Adaptive Training for Robust Spoken Language Understanding

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

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València. Spain
{fgarcia,esanchis,lhurtado,esegarra}@dsic.upv.es

Abstract. Spoken Language Understanding, as other areas of Language Technologies, suffers from a mismatching between the conditions of the training of the models and the real use of the systems. If the semantic models are estimated from the correct transcriptions of the training corpus, when the system interacts with real users, some recognition errors can not be recovered by the understanding system. To achieve an improvement in real environments we propose the use of the output sentences from the recognition process of the training corpus in order to adapt the models. To estimate these models, a labeled and segmented corpus is needed. We propose an algorithm for the automatic segmentation and labeling of the recognized sentences considering the correct segmented and labeled data as reference. Experiments with a spoken dialog corpus show that this approach outperforms the approach based on correct transcriptions.

Keywords: Spoken language understanding, learning from noisy data, adaptive training.

1 Introduction

In many applications of language technologies, a mismatch may occur between the conditions of the training of the models and the real use of the systems. This problem can appear in statistical models, which are some of the most common models used to represent the knowledge sources involved in oral communication. Statistical models have the advantage that they can be trained by using automatic learning algorithms and they can accurately represent the variability of many linguistic components, such as acoustic-phonetic, syntactic, or semantic components [11],[4]. However, in many cases it is not possible to have a training corpus that contains all of the linguistic variability necessary to estimate good linguistic components.

To address this problem, some approaches have been proposed in the literature depending on the kind of models to learn or the possibility of obtaining accurate training samples [5]. For example in Automatic Speech Recognition (ASR), the acoustic-phonetic models must be adapted to noisy environments

and the language models must also be adapted to deal with the problems of coverage. In Spoken Language Understanding (SLU), the lack of enough training data is also a common problem.

One possibility for tackling unobserved inputs in real environments is to supply lattices or graphs of linguistic units as input to the systems in order to represent more variability according to similarities between words or phonemes, for instance Word Confusion Networks are used for robust semantic parsing in [14]. Another way to adapt systems to unobserved inputs in real environments is to use active learning techniques [12], in which some real utterances/sentences are selected to dynamically adapt the models.

In the specific case of SLU, one of the main problems is to obtain a semantically labeled corpus [2], that is large enough to train the semantic models. In most cases, even though the training corpus is obtained through real speech interactions, the semantic labeling is generated by considering the correct transcriptions of the utterances. Therefore, even though the semantic labeled training corpus takes into account the variability associated to spontaneous speech, the noise generated by the speech recognition errors is not considered. With semantic models trained in that way, when the system interacts with real users, some ASR errors cannot be recovered by the understanding module. Similar problems occur when a multilingual SLU is designed. In that case the input sentences to the understanding module are corrupted not only by the recognition process but also by the translation process. Solutions to this problem can be found in [3], where the translation process is enriched by the combination of several translators and generating a graph of words that represents multiple hypotheses, or in [9], where the training samples are translated from the original language to the user language and then translated back to the original language. This way they have a training corpus that includes the specific characteristics of the translation process.

In this paper, we present an adaptive training approach to SLU that uses the output sentences from the ASR process of the training corpus in order to adapt the SLU models to the characteristics of the ASR process. In order words, we use the noisy training data, which is obtained from the recognition process, to estimate the semantic models. We have applied this approach to the DIHANA task [1] with information about train timetables and fares in Spanish, and we have studied the behavior of the system by considering three different ASR engines: an open domain (Google recognizer), and two in-domain recognizers (HTK and Loquendo) where both the language model and the vocabulary must be provided. Our training approach is based on the automatic segmentation and labeling of the ASR output taking the correct semantically segmented/labeled data as reference. To do this, an algorithm that segments and labels the ASR output using the Levenshtein distance to the correct transcription has been developed. Two approaches to SLU have been studied: a Conditional Random Field (CRF) approach [10] and a Two-level stochastic model [13], which is based on Stochastic Finite-State Automata. Experiments with the DIHANA corpus

show that this enriched learning approach outperforms the classical approach based on clean (the correct transcriptions) training data.

2 Semantic representation for the DIHANA task

The domain of the DIHANA task is an information system about railway timetables, fares, and services in Spanish. The DIHANA corpus consists of 900 dialogs that were acquired from 225 users using the Wizard of Oz technique. Thus many characteristcs of spontaneous speech are present in the user utterances. The number of user turns acquired was 6,280 and the vocabulary size was 823 words. The semantic representation chosen for the task is based on frames. A total of 25 semantic labels were defined for the DIHANA task, consisting of 10 types of frames (Affirmation, Negation, Price, Hour, Departure-time..) and 15 attributes (City, Origin-City,Destination-City, Class, Train-Type...).

An example of the semantic representation translated from the original Spanish DIHANA corpus is shown below:

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

3 The understanding system

Our understanding system works in two phases (see Figure 1) [13]. The first phase consists of a sequential transduction of the input sentence in terms of an intermediate semantic language. In the second phase, a set of rules transduces this intermediate representation in terms of frames. Since the intermediate language is close to the frame representation, this phase only requires a small set of rules to construct the frame. This second phase consists of the following: the deletion of irrelevant segments of the input sentence, the reordering of the relevant concepts and attributes that appeared in the user sentence following an order which has been defined a priori, the instantiation of certain task-dependent values, etc.

In order to represent the meaning of the sentences in terms of the intermediate semantic language, a set of 64 semantic units was defined. Each semantic unit represents the meaning of words (or sequences of words) in the sentences. For example, the semantic unit *query* can be associated to *"can you tell me"*, *"please tell me"*, *"what is"*, etc. This way, an input sentence (sequence of words) has a semantic sentence (sequence of semantic units) associated to it, and there is an inherent segmentation. An example is shown in Figure 1.

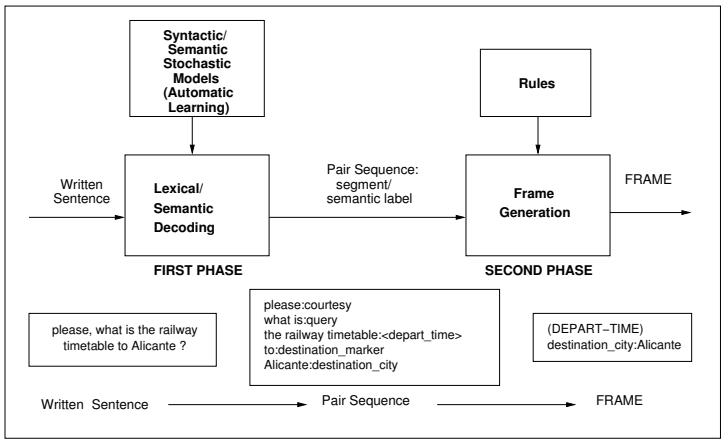

Fig. 1. The understanding process in two phases.

3.1 Semantic models

Two different SLU techniques have been studied to implement the first phase, a generative technique (the Two-level) and a discriminant technique (a classical CRF).

To apply the Two-level technique [13],[6], we assume that each user turn in the training set has a sequence of concepts (semantic units) associated to it, each of these concepts represents a piece of meaning of the user turn, and there is a segment (sequence of consecutive words) in the user sentence that is associated to each of these concepts. This approach consists of learning two types of finite-state models from the training set of pairs (u, c), where u is the sequence of segments and c is the corresponding sequence of concepts.

A model A_{s} for the *semantic language* is estimated from the sequences of concepts c that are associated to the input sentences. A set of models, *concept models* $A_{\mathrm{c_i}}$ (one for each concept c_{i}), is estimated from all the segments of words associated to this concept. The semantic model A_{s} represents the semantic information provided by the training data, and each concept model $A_{\mathrm{c_i}}$ represents the lexical and syntactic information for the corresponding concept c_{i}.

For the understanding process, all the models must be combined in order to take advantage of all the lexical, syntactic, and semantic constraints. To do this, the states of the stochastic automaton A_{s} are substituted by the corresponding stochastic automaton $A_{\mathrm{c_i}}$. Once this integrated automaton A_{t} is built, the understanding process consists of finding the best path in this automaton given the input sentence. In the experimentation, we used a 2-gram model for the A_{s} automaton and for the $A_{\mathrm{c_i}}$ automata.

CRFs have been successfully used for SLU tasks [7]. We defined a set of basic features that includes only lexical information, setting a window such as incorporates the two previous and the two posterior words. A more complete set

of features could be defined for applying the CRFs to SLU tasks [7], however, in this work we have not done a depth study of the best combination of features.

4 Alignment Process

In order to learn the SLU system from noisy samples (that is, with the sentences obtained by the ASR engine) it is necessary to segment and label these new sentences for both approaches, the Two-level and CRFs.

To do this without manual effort, we translate the labeling and segmentation of the correct transcribed sentences to the recognized sentences. This is done by obtaining the Levenshtein distance between the two sentences, which not only supplies the distance but also supplies the word alignment associated to this distance.

Once we have a word-to-word alignment, we can translate the segmentation and labeling to this new sentence. Then, we can learn the concept models by using the original clean data, the new noisy data obtained by the ASR, or a combination of the two. These concept models will be used in the SLU system (Figure 2).

It should be noted that, in SLU, there are some words that are keywords (very relevant) to some concepts, and systematic errors in these specific words can generate many errors in the semantic interpretation. For example in the following sentence:

Correct:	*qué*	*tipo*	*de*	*tren*	*es*	*el*	*más*	*rápido*
ASR output:	*que*	*tipo*	*de*	*tres*	-	*el*	*más*	*rápido*

qué tipo de tren es : <tipo_tren>
el más rápido : tipo_tren

que tipo de tres : <tipo_tren>
el más rápido : tipo_tren

there is an ASR error in the word "*tren*" (train) that has been recognized as the word "*tres*" (three) due to the acoustic similarity in Spanish. If we include the output of the ASR as a semantic training sample, the segment "*que tipo de tres*" will be associated to the concept "*tipo_tren*". This way, it is possible for a similar error to be recovered during the understanding process.

This also occurs with some words specific to the task, such as "*moviendo*" (moving) instead of "*volviendo*" (returning); or "*rosario*" (rosary) instead of "*horario*" (timetable).

5 Experiments

In order to evaluate the effectiveness of the approach we carried out some experiments with the DIHANA corpus. The corpus was split into a training set of 4,887 turns and a test set of 1,340 turns.

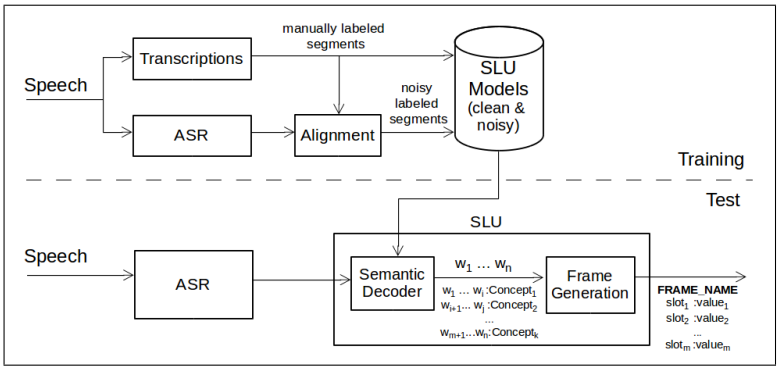

Fig. 2. Scheme of our approach

We studied the behavior of the proposed approach with three different ASR engines: the open domain Google recognizer, and two in-domain recognizers (Loquendo and HTK). The acoustic and language models were learned as follows. In the case of the Google ASR system, there were no options to adapt the models because it is an open domain ASR system with its own acoustic and language models. In the case of Loquendo, which has its own acoustic models, only the language model was learned by using the DIHANA corpus. And in the case of HTK, the acoustic and language models were in-domain models that were learned from the DIHANA training corpus.

For the generation of the semantic corpus using the output sentences from the ASR engines, we have worked in different ways. In the case of Google, since it is an open domain ASR, the new sentences were obtained by just recognizing all the utterances of the corpus. In the case of the other two engines, the language model (LM) was estimated from a part of the training corpus. If we had used the LM obtained from the whole training corpus to recognize it, the ASR results would have been very good, and we would not have a corpus with typical recognition errors. In order to obtain conditions in the new training corpus similar to conditions in the test corpus, we split the training corpus into 10 subsets. Each subset was processed by an ASR system with a LM estimated from the rest of the training sentences. After 10 iterations we had a corpus of sentences that were recognized in conditions similar to the test corpus.

After that, the new recognized sentences were segmented and semantically labeled following the process described in Section 4. The Word Error Rate for the Google recognizer was 27.11, for the Loquendo recognizer was 20.21, and for the HTK recognizer was 17.66. As expected, the more capability to adapt the models to the domain, the less error rate obtained.

We defined two measures to evaluate the accuracy of the models in the SLU process, the percentage of correct semantic units (%csu) and the percentage of correct frame slots (frame name and its attributes) (%cfs), which is equivalent to concept accuracy.

The $\%csu$ measure allows us to evaluate the first phase of our understanding system. This measure is calculated in the same way as the word accuracy used in speech recognition. The $\%cfs$ measure evaluate the overall understanding system and have already been used by other authors [8]. As shown in Section 2, the semantic representation of a sentence is one or more frames. Each frame consists of a name and a sequence of attribute-value pairs. The $\%cfs$ measure is the frame slot accuracy, that is, the number of correctly understood units (frame name and its attribute-value pairs) divided by the number of units in the reference.

Table 1. Results obtained using the different ASR engines.

	Two-level						CRFs					
ASR	Google		Loquendo		HTK		Google		Loquendo		HTK	
SLU model	%csu	%cfs	%csu	%cfs	%csu	%cfs	%csu	%cfs	%csu	%cfs	%csu	%cfs
clean	62.0	72.8	73.8	82.5	80.4	85.6	74.1	77.0	80.3	85.7	84.7	87.8
noisy	76.8	82.6	77.2	85.2	81.3	86.1	83.3	85.2	82.6	87.5	85.3	88.4
clean+noisy	76.8	82.4	77.2	85.3	82.0	86.5	83.9	85.5	82.2	87.4	85.5	89.0

Table 1 shows the results of the understanding process for the test corpus and for the two SLU approaches (Two-level and CRFs). As can be observed, in all the recognizers and in all the understanding systems the results for the $\%cfs$ measure outperforms those of the $\%csu$. That is because in the $\%cfs$ measure irrelevant segments are not considered. These kind of segments are syntactically complex and in the training set there are few samples of each possible realization of them.

The results of SLU systems estimated from noisy data (*noisy* in Table 1) outperform the results obtained by the SLU systems estimated from clean data (*clean* in Table 1) for all the measures defined in both SLU approaches. This differences are more significant for the ASR engines with higher WER; for instance, the *Google noisy* results outperform the *Google clean* results by 9.8 points in the $\%cfs$ metric for Two-level approach and 8.2 points for CRFs approach. The *noisy* results for the ASR with the lowest WER (HTK) only outperform the *clean* results by 0.5 points in the $\%cfs$ metric for Two-level and 0.6 for CRFs approach. As expected, the improvement in SLU is more significant in the ASR engines with lower performance, which scope for improvement is bigger.

Finally, the use of the combined models ($clean+noisy$) returned very similar results to those obtained with *noisy* models. We think that a more sophisticated way of combining *noisy* and *clean* data, for instance using some interpolation techniques, would obtain better results.

6 Conclusions

We have presented an approach for the development of SLU systems by adapting the models to the errors generated in the previous phase of ASR. It is based on the automatic generation of a new segmented and semantically labeled corpus

from the original utterances. Some experiments were performed with the Spanish DIHANA corpus using three different ASR systems and two SLU approaches. The results show that this learning approach can recover and deal with errors generated in the ASR process.

As future work, it would be interesting to study how to better combine models obtained with *clean* training data with models obtained with *noisy* training data.

Acknowledgements This work is partially supported by the Spanish MEC under contract TIN2014-54288-C4-3-R and FPU Grant AP2010-4193

References

1. Benedí, J.M., Lleida, E., Varona, A., Castro, M.J., Galiano, I., Justo, R., López de Letona, I., Miguel, A.: Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. In: LREC 2006. pp. 1636–1639 (2006)
2. Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A., Mostefa, D.: Semantic annotation of the french media dialog corpus. In: Ninth European Conference on Speech Communication and Technology (2005)
3. Calvo, M., García, F., Hurtado, L.F., Jiménez, S., Sanchis, E.: Exploiting multiple hypotheses for multilingual spoken language understanding. CoNLL (2013)
4. De Mori, R., Bechet, F., Hakkani-Tür, D., McTear, M., Riccardi, G., Tür, G.: Spoken language understanding: A survey. IEEE Signal Processing magazine 25(3), 50–58 (2008)
5. Deng, L., Li, X.: Machine learning paradigms for speech recognition: An overview. IEEE Transactions on Audio, Speech, and Language Processing 21(5) (2013)
6. García, F., Hurtado, L., Segarra, E., Sanchis, E., Riccardi, G.: Combining multiple translation systems for Spoken Language Understanding portability. In: Proc. of IEEE Workshop on Spoken Language Technology (SLT). pp. 282–289 (2012)
7. Hahn, S., Lehnen, P., Heigold, G., Ney, H.: Optimizing CRFs for SLU Tasks in Various Languages Using Modified Training Criteria. In: INTERSPEECH (2009)
8. Hahn, S., Lehnen, P., Raymond, C., Ney, H.: A Comparison of Various Methods for Concept Tagging for Spoken Language Understanding. In: LREC (2008)
9. He, X., Deng, L., Tür, D.H., Tür, G.: Multi-style adaptive training for robust cross-lingual spoken language understanding. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (May 2013)
10. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning. pp. 282–289. Citeseer (2001)
11. Raymond, C., Riccardi, G.: Generative and discriminative algorithms for spoken language understanding. Proceedings of Interspeech 2007 pp. 1605–1608 (2007)
12. Riccardi, G., Hakkani-Tür, D.: Active learning: theory and applications to automatic speech recognition. Speech and Audio Processing, IEEE Transactions on 13(4), 504 – 511 (july 2005)
13. Segarra, E., Sanchis, E., Galiano, M., García, F., Hurtado, L.: Extracting Semantic Information Through Automatic Learning Techniques. IJPRAI 16(3) (2002)
14. Tur, G., Deoras, A., Hakkani-Tur, D.: Semantic parsing using word confusion networks with conditional random fields. In: INTERSPEECH (2013)