A methodology for obtaining concept graphs from word graphs

Marcos Calvo, Jon Ander Gómez, Lluís-F. Hurtado, Emilio Sanchis
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera s/n, 46022, València, Spain
{mcalvo,jon,lhurtado,esanchis}@dsic.upv.es

Abstract

In this work, we describe a methodology based on the Stochastic Finite State Transducers paradigm for Spoken Language Understanding (SLU) for obtaining concept graphs from word graphs. In the edges of these concept graphs, both semantic and lexical information are represented. This makes these graphs a very useful representation of the information for SLU. The best path in these concept graphs provides the best sequence of concepts.

1 Introduction

The task of SLU can be seen as the process that, given an utterance, computes a semantic interpretation of the information contained in it. This semantic interpretation will be based on a task-dependent set of concepts.

An area where SLU systems are typically applied is the construction of spoken dialog systems. The goal of the SLU subsystem in the context of a dialog system is to process the information given by the Automatic Speech Recognition (ASR) module, and provide the semantic interpretation of it to the Dialog Manager, which will determine the next action of the dialog. Thus, the work of the SLU module can be split into two subtasks, the first of them is the identification of the sequence of concepts and the segments of the original sentence according to them, and the other is the extraction of the relevant information underlying to these labeled segments. In this work we will focus on concept labeling, but we will also consider the other subtask in our evaluation.

We can distinguish between the SLU systems that work with the 1-best transcription and those that take a representation of the n-best (Hakkani-Tür et al., 2006; Tur et al., 2002). The use of a word graph as the input of the SLU module makes this task more difficult, as the search space becomes larger. On the other hand, the advantage of using them is that there is more information that could help to find the correct semantic interpretation, rather than just taking the best sentence given by the ASR.

In the recent literature, a variety of approaches for automatic SLU have been proposed, like those explained in (Hahn et al., 2010; Raymond and Riccardi, 2007; McCallum et al., 2000; Macherey et al., 2001; Léfèvre, 2007; Lafferty et al., 2001). The methodology that we propose in this paper is based on Stochastic Finite State Transducers (SFST). This is a generative approach that composes several transducers containing acoustic, lexical and semantic knowledge. Our method performs this composition on-the-fly, obtaining as a result a *concept graph*, where semantic information is associated with segments of words. To carry out this step, we use a different language model for each concept and also study the use of lexical categorization and lemmas. The best sequence of concepts can be determined by finding the best path in the concept graph, with the help of a language model of sequences of the concepts.

The rest of this paper is structured as follows. In Section 2, the theoretical model for SLU based on SFST is briefly presented. Then, in Section 3 the methodology for converting word graphs into concept graphs is described. A experimentation to eval-

Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing, pages 75–79,
Donostia–San Sebastián, July 23–25, 2012.

uate this methodology for the SLU task is shown in Section 4. Finally, we draw some conclusions and future work.

2 The SFST approach for SLU

The Bayes classifier for the SLU problem can be expressed as stated in equation 1, where C represents a sequence of concepts or semantic labels and A is the utterance that constitutes the input to the system.

$$\hat{C} = \arg\max_{C} p(C|A) \tag{1}$$

Taking into account the underlying sequence of words W, and assuming that the acoustics may depend on W but not on C, this equation can be rewritten as follows.

$$\hat{C} = \arg\max_{C} \max_{W} p(A|W) \cdot p(W, C) \tag{2}$$

To compute the best sequence of concepts $\hat{C}$ expressed as in equation 2, the proposal made by the paradigm based on SFST is to search the best path in a transducer λ_{SLU} result of composing four SFST:

$$\lambda_{SLU} = \lambda_G \circ \lambda_{gen} \circ \lambda_{W2C} \circ \lambda_{SLM} \tag{3}$$

In this equation λ_G is a SFST provided by the ASR module where the acoustic probabilities $p(A|W)$ are represented, λ_{gen} introduces prior information of the task by means of a lexical categorization, λ_{W2C} provides the probability of a sequence of words and labels it with a semantic label and λ_{SLM} modelizes a language model of sequences of concepts.

3 From word graphs to concept graphs

The output of an ASR can be represented as a word graph. This word graph can be enriched with semantic information, obtaining a *concept graph*. This concept graph constitutes a useful representation of the possible semantics, considering the uncertainty expressed in the original word graph. Finally, finding the best path in the concept graph using a language model of sequences of concepts provides as a result the best sequence of concepts $\hat{C}$, the recognized sentence $\tilde{W}$, and its segmentation according to $\hat{C}$.

3.1 Topology and semantics of the word graph

To perform the transformations for obtaining the concept graph, the input graph given by the ASR should represent the information in the following way. First, its nodes will be labeled with timestamps. Also, for every two nodes i, j such that $i < j - 1$, there will be an edge from i to j labeled with w and weight s if the ASR detected w between the instants i and $j - 1$ with an acoustic score s. Finally, there may exist a λ-transition between any pair of adjacent nodes. The score of this edge should be computed by means of a smoothing method.

Defining the word graph in this way allows us to model on it the distribution $p(A|w)$, where A is the sequence of acoustic frames between the initial and final nodes of any edge, and w the word attached to it. This probability distribution is represented in the theoretical model by λ_G.

3.2 Building the concept graph

The concept graph that is obtained has the following features. First, its set of nodes is the same of the word graph, and its meaning is kept. There is at most one edge between every two nodes i and j $(i < j)$ labeled with the concept c. Every edge is labeled with a pair (W, c), where W is a sequence of words and c the concept that they represent. The weight of the edge is $\max_W(p(A_i^j|W) \cdot p(W|c))$, where A_i^j are the acoustic frames in the interval $[i, j[$ and W the argument that maximizes the former expression.

In this specification appears the probability distribution $p(W|c)$, which can be estimated by using a language model for each available concept.

This concept graph can be built using a Dynamic Programming algorithm that finds for each concept c and each pair of nodes i, j, with $i < j$, the path from i to j on the word graph that maximizes $p(A_i^j|W) \cdot p(W|c)$. In this case, W is the sequence of words obtained by concatenating the words attached to the edges of the path. Each of the "best paths" computed in this way will become an edge in the resulting concept graph.

Thus, in the concept graph it is represented information about possible sequences of words that might have been uttered by the speaker, along with the concepts each of these sequences expresses. This pair is weighted with a score that is the result of combining the acoustic score expressed in the word graph, and the lexical and syntactic score given by the language model, which is dependent on the current concept. Furthermore, this information is enriched with temporal information, since the initial and final nodes of every edge represent the beginning and ending timestamps of the sequence of words. Consequently, this way of building the concept graph corresponds to the transducer λ_{W2C} of equation 3, since we find sequences of words and attach them to a concept. However, we also take advantage of and keep other information, such as the temporal one.

4 Experiments and results

To evaluate this methodology, we have performed SLU experiments using the concept graphs obtained as explained in Section 3 and then finding the best path in each of them. For this experimentation we have used the DIHANA corpus (Benedí et al., 2006). This is a corpus of telephone spontaneous speech in Spanish composed by 900 dialogs acquired by 225 speakers using the Wizard of Oz technique, with a total of 6,229 user turns. All these dialogs simulate real conversations in an automatic train information phone service. The experiments reported here were performed using the user turns of the dialogs, splitting them into a set of 1,340 utterances (turns) for test and all the remaining 4,889 for training. Some interesting statistics about the DIHANA corpus are given in table 1.

Number of words	47,222
Vocabulary size	811
Average number of words per user turn	7.6
Number of concepts	30

Table 1: Characteristics of the DIHANA corpus.

In the DIHANA corpus, the orthographic transcriptions of the utterances are semi-automatically segmented and labeled in terms of semantic units. This segmentation is used by our methodology as a language model of sequences of words for each concept. All the language models involved in this experimentation are bigram models trained using Witten-Bell smoothing and linear interpolation.

In our experimentation, we have considered three different ways for building the λ_{gen} transducer explained in Section 2. The first way consists of considering a transducer that given a word as its input, outputs that word with probability 1. This means that no generalization is being done.

The second λ_{gen} transducer performs a lexical categorization of some of the nouns of the vocabulary. Some extra words have been added to some lexical categories, in order to make the task more realistic, as the lexical coverage is increased. Nevertheless, it also makes the task harder, as the size of the vocabulary increases. We have used a total of 11 lexical categories.

Finally, the third λ_{gen}, transducer we have generated performs the same lexical categorization but it also includes a lemmatization of the verbs. This process is normally needed for real-world systems that work with spontaneous (and maybe telephonic) speech.

We have generated three sets of word graphs to take them as the input for the method. The first of these sets, G_1, is made up by the whole graphs obtained from a word graph builder module that works without using any language model. The *Oracle WER* of these graphs is 4.10. With *Oracle WER* we mean the WER obtained considering the sequence of words $S(G)$ corresponding to the path in the graph G that is the nearest to the reference sentence.

The second set, G_2, is composed by word graphs that only contain the path corresponding to $S(G)$ for each graph $G \in G_1$. These graphs give an idea of the best results we could achieve if we could minimize the confusion due to misrecognized words.

The third set, G_3 is formed by a synthetic word graph for each reference sentence, in which only that sentence is contained. This set of graphs allows us to simulate an experimentation on plain text.

For our evaluation, we have taken two measures. First, we have evaluated the Concept Error Rate (CER) over the best sequence of concepts. The definition of the CER is analogous to that of the WER but taking concepts instead of words. Second, we have also evaluated the slot-level error (SLE). The SLE is similar to the CER but deleting the non-relevant segments (such as courtesies) and substituting the relevant concepts by a canonic value for the sequence of words associated to them.

Tables 2, 3, and 4 show the results obtained using the different λ_{gen} transducers explained before.

Input word graphs	CER	SLE
G_1	31.794	35.392
G_2	11.230	9.104
G_3	9.933	5.321

Table 2: CER and SLE without any categorization.

Input word graphs	CER	SLE
G_1	34.565	38.760
G_2	11.755	8.714
G_3	9.633	4.516

Table 3: CER and SLE with lexical categorization.

From the results of Tables 2, 3, and 4 several facts come to light. First, we can see that, in all the experiments performed with the G_1 set, the CER is lower than the SLE, while with the other sets the CER is larger than the SLE. It is due to the fact that the whole graphs obtained from the word graph builder have more lexical confusion than those from G_2 and G_3, which are based on the reference sentence. This lexical confusion may cause that a well-recognized concept is associated to a misrecognized sequence of words. This would imply that a hit would be considered for the CER calculation, while the value for this slot is missed.

Other interesting fact is that, for the G_1 set, the more complex λ_{gen} transducers give the worse results. This is because in these graphs there is a significant confusion between phonetically similar words, as the graphs were generated without any language model. This phonetic confusion, combined with the generalizations expressed by the lexical categorization and the lemmas, makes the task harder, which leads to worse results. Nevertheless, in a real-world application of this system these generalizations would be needed in order to have a larger coverage of the lexicon of the language. The experiments on G_2 and G_3 show that when the confusion introduced in the graphs due to misrecognized words is minimized, the use of lexical categorization and lemmatization helps to improve the results.

5 Conclusions and future work

In this paper we have described a methodology, based on the SFST paradigm for SLU, for obtaining concept graphs from word graphs. The edges of the concept graphs represent information about possible sequences of words that might have been uttered by the speaker, along with the concept each of these sequences expresses. Each of these edges is weighted with a score that combines acoustic, lexical, syntactic and semantic information. Furthermore, this information is enriched with temporal information, as the nodes represent the beginning and ending of the sequence of words. These concepts graphs constitute a very useful representation of the information for SLU.

Input word graphs	CER	SLE
G_1	36.536	40.640
G_2	11.605	8.445
G_3	9.458	4.064

Table 4: CER and SLE with lemmatization and lexical categorization.

To evaluate this methodology we have performed an experimental evaluation in which different types of lexical generalization have been considered. The results show that a trade-off between the lexical confusion expressed in the word graphs and the generalizations encoded in the other transducers should be achieved, in order to obtain the best results.

It would be interesting to apply this methodology to word graphs generated with a language model, although this way of generating the graphs would not fit exactly the theoretical model. If a language model is used to generate the graphs, then their lexical confusion could be reduced, so better results could be achieved. Other interesting task in which this methodology could help is in performing SLU experiments on a combination of the output of some different ASR engines. All these interesting applications constitute a line of our future work.

Acknowledgments

This work has been supported by the Spanish Ministerio de Economía y Competitividad, under the project TIN2011-28169-C05-01, by the Vicerrectorat d'Investigació, Desenvolupament i Innovació de la Universitat Politècnica de València, under the project PAID-06-10, and by the Spanish Ministerio de Educación under FPU Grant AP2010-4193.

References

José-Miguel Benedí, Eduardo Lleida, Amparo Varona, María-José Castro, Isabel Galiano, Raquel Justo, Iñigo López de Letona, and Antonio Miguel. 2006. Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. In *Proceedings of LREC 2006*, pages 1636–1639, Genoa (Italy).

S. Hahn, M. Dinarelli, C. Raymond, F. Léfèvre, P. Lehnen, R. De Mori, A. Moschitti, H. Ney, and G. Riccardi. 2010. Comparing stochastic approaches to spoken language understanding in multiple languages. *Audio, Speech, and Language Processing, IEEE Transactions on*, 6(99):1569–1583.

D. Hakkani-Tür, F. Béchet, G. Riccardi, and G. Tur. 2006. Beyond ASR 1-best: Using word confusion networks in spoken language understanding. *Computer Speech & Language*, 20(4):495–514.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, pages 282–289. Citeseer.

F. Léfèvre. 2007. Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages 13–16. IEEE.

K. Macherey, F.J. Och, and H. Ney. 2001. Natural language understanding using statistical machine translation. In *European Conf. on Speech Communication and Technology*, pages 2205–2208. Citeseer.

A. McCallum, D. Freitag, and F. Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 591–598. Citeseer.

C. Raymond and G. Riccardi. 2007. Generative and discriminative algorithms for spoken language understanding. *Proceedings of Interspeech2007, Antwerp, Belgium*, pages 1605–1608.

G. Tur, J. Wright, A. Gorin, G. Riccardi, and D. Hakkani-Tür. 2002. Improving spoken language understanding using word confusion networks. In *Proceedings of the ICSLP*. Citeseer.