HYPOTHESES SELECTION FOR RE-RANKING SEMANTIC ANNOTATIONS

Marco Dinarelli[1]*, Alessandro Moschitti*[2]*, Giuseppe Riccardi*[2]

[1] LIMSI-CNRS
[2]Department of Computer Science and Information Engineering
University of Trento
marcod@limsi.fr {moschitti,riccardi}@disi.unitn.it

ABSTRACT

Discriminative reranking has been successfully used for several tasks of Natural Language Processing (NLP). Recently it has been applied also to Spoken Language Understanding, imrpoving state-of-the-art for some applications. However, such proposed models can be further improved by considering: (i) a better selection of the initial n-best hypotheses to be re-ranked and (ii) the use of a strategy that decides when the reranking model should be used, i.e. in some cases only the basic approach should be applied.

In this paper, we apply a semantic inconsistency metric to select the n-best hypotheses from a large set generated by an SLU basic system. Then we apply a state-of-the-art re-ranker based on the Partial Tree Kernel (PTK), which encodes SLU hypotheses in Support Vector Machines (SVM) with complex structured features. Finally, we apply a decision model based on confidence values to select between the first hypothesis provided by the basic SLU model and the first hypothesis provided by the re-ranker.

We show the effectiveness of our approach presenting comparative results obtained by reranking hypotheses generated by two very different models: a simple Stochastic Language Model encoded in Finite State Machines (FSM) and a Conditional Random Field (CRF) model. We evaluate our approach on the French MEDIA corpus and on an Italian corpus acquired in the European Project LUNA. The results show a significant improvement with respect to the current state-of-the-art and previous re-ranking models.

Index Terms: Spoken Language Understanding, Discriminative Reranking, Kernel Methods.

1. INTRODUCTION

Discriminative reranking is a widely used approach for several NLP tasks: Syntactic Parsing [1], Named Entity Recognition [1, 2], Semantic Role Labelling [3], Machine Translation [4], Question Answering [5]. Recently reranking has also been successfully applied †to SLU [6].

Discriminative Reranking is a combination of two different models: a first SLU model is used to generate a ranked list of n-best hypotheses. Then, a reranking model sorts the list based on a different score and the final result is the new top ranked hypothesis.

In previous approaches complex features are extracted from the hypotheses to learn the reranking model, but no model has been applied to search in the hypothesis space generated by the baseline SLU model, i.e. the raw n-best list is simply used. Moreover, to keep low the overall computational cost, the size of n is typically small (few tens). This is a limitation since the larger is the hypotheses space generated, the more likely is to find a better hypothesis. Re-ranking a large set of hypotheses is computationally expensive, thus a strategy to select the best hypotheses to be re-ranked would partially overcome this problem.

Another aspect of reranking that deserves to be deeper studied is its applicability, i.e. a strategy to decide when it should be applied or when the first hypothesis of the basic model can result in a higher accuracy. In other words, although reranking generally improves the baseline model, sometimes this assumption is wrong. Thus finding a strategy to detect this situation can improve the final accuracy.

In this paper, we propose two new models for improving discriminative reranking: (a) we studied a semantic inconsistency metric that can be applied to SLU hypotheses to select those that are more likely to be correct; (b) we apply a model selection based on the confidence scores provided by the baseline SLU model and the reranker. This decides if the original top ranked hypothesis is more accurate than the reranked best hypothesis.

Our re-ranking startegies result to be effective using two baseline models with very different characteristics on different aspects: a FSM-based Stochastic Language Model and a CRF Model. The first is a generative model encoding only few features, the second is a discriminative model learning global probabilities and using many features. We evaluate our approach on two corpora: MEDIA [7] and the Italian corpus acquired in the European Project LUNA [8]. The results show that our approach significantly improves both "traditional" reranking approaches and state-of-the-art SLU models.

The remainder of the paper is organized as follows: in Section 2 we introduce the SLU task. Section 3 describes discriminative reranking for SLU. Section 4 describes the improved strategies for SLU reranking whereas the experiments that evaluate our approaches are described in Section 5. Finally in Section 6 we draw some conclusions.

2. SPOKEN LANGUAGE UNDERSTANDING (SLU)

SLU aims at extracting a meaning representation from natural language sentences. Designing a general meaning representation which can capture the semantics of a spoken language is complex. Therefore, in practice, the meaning representations depend on the task domain modeled in each application. For the corpora used in this work, the semantic representation is defined in an ontology (see [7] for the French MEDIA ontology and [9] for the Italian corpus ontology). Given as input the following natural language sentence:

"*Good morning I have a problem with my printer*"

SLU performs the semantic representation extraction in two steps:

1. **Automatic Concept Labeling**

Null{*Good morning I have*} **Problem**{*a problem*} **Peripheral**{*with my printer*}

2. **Attribute-Value Extraction**
   Problem[general_problem] **Peripheral**[printer]

Problem and **Peripheral** are two domain concepts defined in the ontology and **Null** is the concept for words not bringing any semantic content with respect to the application domain, thus, as shown in the example above, it is removed from the final result. **generic_problem** and **printer** are two normalized values, defined also in the application ontology. Concepts are called also "attributes" and the representation used for SLU is usually called attribute-value representation.

Several models have been proposed for the Automatic Concept Labeling step: Stochastic Finite State Transducers (SFST), Conditional Random Fields (CRF), Support Vector Machines (SVM), Maximum Entropy (EM), Statistical Machine Translation (SMT). In [10], it is provided a comparison of all these models and a combination of them using ROVER [11] on the MEDIA corpus. SLU models are learned from manually annotated data.

The second SLU step is performed with two approaches: a) Rule-based approaches apply Regular Expressions (RE) to map the words realizing a concept into a normalized value. Regular expressions are defined for each attribute-value pair. Given a concept and its realizing surface, if a RE for that concept matches the surface, the corresponding value is returned. An example of surfaces that can be mapped into the value "printer" given the concept "Peripheral" is:

1. *printer*
2. *the printer*
3. *with my printer*
   ...

Note that these surfaces share the same keyword for the given concept, i.e. "*printer*".

b) Probabilistic approaches learn from data the conditional probability of values V, given the concept C and the corresponding sequence of words W: $P(V|W, C)$.

3. DISCRIMINATIVE RERANKING FOR SLU

Discriminative reranking has been introduced in [1]. It has been successfully applied to many NLP tasks like Named Entity Recognition [2], Syntactic Parsing [1], Semantic Role Labeling [3], Question Answering [5], Machine Translation [4] and, more recently, also to Spoken Language Understanding [6].

The first step in the reranking approach is to generate the hypotheses using a baseline model. For this purpose, in this work we use SFST and CRF models, both described in [10]. The SFST model encodes joint probabilities of a trigram conceptual language model:

$$P(W_1^N, C_1^N) = \prod_{i=1}^{N} P(w_i, c_i | w_{i-1}, c_{i-1}, w_{i-2}, c_{i-1}), \quad (1)$$

where W_1^N and C_1^N are words and concept sequences. Conditional Random Fields learn discriminatively global posterior probabilities:

$$p(C_1^N | W_1^N) = \frac{1}{Z} \prod_{n=1}^{N} \exp\left(\sum_{m=1}^{M} \lambda_m \cdot h_m(c_{n-1}, c_n, w_{n-2}^{n+2}) \right) \quad (2)$$

where λ_m are the training parameters. $h_m(c_{n-1}, c_n, w_{n-2}^{n+2})$ are the feature functions capturing conditional dependencies of concepts and words. Z is a probability normalization factor in order to model well defined probability distribution:

$$Z = \sum_{\tilde{c}_1^N} \prod_{n=1}^{N} H(\tilde{c}_{n-1}, \tilde{c}_n, w_{n-2}^{n+2}) \quad (3)$$

where $\tilde{c}_{n-1}$ and $\tilde{c}_n$ are the concepts hypothesized for the previous and current words.

Hypotheses generated by the baseline model are used to train the reranking model. Our reranking framework is the same described in [6], based on SVM and Partial Tree Kernel (PTK) [12]. Since PTK works on trees, hypotheses must be converted in a tree-like structure. Using the same hypothesis example of previous section, the semantic tree structure shown in Figure 1 is produced. Note that it is immediate to find the hypothesis corresponding to a given tree, thus from now on we will use "tree" and "hypothesis" interchangeably.

Trees are used to build pairs $e_k = \langle t_k^1, t_k^2 \rangle$, which are training and classification instances. Positive training instances are pairs where the first element is the best hypothesis in the *n*-best list. The best hypothesis is found measuring the edit distance with respect to the manual annotation. Negative training instances are built simply inverting positive ones. This means that given a list of *n* hypotheses, $2 \cdot n$ instances are generated from each input sentence, *n* comparing the best hypothesis with the others, i.e. positive instances, and *n* inverting positive instances. This approach for building pairs allows the reranker to learn to give higher scores to correct hypotheses. For classification instances, since hypotheses cannot be compared with the reference annotation, in principle all possible pairs of *n* hypotheses must be generated. Nevertheless, using the simplification described in [14], only n instances are generated, allowing a relatively fast classification phase.

Training and classification are performed using the following reranking kernel in SVM [2]:

$$\begin{aligned} K_R(e_1, e_2) &= PTK(t_1^1, t_2^1) + PTK(t_1^2, t_2^2) \\ &- PTK(t_1^1, t_2^2) - PTK(t_1^2, t_2^1), \end{aligned} \quad (4)$$

where e_1 and e_2 are two pairs of trees to be compared (see [6] for details).

It is important to note that the reranking kernel in equation 4, consisting in summing four different kernels, has been successfully proposed in [15, 14] for syntactic parsing reranking, where the basic kernel was a Tree Kernel. The same reranking schema has been used in [4] for reranking different candidate hypotheses for machine translation.

The hypotheses generated by the baseline model are ranked by the score computed by the SVM model, according to the kernel in Eq. 4.

4. IMPROVED RERANKING STRATEGIES

4.1. Hypotheses Selection Criteria (HSC)

An interesting strategy to improve reranking performance is the selection of the best set of hypotheses to be reranked. In previous work [1, 4, 6], no study in this direction has been carried out, i.e. the *n*-best hypotheses generated by the basic model were simply used for rernaking.

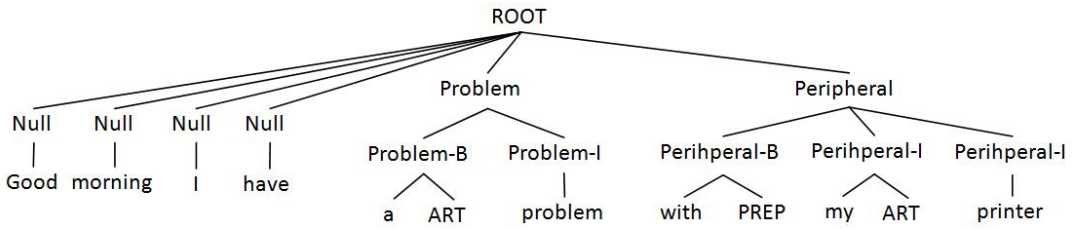

Fig. 1. An example of "*FEATURES*" semantic tree used in PTK

In this work we propose a semantic inconsistency metric (SIM) based on the attribute-value extraction phase that allows to select better n-best hypotheses.

The attribute-value extraction module is based on rules that map words (or word sequences) into the corresponding value. For this purpose, the conceptual information associated with words (annotated during the automatic concept labeling step) is also used.

The rules are defined to extract values from well formed phrases annotated with correct concepts. Thus, when the corresponding words are annotated with a wrong concept, the extracted value will probably result wrong. We use this property to compute a semantic inconsistency value for hypotheses, which in turn allows us to select better, i.e. with high probability to be correct, hypotheses.

We show our SIM using the same example of Section 2. From it, three possible hypotheses may be generated by the baseline model (we suppose to have already removed **Null** concepts):

1. **Action**{*a problem*} **Peripheral**{*with my printer*}
2. **Problem**{*a problem*} **Peripheral**{*with my printer*}
3. **Problem**{*a problem*} **Peripheral**{*with my*} **Peripheral**{*printer*}

Two of these annotations show typical errors of an SLU model:
(i) wrong concepts annotation: in the first hypothesis the phrase "a problem" is erroneously annotated as **Action**;
(ii) wrong concept segmentation: in the third hypothesis the phrase "with my printer" is split in two concepts.

If we apply the attribute-value extraction (AVE) module to these hypotheses the result is:

1. **Action**[] **Peripheral**[printer]
2. **Problem**[general_problem] **Peripheral**[printer]
3. **Problem**[general_problem] **Peripheral**[] **Peripheral**[printer]

We note that **Action** has an empty value since it was incorrectly annotated and, therefore, it is not supported by words from which the AVE module can extract a correct value. In this case, the output of AVE can only be empty. Similarly, for the third hypothesis, the AVE module cannot extract a correct value from the phrase "with the" since it doesn't contain any keyword for a **Peripheral** concept.

For each hypothesis, our SIM simply counts the number of wrong values. In the example above, we have 1, 0 and 1 for the three hypothesis, respectively. Accordingly, the most accurate hypothesis under SIM is the second, which is also the correct one.

Using SIM we generate a huge number of hypotheses with the baseline model and we select only the top *n*-best, to be used in the discriminative reranking.

MEDIA	training		test	
# sentences	12,908		3,005	
	words	concepts	words	concepts
# tokens	94,466	43,078	25,606	11,383
# vocabulary	2,210	99	1,276	78
# OOV rate [%]	–	–	1.39	0.04

Table 1. Statistics of the MEDIA training and evaluation sets used for all experiments.

4.2. Rerank Selection (RRS)

After the reranking model is applied, the top hypothesis is selected to be the final outcome. This solution assumes that the new hypothesis is more accurate then the one provided by the baseline model. In general, as we will see from results, this assumption is not true. This means that we can improve reranking performance by applying a strategy that detects when it is more likely that the best hypothesis of the baseline model is more accurate than the one provided by the reranker.

For this purpose we propose a simple strategy based on the scores computed by the two models involved in reranking: SFST or CRF for the baseline and SVM with PTK for reranking.

Using these scores, we train two thresholds for error rate minimization and we use them to re-select the final best hypothesis (BestHyp.) according to the following decision function:

$$BestHyp. = \begin{cases} HYP_{RR} & \text{if } C_{fst} \leq T_{fst} \; and \; C_{RR} \geq T_{RR} \\ HYP_{fst} & \text{otherwise.} \end{cases}$$

where HYP_{RR} and $HYP_{fst/crf}$ are the best hypothesis of the reranking and baseline models (SFST of CRF), respectively, T_{RR} and $T_{fst/crf}$ are the trained thresholds and $C_{fst/crf}$ and C_{RR} are the scores for the best hypotheses.

5. EXPERIMENTS

This section describes corpora and experiments for the evaluation of our approach.

5.1. Corpora

The corpus MEDIA was collected in the project MEDIA-EVALDA [7] for development and evaluation of spoken understanding models. The corpus is made 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 annotated data are reported in Table 1.

LUNA Italian	training		test	
# sentences	3,171		634	
	words	concepts	words	concepts
# tokens	30,470	18,408	6,436	3,783
# vocabulary	2,386	42	1,059	38
# OOV rate [%]	–	–	3.68	0.0

Table 2. Statistics of the latest version of the LUNA Italian training and evaluation sets used for all experiments.

Text Input	MEDIA		LUNA-IT	
Model	Attr	Attr+Val	Attr	Attr+Val
FST	14.2%	17.0%	24.4%	27.4%
CRF	11.7%	14.2%	21.3%	23.5%
FST+RR	11.9%	14.6%	21.3%	23.7%
CRF+RR	11.5%	14.1%	20.6%	23.1%
FST+RR_{HSC}	11.5%	14.0%	20.7%	22.8%
CRF+RR_{HSC}	11.2%	13.8%	19.9%	21.9%
FST+RR_S	11.3%	13.8%	19.2%	21.5%
CRF+RR_S	**11.1%**	**13.2%**	**19.0%**	**21.1%**

Table 3. Results of SLU experiments on the MEDIA and the Italian LUNA test sets on manual transcriptions (Text Input) for both attribute (Attr) names and attribute values (Attr+Val)

The LUNA Italian corpus, produced in the homonymous European project, is the first Italian dataset of spontaneous speech on spoken dialogs. It is based on help-desk conversations in a domain of software/hardware repairing [8]. The corpus is made of 723 Human-Machine dialogs (HM) acquired with a WOZ approach. The data have been split in training, development and test sets. Statistics for training and test sets are reported in Table 2.

5.2. Results

For the experiments presented in this work we used the AT&T FSM/GRM Tools for FST baseline model [13], for CRF baseline we used CRF++, available at http://crfpp.sourceforge.net/. For reranking models based on SVM and PTK we used SVM-Light-TK, available at http://disi.unitn.it/moschitti. Model parameters, as well as thresholds for the RRS strategy described in previous section, are tuned on the development set of each corpus. The number of hypotheses generated is 10 with the "traditional" reranking model, while when using the HSC strategy we generate 1.000 hypotheses and we keep the 10 best under the inconsistency metric. We performed experiments on both manual and automatic transcriptions of utterances. Automatic transcriptions were generated by ASR systems with a Word Error Rate (WER) of 31.4% and 27.0% on MEDIA and Italian corpus test sets, respectively. SLU results are expressed in terms of Concept Error Rate (CER).

Results of the experiments are reported in Table 3 and 4. The two tables show: a comparison of SFST and CRF baselines, i.e. FST and CRF; FST and CRF using the basic reranking model, without the enhancements proposed in this paper, i.e. FST+RR and CRF+RR; two reranking models based on FST and CRF using the HSC strategy, i.e. FST+RR_{HSC} and CRF+RR_{HSC}, respectively; and two reranking models using the HSC+RRS (RR_S) strategy, i.e. FST+RR_S and CRF+RR_S, respectively.
We note that in general CRF outperforms FST, this is not surprising since CRF is a global model able to take many features into account. This makes more interesting the analysis of results. The comparison

Speech Input	MEDIA		LUNA-IT	
Model	Attr	Attr+Val	Attr	Attr+Val
FST	28.9%	33.6%	36.4%	39.9%
CRF	24.3%	28.2%	31.0%	34.2%
FST+RR	25.4%	29.9%	32.6%	36.2%
CRF+RR	23.6%	27.2%	29.4%	32.6%
FST+RR_{HSC}	24.9%	28.7%	31.5%	34.6%
CRF+RR_{HSC}	22.9%	26.5%	29.0%	32.2%
FST+RR_S	24.5%	28.2%	30.7%	34.0%
CRF+RR_S	**22.7%**	**26.3%**	**28.3%**	**31.4%**

Table 4. Results of SLU experiments on the MEDIA and the Italian LUNA test sets on automatic transcriptions (Speech Input) for both attribute (Attr) names and attribute values (Attr+Val). The WER of the ASR is 31.4% on MEDIA and 27.0% on the LUNA Italian corpus

using various strategies shows that our reranking tends to achieve the maximum accuracy whatever the initial baseline model is. Indeed, although FST and CRF have different baselines (e.g. 14.2% and 11.7% on MEDIA, 24.4% and 21.3% on LUNA-IT on text Input), the results obtained when using our new reranking strategies are similar (11.3% and 11.1% on MEDIA, 19.2% and 19.0% on LUNA-IT). This is also true for speech input, although less evident since on an absolute scale CRF reranking is significantly better.
Moreover, the results are significantly higher than those provided by "traditional" reranking models. For example, for FST reranking, we have a relative improvement of 4.4% on MEDIA when using the HSC strategy (CER goes from 14.6 with FST+RR to 14 with FST+RR_{HSC}), and further 1.4% relative improvement when also using RRS (from 14 with FST+RR_{HSC} to 13.8 with FST+RR_S). The total relative improvement on "traditional" reranking model is 5.4% (from 14.6 to 13.8), bringing the relative improvement on the baseline from 14.1% to 18.8% (from 17.0 to 13.8).
The improvement is much higher on the LUNA-IT task and on speech input, where CER are higher and so there is a bigger room for improvements. For example, CER on attribute-value extraction on LUNA-IT and text input goes from 23.7% to 21.5%, (i.e. 9.2% of relative improvement), while on speech input we achieve a relative improvement of 5.6% on MEDIA and 6.1% on LUNA-IT (all the improvements are measured with respect to the "traditional" reranking model).
Reranking CRF hypotheses leads to similar improvement, but since CRF models are more accurate than FST, on absolute scale CRF reranking is more accurate than FST reranking. In particular, our results can be directly compared with the results reported in [10]. It can be seen that our CRF reranking approach, with 22.7% and 26.3% CER on MEDIA speech input, improves both the CRF baseline and the system combination based on ROVER described in [10], the latter obtained combining 5 SLU models.

6. CONCLUSIONS

In this paper we propose two strategies to improve discriminative reranking for SLU. One is based on a semantic inconsistency metric that can be used to select hypotheses for reranking. The other is based on the confidence values provided by the models involved in reranking and can be used to re-select the final best hypothesis. Both strategies achieve significant improvement on state-of-the-art reranking models.

Additionally we report for the first time results on CRF hypothe-

ses reranking for SLU, these remarkably improve the FST reranking and on MEDIA speech input they are also the new state-of-the-art.

An interesting future work is to implement the attribute value extraction module with a probabilistic model. Using the score provided by the model for each value in a hypothesis to compute the inconsistency measure would probably provide a more robust inconsistency metric.

7. ACKNOWLEDGEMENTS

This work has been partially funded by the European project LUNA, contract no. 33549 and by OSEO under the Quaero program.

8. REFERENCES

[1] Michael Collins. Discriminative reranking for natural language parsing. In *ICML*, pages 175–182, 2000.

[2] Michael Collins and Nigel Duffy. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems 14*, pages 625–632. MIT Press, 2001.

[3] Alessandro Moschitti, Daniele Pighin, and Roberto Basili. Tree kernels for semantic role labeling. *Computational Linguistics*, 34(2):193–224, 2008.

[4] Libin Shen, Anoop Sarkar, and Franz J. Och. Discriminative reranking for machine translation. In *HLT-NAACL*, pages 177–184, 2004.

[5] Alessandro Moschitti, Silvia Quarteroni, Roberto Basili, and Suresh Manandhar. Exploiting syntactic and shallow semantic kernels for question/answer classification. In *Proceedings of ACL'07*, Prague, Czech Republic, 2007.

[6] Marco Dinarelli, Alessandro Moschitti, and Giuseppe Riccardi. Concept segmentation and labeling for conversational speech. In *Interspeech*, Brighton, U.K., 2009.

[7] Hélène Bonneau-Maynard, Christelle Ayache, F. Bechet, A Denis, A Kuhn, Fabrice Lefèvre, D. Mostefa, M. Qugnard, S. Rosset, and J. Servan, S. Vilaneau. Results of the french evalda-media evaluation campaign for literal understanding. In *LREC*, pages 2054–2059, Genoa, Italy, May 2006.

[8] Marco Dinarelli, Silvia Quarteroni, Sara Tonelli, Alessandro Moschitti, and Giuseppe Riccardi. Annotating spoken dialogs: from speech segments to dialog acts and frame semantics. In *Proceedings of SRSL 2009 Workshop of EACL*, Athens, Greece, 2009.

[9] S. Quarteroni, G. Riccardi, and M. Dinarelli. What's in an ontology for spoken language understanding. In *Proc. of Interspeech 2009*, Brighton, U.K., 2009.

[10] Stefan Hahn, Patrick Lehnen, and Hermann Ney. System combination for spoken language understanding. In *Interspeech*, pages 236–239, Brisbane, Australia, 2008.

[11] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER). In *Proceedings 1997 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 347–352, Santa Barbara, CA, December 1997.

[12] Alessandro Moschitti. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In *Proceedings of ECML 2006*, pages 318–329, Berlin, Germany, 2006.

[13] M. Mohri, F. Pereira, and M. Riley. Weighted finite-state transducers in speech recognition. *Computer, Speech and Language*, 16(1):69–88, 2002.

[14] Shen, L. and Sarkar, A. and Joshi, A.K., "Using LTAG Based Features in Parse Reranking", Proceedings of EMNLP2003, 2003.

[15] Michael Collins, Nigel Duffy. New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron. *Proceedings of the Association for Computational Linguistics*, 263–270, 2002.