Unilateral Semi-supervised Learning of Extended Hidden Vector State for Persian Language Understanding

Fattaneh JABBARI, Hossein SAMETI, and Mohammad Hadi BOKAEI
Speech Processing Lab, Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
Fataneh.jabari@gmail.com, sameti@sharif.edu, m.hadibokaei@gmail.com

***Abstract*—The key element of a spoken dialogue system is Spoken Language Understanding (SLU) part. HVS and EHVS are two most popular statistical methods employed to implement the SLU part which need lightly annotated data. Since annotation is a time consuming, we present a novel semi-supervised learning for EHVS to reduce the human labeling effort using two different statistical classifiers, SVM and KNN. Experiments are done on a Persian corpus, the University Information Kiosk corpus. The experimental results show improvements in performance of semi-supervised EHVS, trained by both labeled and unlabeled data, compared to EHVS trained by just initially labeled data. The performance of EHVS improves 13.41% in the case of SVM classifier and 5.16% in the case of KNN. This demonstrates effectiveness and feasibility of the proposed approach.**

Keywords-spoken language understanding; semi-supervised learning; extended hidden vector state; statistical classifier

I. INTRODUCTION

The key element of a spoken dialogue system is spoken language understanding (SLU) part. The main function of this unit is to convert the input utterances into semantic information. Conventionally, this task was done by using hand-crafted semantic grammars. Such methods were fragile, laborious, expensive, error-prone, and needed a lot of expertise. Recently, data-driven (or statistical) approaches have been proposed for the SLU task, which are more robust, portable and less costly to build for a new domain. Finite State Tagger was an early data-driven model which applied on AT&T's CHRONUS system [1] which was straightforward but unable to capture long range dependencies since it was flat concept. BBN's Hidden Understanding Model (HUM) [2] and hierarchical HMMs [3] solve this dilemma by Probabilistic Context Free Grammars (PCFG), which makes finite state networks recursive. These models require fully annotated Treebank data and tractability concerns may raise. The Hidden Vector State (HVS) parser was proposed in [4] to make a tradeoff between these two ends, the flat concept and fully recursive model. It is trainable by a lightly annotated data without the requirement to Treebank. Later [5] uses three techniques to improve the performance of the HVS parser. The main drawback of statistical methods is the need for big amount of labeled data. Since gathering the unlabeled data is easy, semi-supervised learning is a good way to reduce the human labeling effort. Some semi-supervised learning methods include EM with generative mixture models [6], self-training [7], co-training [8], transductive support vector machines (TSVM) [9], etc. To see the details of semi-supervised learning, one may refer to [10].

In this paper, we apply a novel Unilateral Semi-supervised Learning approach on EHVS model by exploiting the unlabeled data by means of a classifier. The remainder of this paper is organized as follows. In Section 2 the HVS parser is introduced and its mathematical framework is discussed. Section 3 describes the EHVS model. Unilateral Semi-supervised Learning on EHVS parser is then introduced in Section 4. The University Information Kiosk corpus is introduced in Section 5. Section 6 reports the experimental results by EHVS and Unilateral Semi-supervised EHVS. Finally, Section 7 concludes the paper.

II. HIDDEN VECTOR STATE PARSER

The HVS parser is a statistical parser and is considered as an expansion of the discrete Markov model. Considering the semantic annotation as a parse tree, the vector states of each word could be achieved by starting from the pre-terminal node and pass through the tree to reach the root. Considering these vector states as hidden variables, the parse tree is equivalent to a first order vector state Markov model, namely HVS. Each vector state is similar to a snapshot of a push-down automaton and the transition between different snapshots can be viewed as stack shift operations. The aim of the HVS is to find the most likely vector state sequence of semantic concepts C^* given the word sequence W:

$$C^* = \text{argmax}_c\, P(C \mid W) = \text{argmax}_c\, P(W \mid C)\,.\,P(C)\,, \qquad (1)$$

where $P(C)$ is semantic model and is formulated as in (2):

$$P(C) = \prod_{t=1}^{T} P(pop_t \mid c_{t-1})\,.\,P(c_t[1] \mid c_t[2,...,n])\,, \qquad (2)$$

where pop_t is the stack shift operation and takes values in range *0* to *n*. In many real applications *n* is set to be four. c_t represents the vector state at time *t*, $c_t[1]$ is the pre-terminal semantic concept for word w_t and $c_t[n]$ is the root node of the semantic parse tree. Moreover, $P(W|C)$ is the lexical model. More details regarding HVS are described in [4].

978-1-61284-729-0/11/$26.00 ©2011 IEEE

III. EXTENDED HIDDEN VECTOR STATE PARSER

The EHVS contains three properties more than the HVS [5]. These features are Negative Example, Left-Right Branching [11], and Input Parameterization [12].

IV. UNILATERAL SEMI-SUPERVISED LEARNING OF EHVS

Because labeling data in statistical methods is quite time consuming and difficult, we apply a novel semi-supervised machine learning method to the EHVS parser to alleviate this problem. The goal of the semi-supervised learning is to exploit the unlabeled data to improve the performance of the model and hence reduce the human labeling effort. We propose a novel Unilateral Semi-supervised Learning method for EHVS parser based on pattern classification. Considering the abstract annotation of each sentence as a class label for that sentence, the semantic tagging problem can be considered as a traditional classification problem. First we train an initial statistical classifier by the labeled sentences; the classes here are the semantic annotation of these sentences. Then this classifier is used to label (or classify) the unlabeled data automatically. These automatically labeled data are then used to improve the performance of the EHVS parser in an iterative manner. Given some amount of human-labeled training data S_l and some unlabeled data S_u , the overall procedure of this Unilateral Semi-supervised EHVS parser is as follows which is also shown in Fig. 1:

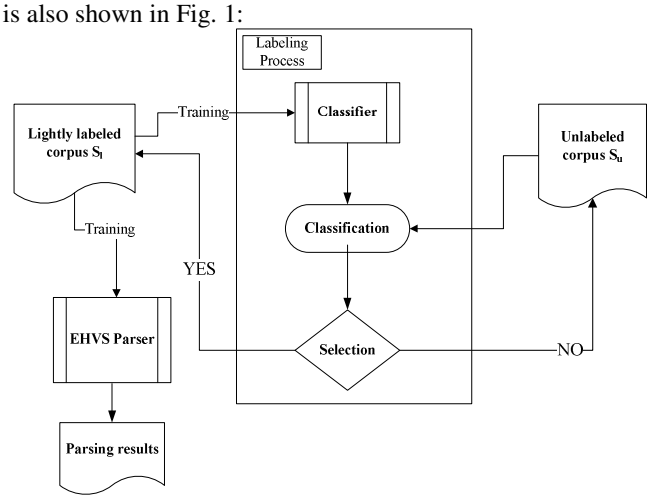

Figure 1. Overall procedure of the Unilateral Semi-supervised EHVS Parser

1. Construct an initial statistical classifier and EHVS parser based on labeled training sentences in S_l.
2. Label the unlabeled sentences in S_u using the trained classifier.
3. Select the automatically labeled sentences by the classifier which have confidence score greater than a specified threshold and add them to the labeled sentences in S_l. Return the labeled sentences with confidence score lower than the threshold unlabeled to the S_u.
4. Retrain the classifier and EHVS parser with new S_l.
5. While the performance of EHVS does not decrease and automatically labeled sentences with desired confidence score are available, repeat from 2.

This approach is named Unilateral Semi-supervised Learning because two models are used for learning but just the classifier helps to improve the performance of both models and EHVS helps to neither of them, thus this is completely different from co-training. Two important issues should be considered in this approach. The first critical issue for the effectiveness of the algorithm is the choice of the classifier, because the improperness of the classifier can lead to adding noisy data to the model and reduces the model performance. The second important issue is to define a reasonable confidence measure according to the chosen classifier in order to determine the correctness of the labeled sentences.

Here, we examine two types of statistical pattern classifiers, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). SVM is a supervised binary classifier. Given a set of training data and their corresponding targets; the SVM algorithm finds a separating hyper-plane with the maximal margin. Sequence tagging in this problem is a multi-class problem. To use SVM classifier in a multi-class problem, we use one-against-one or pair wise coupling classification method where each pair of classes is separated by trained classifier. Another factor in SVM is the selection of kernel function. It should be selected according to the dimensionality of feature space and data size. In this paper linear kernel is used, since the number of features are large enough compared to the dataset size. Another classifier applied here is KNN which is a non-parametric density estimation technique. In this method test samples are classified based on their similarity with training data such that, for the test sample s, k closest neighbors of it from the training set are selected and then assign the most frequent class within these k neighbors to it.

Once the unlabeled sentences are labeled by the classifier, we should determine a proper confidence score for each labeled sentence to determine the correctness of it, because adding sentences with incorrect annotation degrades the performance of the model. For the SVM classifier, we use the estimation of maximum class probability for multi-class classification by pair wise coupling [13], which can be provided by LIBSVM [14] toolkit. Confidence score of Labeled sentence s is computed as follows:

$$CM(s) = \max_j P(l_j \mid s) \quad , \tag{3}$$

where l_j is the j^{th} label or class and the maximum operation is computed among all possible classes. As mentioned earlier, the classes are all possible semantic annotations that appear in training set.

To measure the confidence of the KNN classification an estimator of posterior probability $P(l_i \mid s)$ can be used as [15] :

$$P(l_i \mid s) = \frac{k_i}{k} \quad , \tag{4}$$

where k_i is the number of samples which belong to class l_j among the k nearest neighbors of the test sample s. However, this estimator is applicable when k and number of training samples approaches to infinity. Cover and Hart [15] proved that in the limiting case, the distance of a test point to its nearest neighbor approaches to zero with probability one. Because of finite sample size in many practical applications, using this estimator is not appropriate, and it is better to utilize the distance information in confidence computation. So, a distance weighted score is used [16]:

$$CM(s) = \frac{(\sum_{y_i \text{ of class } l_j} v_i)}{(\sum_{i=1}^{k} v_i)} \quad , \tag{5}$$

where v_i is the weight related to distance of test sample s and its i^{th} nearest neighbor. The weight v_i can be calculated by a linear interpolation between the nearest neighbor of test sample s, i.e. y_1, and it's farthest neighbor y_k :

$$v_i = \frac{d(s, y_k) - d(s, y_i)}{d(s, y_k) - d(s, y_1)} \quad , \tag{6}$$

where d denotes the distance between two instances. This confidence score is bounded to [0, 1] and is equal to one when all of its k nearest neighbors belong to the same class.

V. University Information Kiosk Corpus

The University Information Kiosk corpus is a small corpus containing 268 spoken sentences annotated abstractly and hierarchically with semantic annotations. The vocabulary size of the entire corpus is 184 and vocabulary of lemmatization of the sentences contains 138 lemmas. It contains 12 different semantic annotations from 7 main semantic categories: COURSE, FIELD, OFFICE-NUM, LAB-NUM, OFFICE-TEL, LAB-TEL, and GROUP. The following table shows some examples of some categories.

TABLE I. Table of University Information Corpus with examples of Some Main categories

Category	***Persian sentence***
	English translation
	Semantic annotation
COURSE	مي‌خواستم بدونم آقاي دكتر محمدي اين ترم چه درس‌هايي را ارائه داده است.
	I want to know which courses Mr. Mohammadi is presenting this term.
	COURSE(TITLE, NAME)
FIELD	تحقيقات آقاي محمدي در چه زمينه‌هايي است؟
	What are the research fields of Mr. Mohammadi?
	FIELD(TITLE, NAME)
OFFICE-NUM	لطف كنيد شماره اتاق آقاي محمدي را به من بدهيد.
	Please give me the room number of Mr. Mohammadi.
	NUM (OFFICE (TITLE, NAME))

IV. Experiments And Results

Experiments are done on the University Information Kiosk corpus which has 268 sentences. We take 10% of the data randomly as test set and this is fixed through all iterations of the experiments on EHVS. The remaining 240 sentences are divided into two parts randomly; one part is considered as labeled sentences and the other part is considered as unlabeled ones. It is important to note that in test set, labeled set, and unlabeled set instances from all main categories are included. The performance of the EHVS parser is measured by Concept-Accuracy measure. It actually computes the minimum number of substitutions (S), deletions (D), and insertions (I) to convert a semantic tree to another one. It is computed as below:

$$\text{Concept-Accuracy} = \frac{(N - D - I - S)}{N} * 100\% \cdot \tag{7}$$

The EHVS parser was implemented using the Graphical Model Toolkit (GMTK) [17]. The input feature a vector to EHVS was composed of the observation words themselves, their lemmas, and their related part-of-speech tags. Generations of the left-right branching semantic trees were also allowed in the experiments.

To evaluate the proposed method performance, the baseline EHVS parser was trained using labeled sentences and its Concept-Accuracy on the test set was computed. Then to improve the performance of the initial EHVS parser, the SVM/KNN classifiers were trained using the labeled set as the training set. These classifiers automatically label the sentences in unlabeled set. Next, the automatically labeled sentences with confidence score greater than the specified threshold was selected to be added to the labeled corpus and those with confidence score lower than that remained unlabeled in unlabeled set. Afterward, EHVS and SVM/KNN were re-trained by all previously and newly labeled sentences. The procedure was stopped when the EHVS performance was decreased or no more automatically labeled sentences with desired confidence score was available. Threshold 0.5 seems reasonable for both classifiers. In the case of SVM, the confidence measure is class probability and probability more than 0.5 is sensibly acceptable, and in the case of KNN the confidence score is normalized to [0, 1] and 0.5 seems to be a fair threshold.

To use SVM/KNN for automatic labeling, each utterance should be converted to a binary feature vector. Three types of feature vectors can be used to train the SVM/KNN classifier:

a. Each element of the feature vector corresponds to a word in vocabulary of the given corpus. If the word exists in the sentence, the corresponding feature of that word becomes one in feature vector, otherwise it will be zero.

b. Each element of the feature vector corresponds to a word in the vocabulary of the lemmatized corpus. If the lemma exists in the given utterance its related feature will be one, otherwise it will be zero.

c. In addition to the feature type b, part-of-speech tags of the utterances are included.

The leave-one-out cross-validation (LOOCV) was applied on labeled corpus with 120 sentences to select the proper feature set for each of the classifiers. Table 2 shows the SVM/KNN accuracy achieved by each type of feature vectors using LOOCV. Because the classifier accuracy for feature type b was more than the other two types, it was used to train both classifiers for automatic labeling.

TABLE II. SVM/KNN accuracy on labeled corpus by LOOCV for all feature types

Feature Type	**SVM accuracy by LOOCV on labeled set**	**KNN accuracy by LOOCV on labeled set (k=10)**
a	86.66%	60.83%
b	91.66 %	70.00%
c	88.33%	55.83%

Table 3 shows the performance of Unilateral Semi-supervised EHVS Parser. Baseline-EHVS is the one with only 120 labeled training sentences S_l and Best-EHVS shows the performance of EHVS when training from both S_l and S_u when all sentences in S_u are also manually labeled. $EHVS_i$ means the EHVS at i^{th} iteration of semi-supervised approach.

TABLE III. PERFORMANCE OF UNILATERAL SEMI-SUPERVISED EHVS PARSER

EHVS Parser	EHVS Concept Accuracy by SVM	EHVS Concept Accuracy by KNN
Baseline-EHVS	32.98%	32.98%
$EHVS_1$	44.33%	34.02%
$EHVS_2$	46.39%	34.02%
$EHVS_3$	46.39%	37.11%
$EHVS_4$	23.40%(stop!)	38.14%
$EHVS_5$	-	37.11%(stop!)
Best-EHVS	55.32%	55.32%

As the experimental results shows, the performance of Baseline-EHVS was improved by applying the proposed semi-supervised method using the automatically labeled sentences by SVM classifiers. Here, 46.39% of Concept-Accuracy was achieved which means 13.41% improvement was gained in comparison with the Baseline-EHVS. Similarly, using the KNN, the maximum Concept-Accuracy achieved was 38.14% which is again greater than the Baseline-EHVS performance.

Although the changes in performance of SVM/KNN are not important in the proposed algorithm, we investigate the changes in performance of the classifiers during the iterations. Table 4 shows the changes of SVM/KNN performance on training set as the algorithm proceeds. Base-Set is the initial human-labeled train set which contains 120 abstractly annotated sentences. $Train\text{-}Set_i$ contains the previously and newly labeled sentences at the i^{th} iteration.

TABLE IV. LOOCV PERFORMANCE OF SVM/KNN AT ITERATIONS OF UNILATERAL SEMI-SUPERVISED LEARNING

Training sets	SVM		KNN	
	# of sentences	Accuracy	# of sentences	Accuracy
Base- Set	120	91.66%	120	70.00%
$Train\text{-}Set_1$	200	94.94%	166	80.12%
$Train\text{-}Set_2$	208	95.28%	174	79.89%
$Train\text{-}Set_3$	213	95.30%	176	79.55%
$Train\text{-}Set_4$	_	_	180	80.00%

V. CONCLUSION AND FUTURE WORKS

In this paper a novel Unilateral Semi-supervised Learning was proposed to improve the performance of the EHVS parser. In this method the human labeling effort for the task of statistical spoken language understanding reduces by applying a classifier in order to produce semantic annotations for the unlabeled data automatically in an iterative manner, and hence improve the performance of the initial model. The experimental results show the effectiveness and feasibility of the proposed approach since it improves the Concept-Accuracy of the initial EHVS model. The future works include application of active learning and its combination with semi-supervised learning to further improve the performance and reduce the human labeling task. Moreover, for more accurate labeling, the results of different classifiers along with their corresponding confidence scores can be combined.

REFERENCS

[1] E. Levin, and R. Pieraccini, "CHRONUS, the next generation," In: Proceedings of the DARPA Speech and Natural, Language Workshop, Austin, Texas, pp. 269–271, 1995.

[2] S. Miller, R. Schwartz, R. Bobrow, and R. Ingria, "Statistical Language Processing Using Hidden Understanding Models," In: Proceedings of the workshop on Human Language Technology , Plainsboro, NJ, pp. 278-282, 1994.

[3] S. Fine, Y. Singer, and N. Tishby, "The hierarchical hidden Markov model: Analysis and applications," Machine Learning, Springer, vol. 32, no. 1, pp. 41–62, 1998.

[4] Y. He, and S. Young, "Semantic processing using the Hidden Vector State Model," Computer Speech and Language, Elsevier, vol. 19, no. 1, pp. 85–106, 2005.

[5] J. Svec, and F. Jurcicek, "Extended Hidden Vector State Parser", In: Proceedings of the 12th International Conference on Text, Speech and Dialogue, Plzen, Czech Republic, pp. 403-410, 2009.

[6] K. Nigam, AK. McCallum, S. Thrun, and TM. Mitchell, "Text classification from labeled and unlabeled documents using EM". Mach Learn 2000; vol. 39, no. 2, pp.103-34.

[7] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods". In: Uszkoreit H, editor. Proceedings of the 33rd annual meeting of the association for computational linguistics. Morristown, NJ, USA: Association for Computational Linguistics; 1995. pp. 189-96.

[8] A. Blum, T. Mitchell, "Combining labeled and unlabeled data with co-training". In: Bartlet P, Mansour Y, editors. Annual workshop on computational learning theory, Proceedings of the eleventh annual conference on computational learning theory. New York, NY, USA: ACM Press; 1998 . pp. 92-100.

[9] L. Xu, D. Schuurmans, "Unsupervised and semi-supervised multi-class support vector machines". In: Veloso MM, Kambhampati S, editors. Proceedings of the twentieth national conference on artificial intelligence. Menlo Park, California, USA: The AAAI Press; 2005. pp. 904-10.

[10] X. Zhu, "Semi-supervised learning literature survey". Technical Report 1530, Computer Sciences Department, University of Wisconsin-Madison; 2005.

[11] F. Jurcicek, J. Svec, and L. Muller, "Extension of HVS Semantic Parser by Allowing Left-Right Branching," In: Proceedings of IEEE ICASSP, Las Vegas, Nevada, USA, pp. 4993-4996, 2008.

[12] J. Svec, F. Jurcicek, and L. Muller, "Parameterization of the input in training the HVS semantic parser," In: Proceedings of the 10th international conference on Text, speech and dialogue, Plzen, Czech Republic, pp. 415-422, 2007

[13] T.F. Wu , C.J. Lin and R.C. Weng, "Probability estimates for multi-class classification by pairwise coupling", Mach. Learning Res., vol. 5, p.975,2004.

[14] C.C. Chang and C.J. Lin, "LIBSVM : a library for support vector machines", 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[15] T.M. Cover, P.E. Hart, "Nearest neighbor pattern classification". IEEE Transactions on Information Theory 13, pp. 21-27

[16] S.A. Dudani, "The Distance-Weighted k-Nearest Neighbor Rule".IEEE Transactions on Systems, Man, and Cybernetics 6, pp. 325-327, 1976

[17] J. Bilmes, and G. Zweig, "The graphical models toolkit: An open source software system for speech and time-series processing," In: Proceedings of IEEE ICASSP, pp. IV-3916-IV-3919, 2002.