisk are those shown to have negative impact.

$$L_{likeli}(\Lambda_S) = \sum_{i=1}^{N}\sum_{k=1}^{K} \left(p(x_i \mid C_k)\log p_{svm}(C_k \mid x_i)\cdot I(x_i \in C_k) \right) \qquad (4)$$

where *I* is an indicator function with the value being 1 if sample x_i belongs to class C_k. The loss function is defined as

$$L_{loss}(\Lambda_S) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} d_k(x_t;\Lambda_S)\cdot I(x_i \in C_k) \qquad (5)$$

where d_k is the misclassification function, and is calculated as

$$d_k(x_i;\Lambda_S) = -q_k(x_i;\Lambda_S) + Q_k(x_i;\Lambda_S) \ (6)$$

where the discriminant function is defined as the log-likelihood of class C_j upon observing pattern x_i as

$$q_j(x_i;\Lambda_S) = \log p_{svm}(C_j \mid x_i) \qquad (7)$$

The anti-discriminant function with respect to class C_k upon observing pattern x_i is then calculated as

$$Q_k(x_i;\Lambda_S) = \log\left\{ \frac{1}{K-1}\sum_{j,j\neq k} \exp(\eta\cdot q_j(x_i;\Lambda_S)) \right\}^{\frac{1}{\eta}} \quad (8)$$

4. Performance Evaluation

To evaluate the SRL performance of the proposed method, the CoNLL-2005 SRL corpus was used. The ATIS-3 (Air Travel Information System) corpus, obtained from Penn Treebank-3 (LDC99T42), was used for evaluation on the performance of intent classification. For efficient comparison to the proposed method, all the experiments were trained using sections 15-18 of CoNLL-2005 SRL corpus. SVM^{light} [14] was adopted to implement the base SVM classifiers using polynomial kernel of degree 4.

4.1. Contributions of Structural Features for SRL

In this section the influence of four important structural features (Path, ApexPath, Position, and Bernoulli event features) on the performance of Semantic Role Labeling were evaluated because this kind of features comprise the majority of the entire feature space, where Bernoulli Event features are newly introduced in this research. Each of the four semantic role labelers is trained using the whole feature space except one of the four structural features without feature selection,

Table 3. *Detailed Results of Semantic Role Labeling.*

WSJ	Accuracy	Precision	Recall	F_1
A0	98.23%	69.96%	83.29%	76.05
A1	98.61%	64.36%	72.06%	67.99
A2	99.77%	65.57%	60.45%	62.91
A3	99.96%	55.88%	37.25%	44.71
AM-ADV	99.88%	56.79%	54.12%	55.42
AM-DIR	99.96%	49.21%	59.62%	53.91
AM-DIS	99.75%	64.74%	75.85%	69.85
AM-LOC	99.97%	42.86%	09.38%	15.38
AM-MNR	99.87%	57.63%	66.67%	61.82
AM-MOD	99.96%	94.32%	97.81%	96.03
AM-NEG	99.99%	55.56%	55.56%	55.56
AM-TMP	99.77%	71.00%	80.89%	75.63
R-A0	99.78%	64.81%	77.83%	70.72
R-A1	99.79%	52.13%	72.73%	60.73
R-A2	99.97%	42.11%	36.36%	39.02
R-AM-LOC	99.98%	66.67%	47.37%	55.38
R-AM-MNR	99.99%	11.11%	12.50%	11.76
R-AM-TMP	99.93%	60.84%	82.11%	69.90
V	99.82%	96.30%	99.92%	98.08

while sample selection is applied to all the four labelers for the purpose of training efficiency.

Table 2 shows the contributions of each feature type to each semantic role. The values presented in this table indicate the performance improvement when the feature type denoted is employed. If the value is negative, it means that the introduction of the feature type has a negative influence on the performance of that semantic role. All the four structural features have detrimental effect on the classification of roles A3, AM-DIR, AM-LOC, AM-NEG, R-A2, and R-AM-LOC because there is no enough training and test samples for these roles. Even so, all the four structural feature types have improved the performance of all the other semantic roles including major roles A0, A1, A2, and AM-TMP.

4.2. Results of Semantic Role Labeling

Table 3 shows the results of final semantic role labeler using the proposed feature selection method. The number of features selected in this experiment is 40,000 as suggested in the preliminary experiment shown in Table 1. It can be seen from Table 3 that although all the classifiers have a very high accuracy in classifying their respective roles, the F-score of some roles such as AM-LOC and R-AM-MNR still have very poor performance.

4.3. Results of Intent Classification

The results of intent classification were shown in Table 4, where values in parentheses denote that the experiment was conducted using transcribed speech. “Supervised” indicates using manually annotated data for classifier training. It can be seen with the introduction of Bernoulli event features of word and POS sequences, the performance of intent classification can be further improved using the already improved SRL results with discriminative feature selection.

5. Conclusions

In this study we have presented an approach to improving the performance of Semantic Role Labeling using a subset of features with discriminative feature selection. Classifiers are known to degrade their performance when facing a huge number of features as presented in this research and previous work. To improve the performance of the classifiers for Semantic Role Labeling, the proposed discriminative feature selection method considers not only seeking better representative feature subset, but also the discriminativity of a feature among different classes.

Table 4. *Results of Intent Classification.*

F-Score	Baseline	Plus Bernoulli Features
Proposed Method	71.3 (77.5) %	74.7 (80.4) %
Supervised	83.9 (89.1) %	84.3 (89.7) %

The experimental results showed that the proposed approach improves SRL performance with the introduction of Bernoulli event features and discriminative feature selection. For spoken language understanding, the experiments on ATIS corpus using a semi-supervised approach also shows the effectiveness of using Bernoulli event features for the task of intent classification.

6. References

[1] G. Tur, D. Hakkani-Tur, and A. Chotimongkol, "Semi-supervised learning for spoken language understanding using semantic role labeling," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 232–237.

[2] R. De Mori, F. Bechet, D. Hakkani-Tur, M. McTear, G. Riccardi, and G. Tur, "Spoken language understanding," in *IEEE Signal Processing Magazine*. vol. 25: IEEE, 2008, pp. 50-58.

[3] X. Carreras and L. Marquez, "Introduction to the CoNLL-2005 shared task: Semantic role labeling," in *Proceedings of CoNLL-2005*, 2005, pp. 152-164.

[4] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*: Springer, 1998.

[5] G. Isabelle, Andr, and Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[6] L. Wang, "Feature Selection with Kernel Class Separability," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, pp. 1534-1546, 2008.

[7] A. L. Berger, V. J. Della Pietra, and S. A. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics,* vol. 22, pp. 39-71, 1996.

[8] K. Kira and L. A. Rendell, *The feature selection problem: Traditional methods and a new algorithm*: John Wiley & Sons Ltd., 1992.

[9] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The ATIS spoken language systems pilot corpus," in *Proceedings of the workshop on Speech and Natural Language*, 1990, pp. 96-101.

[10] D. A. Dahl, M. Bates, M. Brown, W. Fisher, K. Hunicke-Smith, D. Pallett, C. Pao, A. Rudnicky, and E. Shriberg, "Expanding the scope of the ATIS task: The ATIS-3 corpus," in *Proceedings of ARPA Human Language Technology Workshop'92*, 1992, pp. 45–50.

[11] V. Punyakanok, P. Koomen, D. Roth, and W. Yih, "Generalized inference with multiple semantic role labeling systems," in *Proceedings of CoNLL-2005*, 2005, pp. 181-184.

[12] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2001, pp. 311-318.

[13] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of ARPA Workshop on Human Language Technology*, 2002, pp. 138-145.

[14] T. Joachims, "SVMLight: Support Vector Machine," *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund,* 1999.