García Granada, F.; Hurtado Oliver, LF.; Sanchís Arnal, E.; Segarra Soriano, E. (2011). An active learning approach for statistical spoken language understanding. En Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer Verlag (Germany). 7042:565-572. doi:10.1007/978-3-642-25085-9_67.

An Active Learning Approach for Statistical Spoken Language Understanding

Fernando García, Lluís-F. Hurtado, Emilio Sanchis, and Encarna Segarra

Grup d'Enginyeria del Llenguatge Natural i Reconeixement de Formes,
Department de Sistemes Informàtics i Computació,
Universitat Politècnica de València, València, Spain
{fgarcia,lhurtado,esanchis,esegarra}@dsic.upv.es
http://dsic.upv.es/users/elirf

Abstract. In general, large amount of segmented and labeled data is needed to estimate statistical language understanding systems. In recent years, different approaches have been proposed to reduce the segmentation and labeling effort by means of unsupervised o semi-supervised learning techniques. We propose an active learning approach to the estimation of statistical language understanding models that involves the transcription, labeling and segmentation of a small amount of data, along with the use of raw data. We use this approach to learn the understanding component of a Spoken Dialog System. Some experiments that show the appropriateness of our approach are also presented.

Keywords: active learning, unaligned corpus, spoken language understanding, spoken dialog systems

1 Introduction

One of the most important drawbacks in almost all the corpus-based approaches to the development of Spoken Language Understanding (SLU) systems is the effort that is necessary to manually transcribe, segment and label a training corpus, process that is essential in this kind of approaches. Manual segmentation and labeling, apart from the time-consuming work, has the disadvantage that sometimes it is difficult to decide a-priori which limits of the segments are more accurate to represent a specific semantic label and to better discriminate from other semantic labels. Despite of this laborious and time-consuming process of preparation of training data, statistical models have been widely used in recent years in the Spoken Language Understanding (SLU) area, mainly in the framework of spoken dialog systems, and they have shown good performances [7], [4], [1], and [3].

Moreover, automatically training an understanding model from a segmented and labeled corpus is a static learning process and it is not possible to adapt the model to new kinds of interactions or to new ways to express the concepts. This is why in recent years different techniques have been proposed to reduce the labeling effort by means of unsupervised o semi-supervised learning techniques

and to have the possibility of dynamically adapt the models when the system is interacting with the users in order to allow for an active learning process [6], [8] and [2]. Active learning aims at reducing the number of training examples to be labeled by selectively sampling a subset of the unlabeled data. This is done by inspecting the unlabeled examples and selecting the most informative ones, with respect to a given cost function. Active learning is well-motivated in many modern machine-learning problems, where unlabeled data may be abundant or easily obtained; however, the labeling process is difficult, time-consuming, and expensive.

In this paper we present an approach to SLU that is based on automatic learning of statistical models. In previous versions of our SLU system [7], all the transcribed training corpus was manually segmented and labeled in terms of semantic labels. In the present approach we propose to apply an active learning process to estimate a SLU system which requires only the transcription, segmentation and labeling of a small set of training user utterances.

We propose a two-step approach to the estimation of statistical language understanding models that involves the transcription, segmentation and labeling of a small amount of data (recognized user utterances), along with the use of raw (untranscribed, unsegmented and unlabeled) recognized user utterances. In the first step, from a small corpus of unaligned pairs of recognized sentences and their corresponding semantic representation (frames), we have applied a semi-supervised process [5] obtaining an automatic segmentation of the corpus. From the segmented and labeled sentences of that small corpus, a baseline statistical language understanding model is estimated using an automatic method [7]. In the second step, we incrementally update this baseline language understanding model with more segmented and labeled sentences following an active learning process. A set of new recognized user utterances is automatically segmented and labeled with the baseline statistical language understanding model. According to a confidence measure criterium obtained during the understanding process, a small number of these new sentences (the least reliable ones) are manually transcribed, segmented and labeled by an expert, and together with the automatically segmented and labeled sentences, are used to retrain the baseline statistical language understanding model. This process is repeated for another set of raw sentences, but, this time, the retrained statistical language understanding model is used.

The SLU model used [7] is based on a two-level statistical model, in which both the probabilities of sequences of semantic labels and the lexical realization (that is, the sequences of words associated) of each semantic label are represented. Some confidence measures generated in this decoding process are used to automatically detect sentences that can be candidates for manual labeling. This way only a few of the new sentences are manually labeled, while the sentences that are decoded with high confidence are automatically included in the new training corpus.

Some experiments were performed over a task of information about train timetables and prices in Spanish. The experiments show the accuracy of the

proposed learning methods that provides similar results to those obtained from a completely segmented and labeled training corpus. Thus we have the possibility of having a system that can be dynamically adapted while it is used by real users, whereas the effort employed to obtain the models is not comparable with the effort of manually to transcribe, segment and label the full training corpus.

This paper is organized as follows, Section 2 describes the SLU process using the two-level statistical model. Section 3 describes the initial automatic semi-supervised segmentation process and the process of incrementally updating the SLU model through an active learning approach. Section 4 presents the evaluation of our proposal on the Corpus of Dihana, a Spoken Dialog System to access a railway information system using Spontaneous Speech in Spanish. And finally, Section 5 presents the conclusions.

2 Speech understanding

We have proposed a method for speech understanding based on the use of stochastic models automatically learned from data. The main characteristic of our method is the integration of syntactic and semantic restrictions into one finite-state automaton. To learn syntactic and semantic models a corpus of segmented and labeled sentences is required. Each sentence in the corpus must be segmented and a label (from a set of semantic labels V defined for the task) must be assigned to each segment. The label assigned to each segment represents the semantic information provided by this segment.

From the segmented and labeled corpus two types of finite-state models are learned. A model A_{s} for the *semantic language* is estimated from the sequences of semantic labels associated to the input sentences. A set of models, *syntactic models* $A_{\mathrm{v_i}}$ (one for each semantic label $v_{\mathrm{i}} \in V$), is estimated from all the segments of words assigned to this semantic label.

In order to perform the understanding process, a global automaton A_{t} is generated by combining the semantic model with the syntactic ones. The states of the semantic automaton A_{s} are substituted by their corresponding stochastic automata $A_{\mathrm{v_i}}$.

Given the input sentence $w = w_1 w_2 \dots w_n \in W^*$, the understanding process consists of finding the sequence of semantic labels $v = v_1 v_2 \dots v_k \in V^*$ which maximizes the probability:

$$\widehat{v} = \operatorname*{argmax}_{v} P(w|v)P(v)$$

Where, $P(v)$ is the probability of the sequence of semantic labels v and $P(w|v)$ is the probability of the sequence of words w given the sequence of semantic labels v. We approach this latter probability as the maximum for all possible segmentations of w in $|v|$ segments.

$$P(w|v) = \max_{\forall l_1, l_2, \dots l_{k-1}} \{P(w_1, ..., w_{l_1}|v_1) \cdot P(w_{l_1+1}, ..., w_{l_2}|v_2) \cdot ... \cdot P(w_{l_{k-1}+1}, ..., w_n|v_k)\}$$

The understanding process is performed using the Viterbi algorithm, which supplies the best path through A_t that is able to produce the input sentence w. From this path the sequence of semantic labels and the most likely segmentation of the input sentence associated to it can be easily obtained. More details of our approach to speech understanding can we found in [7].

2.1 Semantic representation for the DIHANA task

Although our method is generic, a specific set of semantic labels must be defined for each task. In addition, once the segmentation of the sentence is performed a second phase is required. This second phase is devoted to reordering the semantic labels following a canonical order and instantiating some values, mostly related to hours and dates.

During the DIHANA project a corpus of 900 dialogs was acquired using the Wizard of Oz technique. Four dialogs were acquired for each of the 225 users who cooperated in the acquisition process. The chosen task was the access to an information system using spontaneous speech. The information system provided information about railway timetables, fares, and services. The system was accessed by telephone in Spanish. The number of user turns acquired was 6 280 and the vocabulary size was 823 different words.

The semantic representation chosen for the task was based on frames. The understanding module takes the sentence supplied by the automatic speech recognizer as input and generates one or more frames (which are concepts with their corresponding attributes) as output. The frames are obtained after reordering the semantic labels from the best segmentation of the sentence and instantiating certain values as stated above. A total amount of 25 semantic labels were defined for DIHANA task. In order to label segments without semantic, a *null* label was also added the the label set.

Ten labels related to frame concepts, divided in two different types, were defined:

1. *Task-independent concepts*: *(ACCEPTANCE)*, *(REJECTION)*, and *(NOT-UNDERSTOOD)*.
2. *Task-dependent concepts*: *(HOUR)*, *(DEPARTURE-HOUR)*, *(ARRIVAL-HOUR)*, *(PRICE)*, *(TRAIN-TYPE)*, *(SERVICES)*, and *(TRIP-DURATION)*.

The task-independent concepts represent generic interaction acts which could be used for any task. The task-dependent concepts represent the information the user can ask for. In an user turn, each task-dependent concept can include one or more attributes from a set of fifteen. These attributes represent the constraints that the user can place on his query.

The fifteen attributes defined for the DIHANA task are: *City*, *Origin-City*, *Destination-City*, *Class*, *Train-Type*, *Num-Relative-Order*, *Price*, *Services*, *Date*, *Arrival-Date*, *Departure-Date*, *Hour*, *Departure-Hour*, *Arrival-Hour*, and *Trip-Type*.

Two examples of the semantic representation, translated from the original Spanish DIHANA corpus, are shown below:

"I want to know the timetable on Friday to Barcelona, on June 18th"
(HOUR)
Destination: Barcelona
Departure-Date: (Friday)[18-06]

"yes, the fares from Valencia"
(ACCEPTANCE)
(PRICE)
Origin: Valencia

3 The active learning process

The goal of the active learning process is to obtain good models by labeling only a small part of the training samples. It also permits the models be dynamically adapted when real users interact with the system. As this process is a kind of bootstrapping process we need to start from an initial model that must be learned using a small set of labeled training samples. Even in this preliminary step of the learning process we avoid the effort of the manual segmentation of the corpus, that is, we only need the pair (sentence, semantic representation in terms of frames) without the explicit association of semantic labels to the segments of the sentence. To do this, we have developed a semi-supervised learning algorithm [5] that associates to each semantic label a set of segments of different lengths based on the co-occurrences of segments and semantic labels. That is, given a fixed length l, $P(v_k|u_l)$ is calculated for every segment of length l, u_l, and every semantic label, v_k, in the training corpus. Then, those segments with $P(v_k|u_l) > theshold$ are considered to belong to v_k.

As the training corpus is small, it is necessary to increase the coverage in order to include more linguistic variability that it is not present in the corpus. To do so, a procedure of categorization, lematization, and semantic generalization based on dictionaries is applied. This is the case for example of the segment "*quiero ir a*" (*I want to go to*) that is generalized to "*querer ir a*" (*to want to go to*) that includes the Spanish conditional form "*querría ir a*" (*I would want to go to*).

Increasing the length of segments, we can better discriminate between words that are semantically ambiguous by adding context to the segment. For example the word "*Valencia*" in an isolated way can not be associated to a semantic label, while the bigram "*to Valencia*" can easily be associated to the semantic label "destination-city". In our experiments, we have considered segments until length 3.

After applying this semi-supervised algorithm, a first segmented and labeled corpus is obtained. From this training corpus we can learn the semantic models as explained in Section 2, and start the active learning process. This process is

based on detecting what new samples are not well represented in our models, and only these samples will be manually transcribed, analyzed and, if it is necessary, relabeled. That is, by using our current semantic models we analyze a new set of sentences from the automatic speech recognizer and those sentences that are selected by considering a confidence score will be manually corrected.

The confidence measure we have used is based on the probability of the appearance of sequences of words when a semantic label is found. For each pair (u_i, v_i), a linear combination of two measures is used to determine if the assignment of the semantic label v_i to the segment u_i has been done properly during the decoding process:

- $\frac{logP(u_i|v_i)}{|u_i|}$ is the probability of the segment u_i within the semantic label v_i normalized according to the number of words in the segment. This measure is more sensitive to syntactic variations.
- $\frac{log\prod_{w_j \in u_i} P(u_i|v_i)}{|u_i|}$ is the same probability but considering only the unigram probability. This measure is more sensitive to out-of-vocabulary words.

Sentences containing one or more segments with a low value for the linear combination of these measures are manually revised.

4 Experiments

Some experiments were carried out in order to evaluate the appropriateness of the described technique. We used the 80% of the corpus as training and development set and the 20% as test set. In all the experiments, the output of the recognition module of the test sentences was used as the input of the understanding process. The speech recognizer used in the experimentation had a 74% of word accuracy.

We defined two measures to evaluate the performances of the understanding module:

- %cf, is the percentage of correct frames, i.e. the percentage of obtained frames that are exactly the same as the corresponding reference frame.
- %cfs, is the percentage of correct frame units (concepts and attributes).

Two different experiments were done. In the first experiment, using the manually transcribed, segmented, and labeled corpus we trained an understanding model (Section 2). This experiment gives an upper bound of our understanding technique to compare with the results of subsequent experiments. The second experiment measures the behavior of the semi-supervised algorithm and the active learning process (Section 3).

For the second experimentation, four subsets were created splitting the training corpus in order to apply the active learning technique (T25_1, T25_2, T25_3, and T25_4), each one of them contained the 25% of the training corpus. The models learned in each step were stochastic finite-state automaton. The process was as follows:

1. We considered the sentences supplied by the speech recognizer for the first training subset and the semantic representation (in terms of frames) associated to each one of them. An automatic segmentation and labeling process was made using the semi-supervised algorithm. With this labeled data, the first understanding model was trained (T25_1).
2. Using this understanding model, a process of segmentation and labeling of the second training subset was performed.
3. Considering the confidence scores generated in the understanding process, a part of the sentences in the second subset was selected in order to be manually transcribed, segmented, and labeled. Instead of finding a threshold of the confidence scores, we selected the 20% of the segments with the lower confidence score.
4. After the last step a new training corpus was generated. This new corpus consists of the first training subset, the sentences in the second subset that were automatically labeled by the understanding process, and the small part of the second subset (20%) that were manually corrected. With this new corpus a new understanding model was learned (T25_2).
5. We repeated the process for the third and the fourth training subsets (T25_3, T25_4).

The results of the first experiment were 63.8% for the cf measure and 78.2% for the cfs measure. The cf value is higher than the cf value, that is because the cf measure is more strict: an error in one frame unit produces an error in the whole sentence.

Table 1 shows the results of the active learning process. As we can see both measures improve with the increasing of the amount of training data. The results are slightly worse than the results in the reference experiment (Ref column), but the effort of manual segmentation and labeling is much smaller.

Table 1. Results of the active learning process

	T25_1	T25_2	T25_3	T25_4	Ref
%cf	53.1	54.8	56.9	57.9	63.8
%cfs	70.5	73.1	74.5	75.3	78.2

From a training corpus of 5,024 user turns, 1,256 were semantically labeled for the initial semi-supervised process, and 750 additional turns were transcribed, segmented and labeled during the active learning process. This implies a transcription and segmentation of 15% of the training corpus, and semantic labeling of the 40% of the training corpus. System performance has been reduced by less than 3% compared to models using the entire transcribed, segmented, and labeled training corpus.

5 Conclusions

In this paper, we have presented an active learning approach to the estimation of statistical language understanding models which involves the transcription, labeling, and segmentation of only a small amount of data, along with the use of raw data. We have used this approach to learn the understanding component of a Spoken Dialog System for railway information retrieval in Spanish. Experiments show that the results obtained with the proposed method are quite similar to those obtained from a completely segmented and labeled corpus. However, the effort employed to obtain the models is much lower than the effort required for completely transcribing, segmenting, and labeling the training corpus.

Acknowledgements

Work partially supported by the Spanish MICINN under contract TIN2008-06856-C05-02, and by the Vicerrectorat d'Investigació, Desenvolupament i Innovació of the Universitat Politècnica de València under contract 20100982.

References

1. De Mori, R., Bechet, F., Hakkani-Tur, D., McTear, M., Riccardi, G., Tur, G.: Spoken language understanding: A survey. IEEE Signal Processing magazine 25(3), 50–58 (2008)
2. Gotab, P., Bechet, F., Damnati, G.: Active learning for rule-based and corpus-based spoken labguage understanding moldes. In: IEEE Workshop Automatic Speech Recognition and Understanding (ASRU'09). pp. 444–449 (2009)
3. Gotab, P., Damnati, G., Becher, F., Delphin-Poulat, L.: Online slu model adaptation with a partial oracle. In: Proc. of InterSpeech 2010. pp. 2862–2865. Makuhari, Chiba, Japan (2010)
4. He, Y., Young, S.: Spoken language understanding using the hidden vector state model. Speech Communication 48, 262–275 (2006)
5. Ortega, L., Galiano, I., Hurtado, L.F., Sanchis, E., Segarra, E.: A statistical segment-based approach for spoken language understanding. In: Proc. of InterSpeech 2010. pp. 1836–1839. Makuhari, Chiba, Japan (2010)
6. Riccardi, G., Hakkani-Tur, D.: Active learning: theory and applications to automatic speech recognition. Speech and Audio Processing, IEEE Transactions on 13(4), 504 – 511 (july 2005)
7. Segarra, E., Sanchis, E., Galiano, M., García, F., Hurtado, L.: Extracting Semantic Information Through Automatic Learning Techniques. International Journal of Pattern Recognition and Artificial Intelligence 16(3), 301–307 (2002)
8. Tur, G., Hakkani-Tr, D., Schapire, R.E.: Combining active and semi-supervised learning for spoken language understanding. In: Speech Communication. vol. 45, pp. 171–186 (2005)