Concept Segmentation and Labeling for Conversational Speech

Marco Dinarelli, Alessandro Moschitti, Giuseppe Riccardi

Department of Computer Science and Information Engineering (DISI), University of Trento, Italy
{dinarelli,moschitti,riccardi}@disi.unitn.it

Abstract

Spoken Language Understanding performs automatic concept labeling and segmentation of speech utterances. For this task, many approaches have been proposed based on both generative and discriminative models. While all these methods have shown remarkable accuracy on manual transcription of spoken utterances, robustness to noisy automatic transcription is still an open issue. In this paper we study algorithms for Spoken Language Understanding combining complementary learning models: Stochastic Finite State Transducers produce a list of hypotheses, which are re-ranked using a discriminative algorithm based on kernel methods. Our experiments on two different spoken dialog corpora, MEDIA and LUNA, show that the combined generative-discriminative model reaches the state-of-the-art such as Conditional Random Fields (CRF) on manual transcriptions, and it is robust to noisy automatic transcriptions, outperforming, in some cases, the state-of-the-art.

Index Terms: Spoken Language Understanding, Discriminative Learning, Kernel Methods

1. Introduction

In Spoken Dialog Systems, the Language Understanding module performs the task of translating a spoken sentence into its meaning representation based on semantic constituents. These are the units for meaning representation, called also concepts. Concepts are instantiated by sequences of words and the Spoken Language Understanding (SLU) module finds the association between words and concepts using machine learning algorithms.

In the last decade two major approaches have been proposed to find this correlation: (i) generative models, whose parameters refer to the joint probability of concepts and constituents; and (ii) discriminative models, which learn a classification function based on conditional probabilities of concepts given words.

A simple but effective generative model is the one based on Stochastic Finite State Transducers (SFST) [1]. It performs SLU as a translation process from words to concepts using FST.

An example of discriminative model used for SLU is the one based on Support Vector Machines (SVMs) [2], as shown in [1]. In this approach, data are mapped into a vector space and SLU is performed as a classification problem using Maximal Margin Classifiers [3].

A relatively more recent approach to SLU is based on Conditional Random Fields (CRF) [4]. CRFs are undirected-graph models conditionally trained and they belong to the class of discriminative models. CRFs take into account many non-independent features of the input to predict the best concept sequence, like discriminative models. Since they are conditionally trained, they don't need to train explicitly features dependencies, like generative models would do.

Generative models have the advantage to be more robust to overfitting while discriminative models are more robust to irrelevant features. Both approaches are particularly suitable for the SLU task [1]. Although these models have proven to be very effective on manual transcriptions of speech sentences, their robustness on noisy input, like automatic transcription of a recognizer, is an open issue. Therefore studies on effective approaches for automatic speech recognition and understanding is an interesting research field.

Discriminative and generative models have very different characteristics and ways of encoding prior knowledge; we believe that models taking into account characteristics of both approaches are particularly promising to improve the robustness on noisy automatic transcription.

In this paper, we propose a method for SLU based on the joint use of generative and discriminative models: FSTs are used to generate a list of SLU hypotheses, which are re-ranked using SVMs and kernel methods. In order to capture arbitrary long distance dependencies between words and concepts, we adopted kernel methods for structured data, in particular Tree Kernels (TK) [5].

We experimented with our approach on two different corpora: the French MEDIA corpus [6] and a new corpus acquired in the European project LUNA[1] [7].

The results show that our approach can improve the state-of-the-art on both manual and automatic transcriptions of spoken sentences, showing a very good robustness to noisy input. Additionally, our method is easily improvable new effective structural features designing.

The rest of the paper is organized as follows: in Section 2, we describe how SLU hypotheses are generated and used to train a discriminative re-ranker. In Section 3, we describe the corpora and the experiments used to evaluate our model. In Section 4, we show the results of the evaluation of our approach compared with state-of-the-art models and finally, in Section 5, we provide conclusion and future work.

2. Re-ranking SLU Hypotheses with SVMs and Kernel Methods

2.1. Generative Model: Stochastic Conceptual Language Model (SCLM)

The first step of our approach is to produce a list of SLU hypotheses using a Stochastic Conceptual Language Model. This model is the same described in [1] with the only difference that we train the language model using the SRILM toolkit [8] and we then convert it into a Stochastic Finite State Transducer (SFST). This allows us to use a wide group of language models, back-off or interpolated with many kind of smoothing techniques [9].

Given the input sentence:

[1]Contract n. 33549

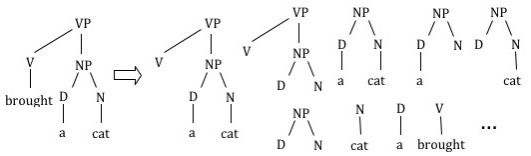

Figure 1: Syntactic Tree fragments

Corpus	Train set		Test set	
LUNA	**words**	**concepts**	**words**	**concepts**
Dialogs	183		67	
Turns	1.019		373	
Tokens	8.512	2.887	2.888	984
Vocab.	1.172	34	-	-
OOV rate	-	-	3.2%	0.1%

Table 1: Statistics on the LUNA WOZ corpus

ho un problema col monitor
(I have a problem with my screen)
a possible semantic annotation is:
null{*ho*} ***PROBLEM***{*un problema*} ***HARDWARE***{*col monitor*}, where ***PROBLEM*** and ***HARDWARE*** are two domain concepts and ***null*** is the concept used for words not meaningful for the task. In order to have a one-to-one association between words and concepts, the concepts are segmented using begin (*B*) and inside (*I*) concept markers:
null{*ho*} ***PROBLEM-B***{*un*} ***PROBLEM-I***{*problema*} ***HARDWARE-B***{*col*} ***HARDWARE-I***{*monitor*}

This annotation is performed by the model using the combination of three transducers:

$$\lambda_{SLU} = \lambda_W \circ \lambda_{W2C} \circ \lambda_{SLM},$$

where λ_W is the transducer representation of the input sentence, λ_{W2C} is the transducer mapping words to concepts and λ_{SLM} is the Stochastic Conceptual Language Model trained with SRILM toolkit and converted in FST. The SCLM represents joint probability of word and concept sequences:

$$P(W, C) = \prod_{i=1}^{k} P(w_i, c_i | h_i),$$

where $W = w_1..w_k$, $C = c_1..c_k$ and $h_i = w_{i-1}c_{i-1}..w_1c_1$.

2.2. Discriminative Re-ranking

Our discriminative re-ranking is essentially an SVM (i.e. a classifier) [2] trained with pairs of conceptually annotated sentences produced by the FST described in previous section. SVM learns to select which annotation has an error rate lower than the others so that the *m*-best annotations can be sorted based on their correctness. In this section we focus on the kernels used to implement our re-ranking model.

2.3. Tree kernels

Tree kernels represent trees in terms of their sub-structures (fragments). The kernel function detects if a tree subpart (common to both trees) belongs to the feature space. For such purpose, the desired fragments need to be described. We consider an important characterization: the syntactic tree fragments (STF).

An STF is a general subtree whose leaves can be non-terminal symbols. For example, Figure 1 shows 9 STFs (out of 17) of the subtree rooted in `VP` (of the left tree). The STFs satisfy the constraint that grammatical rules cannot be broken. For example, `[VP [V NP]]` is an STF, which has two non-terminal symbols, `V` and `NP`, as leaves whereas `[VP [V]]` is not an STF.

2.4. Counting Shared SubTrees

The main idea of tree kernels is to compute the number of common substructures between two trees T_1 and T_2 without explicitly considering the whole fragment space. To evaluate the above kernels between two trees T_1 and T_2, we need to define a set $\mathcal{F} = \{f_1, f_2, \ldots, f_{|\mathcal{F}|}\}$, *i.e.* a tree fragment space and an indicator function $I_i(n)$, equal to 1 if the target f_i is rooted at node n and equal to 0 otherwise. A tree-kernel function over T_1 and T_2 is $TK(T_1, T_2) = \sum_{n_1 \in N_{T_1}} \sum_{n_2 \in N_{T_2}} \Delta(n_1, n_2)$, where N_{T_1} and N_{T_2} are the sets of the T_1's and T_2's nodes, respectively, and $\Delta(n_1, n_2) = \sum_{i=1}^{|\mathcal{F}|} I_i(n_1) I_i(n_2)$. The latter is equal to the number of common fragments rooted in the n_1 and n_2 nodes.

In the following sections we report the equation for the efficient evaluation of Δ for ST kernel.

2.5. Syntactic Tree Kernels (STK)

The Δ function depends on the type of fragments that we consider as *basic* features. For example, to evaluate the fragments of type STF, it is defined as:

1. if the productions at n_1 and n_2 are different then $\Delta(n_1, n_2) = 0$;
2. if the productions at n_1 and n_2 are the same, and n_1 and n_2 have only leaf children (*i.e.* they are pre-terminals symbols) then $\Delta(n_1, n_2) = 1$;
3. if the productions at n_1 and n_2 are the same, and n_1 and n_2 are not pre-terminals then

$$\Delta(n_1, n_2) = \prod_{j=1}^{nc(n_1)} (\sigma + \Delta(c_{n_1}^j, c_{n_2}^j)) \qquad (1)$$

where $\sigma \in \{0, 1\}$, $nc(n_1)$ is the number of children of n_1 and c_n^j is the j-th child of the node n. Note that, since the productions are the same, $nc(n_1) = nc(n_2)$. $\Delta(n_1, n_2)$ evaluates the number of STFs common to n_1 and n_2 as proved in [5]. Moreover, a decay factor λ can be added by modifying steps (2) and (3) as follows[2]:

2. $\Delta(n_1, n_2) = \lambda$,
3. $\Delta(n_1, n_2) = \lambda \prod_{j=1}^{nc(n_1)} (\sigma + \Delta(c_{n_1}^j, c_{n_2}^j))$.

The computational complexity of Eq. 1 is $O(|N_{T_1}| \times |N_{T_2}|)$ but as shown in [10], the average running time tends to be linear, *i.e.* $O(|N_{T_1}| + |N_{T_2}|)$, for natural language syntactic trees.

2.6. Re-ranking Models

The FST model generates the *m* most likely concept annotations. These are used to build annotation pairs, $\langle s^i, s^j \rangle$, which are positive instances if and only if s^i has less concept annotation errors than s^j, with respect to the manual annotation of the

[2] To have a similarity score between 0 and 1, we also apply the normalization in the kernel space, *i.e.*:
$K'(T_1, T_2) = \frac{TK(T_1, T_2)}{\sqrt{TK(T_1, T_1) \times TK(T_2, T_2)}}$.

Corpus Media	Train set		Test set	
	words	concepts	words	concepts
Turns	12,922		3,518	
# of tokens	94,912	43,078	26,676	12,022
Vocabulary	5,307	80	-	-
OOV rate	-	-	0.01%	0.0%

Table 2: Statistics on the MEDIA corpus

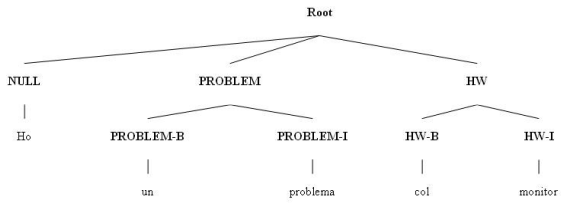

Figure 2: The *Semantic Tree* used for STK or PTK

corpus. Thus, a trained binary classifier can decide if s^i is more accurate than s^j. Each candidate annotation s^i is described by a word sequence with its concept annotation. Considering the example of Section 2.1 (*ho un problema col monitor*), a pair of annotations $\langle s^i, s^j \rangle$ could be

s^i: ***NULL*** *ho* ***PROBLEM-B*** *un* ***PROBLEM-I*** *problema* ***HARDWARE-B*** *col* ***HARDWARE-I*** *monitor*

s^j: ***NULL*** *ho* ***ACTION-B*** *un* ***ACTION-I*** *problema* ***HARDWARE-B*** *col* ***HARDWARE-B*** *monitor*

The second annotation is less accurate than the first since *"problema"* (problem) is annotated as ***ACTION*** and *"col monitor"* (with the screen) is split in two different concepts.

Given pairs of annotated sentences, let e_k be the pair $\langle s_k^1, s_k^2 \rangle$, our re-ranking model is based on the following kernel:

$$K_R(e_1, e_2) = K(s_1^1, s_2^1) + K(s_1^2, s_2^2) - K(s_1^1, s_2^2) - K(s_1^2, s_2^1) \quad (2)$$

where *K* in this case is STK.

This schema, consisting in summing four different kernels, has been already applied in [5] for syntactic parsing re-ranking, where the basic kernel was a tree kernel, and in [11], where, to re-rank Semantic Role Labeling annotations, a tree kernel was used on a semantic tree.

In order to use the kernels described above on our annotated sentences, a suitable representation must be designed. For this reason, from each annotated sentence, we create a *Semantic-Tree* like the one in Fig. 2. The root of the tree is an arbitrary symbol whereas the first level of the tree contains the concepts of the sentence. Each node of the first level has children representing the concept chunked using markers *B* and *I* as described in Section 2.1. Finally, the leaves of each subtree are the words instantiating the corresponding concept.

Note that the tree in Fig. 2 is different from trees in Fig. 1. The first is a *Semantic-Tree*, while the others are syntactic trees. Note that the semantic annotation we used to build our trees is made upon syntactic chunking, so the semantic annotation implicitly takes into account the syntactic structure of the sentence.

Model	MEDIA (CER)		LUNA (CER)	
	Attr.	Attr.-Value	Attr.	Attr.-Value
FST	13.7%	17.9%	23.2%	27.3%
CRF	11.5%	15.7%	20.4%	24.6%
SVM-RR	12.1%	16.4%	18.4%	22.5%

Table 3: Results of SLU experiments on MEDIA and LUNA test set manual transcriptions

3. Experimental Setup

3.1. Corpora

We used two different speech corpora:

The Luna corpus is the first conversational spoken dialog corpora including problem-solving interactions. Such conversations are recorded from the software/hardware help-desk call-center of one industry partner of the LUNA consortium [12]. The data are organized in transcriptions and annotations of speech based on a new multi-level protocol. Data acquisition is still in progress. Currently, 250 dialogs acquired with a WOZ approach and 180 Human-Human (HH) dialogs are available. In this work we used only WOZ dialogs. Statistics on LUNA corpus are reported in Table 1.

The corpus MEDIA was collected within the French project MEDIA-EVALDA [6] for development and evaluation of spoken understanding models and linguistic studies. The corpus is composed of 1,257 dialogs, from 250 different speakers, acquired with a Wizard of Oz (WOZ) approach in the context of hotel room reservations and tourist information. Statistics on transcribed and conceptually annotated data are reported in Table 2.

3.2. Experiments

Given the small size of LUNA WOZ corpus, we did not carried out parameter optimization on a development set but we used default or a priori parameters. We experimented with LUNA WOZ and our re-ranker obtained by using SVMs with STK on our semantic structures described in Section 2.

We trained all the SCLMs used in our experiments with the SRILM toolkit [8] and we used an interpolated model for probability estimation with Kneser-Ney discount [9]. We then converted the model in an FST using SRILM toolkit.

To train the re-ranker, we used the SVM-Light-TK toolkit (available at dit.unitn.it/moschitti), which includes tree kernels in SVM-Light [13].

We compared our results with a CRF model trained using CRF++, a free tool available at http://crfpp.sourceforge.net/. In particular our model was trained using features of the input sentence like Word Categories and morpho-syntactic features as in [14].

We ran SLU experiments on manual and automatic transcriptions. The latter are produced by a speech recognizer with a WER of 41.0% and 31.4% on the LUNA and the MEDIA test sets, respectively.

4. Results

All the results of our experiments are reported in tables 3 and 4 in terms of Concept Error Rate (CER). CER is a measure based on the Levensthein alignment of sentences and it is computed as the ratio between inserted, deleted and confused concepts and the number of concepts in the reference sentence.

Model	MEDIA (CER)		LUNA (CER)	
	Attr.	Attr.-Value	Attr.	Attr.-Value
FST	28.6%	32.7%	42.7%	46.9%
CRF	24.0%	28.9%	41.8%	45.7%
SVM-RR	25.0%	29.7%	38.9%	43.4%

Table 4: Results of SLU experiments on MEDIA and LUNA test set automatic transcriptions (WER 31.4% for MEDIA, 41% for LUNA)

Table 3 shows the results of SLU experiments on the MEDIA and LUNA test sets using manual transcription of spoken sentences. We note that the baseline models (FST and CRF) using a small corpus like the one of LUNA show a larger error rate. Thus the re-ranking approach can learn and correct many mistakes, significantly outperforming CRF on both attribute and attribute-value annotation (2% points and 2.1% points respectively). The FSTs baseline of 23.2% is largely improved by the re-ranking model of 4.9% points (21.1% relative improvement) on attribute annotation.

In contrast, on a big corpus like MEDIA, the baseline models can be accurately learned thus less errors can be corrected. As a consequence, our re-ranking approach is a little less accurate than the CRF model (0.6 and 0.7 percent points on attribute and attribute-values, respectively), but it still improves the FSTs baseline of 1.6% points (11.7% relative improvement).

The same behavior is reproduced for SLU experiments on automatic transcriptions, shown in Table 4.

We note that on the LUNA corpus CRFs are more accurate than FSTs (0.9% points on attributes and 1.2% points on attribute-values), but they are significantly improved by the re-ranking model (2.9% points on attributes and 2.3% points on attribute-values), which yields an improvement of 3.8% points on the FSTs baseline for attribute annotation.

On the MEDIA corpus, the re-ranking model is again very accurate improving the FSTs baseline of 3.6% points (12.6% relative improvement) on attribute annotation, but the most accurate model is again CRF (1% points better than the re-ranking model on attribute annotation).

The different behavior of the re-ranking model between the LUNA and MEDIA corpora is due partially to the task complexity (34 concepts in LUNA, 80 in MEDIA), but more to the fact that CRFs have been studied and experimented more in-dept (see [14]) than the re-ranking approach for this task. This allowed the CRF parameters and features to be greatly optimized. We believe that the re-ranking model can be relevantly improved by carrying out parameter optimization and new structural feature design.

It should be noted that some structural features for manual transcription have been studied in [15]. However, in such work no study on the robustness to the noisy ASR output has been carried out. In this paper, we showed that our structural feature (i.e. the semantic tree) is very effective for this output and that re-ranking models based on SVMs and Tree Kernels are more robust than CRF on small and novel corpora.

5. Conclusions

In this paper we propose an approach to Spoken Language Understanding based on the joint use of a generative and a discriminative model. This approach reaches state-of-the-art, *i.e.* CRF, outperforming it on new corpora. The re-ranking model seems to be particularly robust on automatic transcriptions since it is more accurate than CRF on the LUNA corpus and it improves significantly the FSTs baseline on the MEDIA corpus (12.6% relative improvement). All the results obtained for this work, and in particular the results on the MEDIA corpus, are particularly meaningful since the re-ranking model can be improved along different research lines, e.g. feature design, parameter tuning. For future work we plan to carry out:

- Re-ranking of hypotheses generated with different models, similar to system combinations made with ROVER in [14]
- To use more hypotheses for training/classification (we only used 10 hypotheses for this work).
- To use different kernels and their combinations for training the re-ranker.

6. References

[1] C. Raymond and G. Riccardi, "Generative and discriminative algorithms for spoken language understanding," in *Proceedings of Interspeech2007*, Antwerp,Belgium, 2007.

[2] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[3] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[4] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of ICML2001*, US, 2001.

[5] M. Collins and N. Duffy, "New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete structures, and the voted perceptron," in *ACL02*, 2002, pp. 263–270.

[6] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa, "Semantic annotation of the french media dialog corpus," in *Proceedings of Interspeech2005*, Lisbon, Portugal, 2005.

[7] C. Raymond, G. Riccardi, K. J. Rodrigez, and J. Wisniewska, "The luna corpus: an annotation scheme for a multi-domain multilingual dialogue corpus," in *Proceedings of Decalog2007*, Trento, Italy, 2007.

[8] A. Stolcke, "Srilm: an extensible language modeling toolkit," in *Proceedings of SLP2002*, Denver, USA, 2002.

[9] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Technical Report of Computer Science Group*, Harvard, USA, 1998.

[10] A. Moschitti, "Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees," in *Proceedings of ECML 2006*, Berlin, Germany, 2006, pp. 318–329.

[11] A. Moschitti, D. Pighin, and R. Basili, "Tree kernels for semantic role labeling," *Computational Linguistics*, vol. 34, no. 2, pp. 193–224, 2008. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.2.193

[12] M. Dinarelli, S. Quarteroni, S. Tonelli, A. Moschitti, and G. Riccardi, "Annotating spoken dialogs: From speech segments to dialog acts and frame semantics," in *Proceedings of SRSL 2009, the 2nd Workshop on Semantic Representation of Spoken Language*, Athens, Greece, March 2009.

[13] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds., 1999, pp. 169–184.

[14] S. Hahn, P. Lehnen, C. Raymond, and H. Ney, "A comparison of various methods for concept tagging for spoken language understanding," in *International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May 2008.

[15] M. Dinarelli, A. Moschitti, and G. Riccardi, "Re-ranking models for spoken language understanding," in *Proceedings of EACL 2009*, Athens, Greece, March 2009.