Generative and Discriminative Algorithms for Spoken Language Understanding

Christian Raymond, Giuseppe Riccardi

Generative and Discriminative Algorithms for Spoken Language Understanding

Christian Raymond, Giuseppe Riccardi

Department of Information and Communication Technology
University of Trento, Italy
raymond@dit.unitn.it, riccardi@dit.unitn.it

Abstract

Spoken Language Understanding (SLU) for conversational systems (SDS) aims at extracting concept and their relations from spontaneous speech. Previous approaches to SLU have modeled concept relations as stochastic semantic networks ranging from generative approach to discriminative. As spoken dialog systems complexity increases, SLU needs to perform understanding based on a richer set of features ranging from *a-priori* knowledge, long dependency, dialog history, system belief, *etc.* This paper studies generative and discriminative approaches to modeling the sentence segmentation and concept labeling. We evaluate algorithms based on Finite State Transducers (FST) as well as discriminative algorithms based on Support Vector Machine sequence classifier based and Conditional Random Fields (CRF). We compare them in terms of concept accuracy, generalization and robustness to annotation ambiguities. We also show how non-local non-lexical features (e.g. *a-priori* knowledge) can be modeled with CRF which is the best performing algorithm across tasks. The evaluation is carried out on two SLU tasks of different complexity, namely ATIS and MEDIA corpora.

Index Terms: spoken language understanding (SLU), conditional random fields (CRF), classifiers based sequence labeling, finite state transducers (FST).

1. Introduction

Spoken Language Understanding (SLU) aims at extracting concept and their relations from spontaneous speech. Conceptual structures have different levels of complexity, depending on the application domain. In data-driven approaches to SLU, a mapping between input strings (speech transcriptions) and output structures is learned. In the context of SDS, shallow parsing has been studied as an alternative to full-sentence parsing to cope with parser robustness issues and conversational speech disfluencies. For SLU, shallow parsing corresponds to the first step of understanding and aims at extracting sequence of semantic units called concepts. There are two basic problems in SLU, the semantic unit learning [1] and the concept segmentation. In this paper we address the second problem.
The concept segmentation is a sequence labeling task with words (or word lattices) as input and concept as output labels. Several methods have been applied for sequence labeling from generative to discriminative. [2, 3] use a finite state semantic tagger to get a flat-concept representation of the semantic. [4] extend the flat-concept model with the Hidden Vector State (HVS) model. This is a discrete Markov model in which context is encoded as a stack-oriented state vector in order to capture hierarchical structure in the data. Nevertheless discriminative approach, like sequential classification approach, or CRF, can handle many correlated features which would be difficult to integrate into generative models. [5] use combinations of forward and backward moving sequential SVM classifiers which is the system which performed the best in the CoNLL2000 Shared Task, Chunking and BaseNP Chunking task. [6] show the ability of CRF to integrate easily non-local feature and [7] show that CRF improve accuracy in comparison of a generative approach on ATIS. In this paper, we study three different methods, one generative and two discriminative, SVM based and CRF, applied to SLU tasks of different complexity. We show that the performance of models depends on the complexity of the task and training data set annotation ambiguity or errors.
The paper is organized as follows. Section 2 presents the models used. Section 3 introduces the datasets used to compare methods. Section 4 compares the models with experimental results, and Section 5 concludes the paper.

2. Statistical methods for sequence labeling

The sequence labeling problem for SLU is defined as the mapping between the word sequence W output by the automatic speech recognition (ASR) to a sequence of concepts C needed to perform understanding. An example for the ATIS task could be:

$$\underbrace{list}_{null}\ \underbrace{twa}_{airline\_code}\ \underbrace{flights}_{null}\ \underbrace{from}_{null}\ \underbrace{washington}_{fromloc.city}\ \underbrace{to}_{null}\ \underbrace{philadelphia}_{toloc.city}$$

Previous work used three main machine-learning approaches to sequence labeling. The first approach relies on k-order generative probabilistic models of paired input sequences and label sequences [2, 3, 4]. The second approach views the sequence labeling problem as a sequence of classification problems, one for each of the labels in the sequence [5]. The third approach relies on maximum entropy models like MEMMs or CRFs [8]. We choose one representing algorithm for each machine learning approach and we present them in the following sections.

2.1. Stochastic Finite State Transducers (SFST)

SFST based SLU is a translation process in which stochastic language models are implemented by Finite State Machines (FSM). There is an FSM for each elementary concept. These FSMs are transducers that take words as input and output the concept tag conveyed by the accepted phrase. They could be induced from an annotated corpus and/or manually designed. All these transducers are grouped together into a single transducer,

called λ_{w2c}, which is the union of all of them. A stochastic conceptual language model is computed as the joint probability $P(W, C)$:

$$P(W,C) = \prod_{i=1}^{k} P(w_i c_i | h_i)$$
$$\text{where } h_i = \{w_{i-1}c_{i-1}, \ldots, w_1 c_1\}$$

where $C = c_1, c_2, \ldots, c_k$ is the sequence of concepts and $W = w_1, w_2, \ldots, w_k$ is the sequence of words. h_i is approximated by $\{w_{i-1}c_{i-1}, w_{i-2}c_{i-2}\}$ as 3-gram model. This model called λ_{SLM} is also encoded as an FSM. In order to increase the generalization power of the model, some categories of words are replaced by labels (*e.g.* city, month, *etc.*). Given a new sentence W and its FSM representation λ_W, the translation process is to find the best path of the transducer resulting of the next composition:

$$\lambda_{SLU} = \lambda_W \circ \lambda_{w2c} \circ \lambda_{SLM}$$

In the SFST model, the best sentence segmentation (concept boundaries and labels) is computed over all possible hypotheses in λ_{SLU}.Note that this model supports λ_W generated from word lattices. All operations are done using the AT&T FSM/GRM Library [9].

2.2. Classifier-based sequence labeling

The sequential classification views the labeling problem as a sequence of classification problems, one for each of the labels in the sequence. This algorithm handles correlated and non-local features, but unlike generative models it cannot trade off decisions at different positions against each other. Yamcha [5] is the system which performed the best in the CoNLL2000 Shared Task, Chunking and BaseNP Chunking task. It uses heuristic combinations of forward-moving and backward-moving sequential SVMs classifiers taking as feature previous decisions it made. Once, SVMs are binary classifiers, they extend SVMs to K-class classifiers using pairwise classification, they build $K * (K-1)/2$ classifiers considering all pairs of classes, and final decision is given by their weighted voting.

2.3. Conditional Random Fields

CRFs bring together the best of generative and classification models. Like classification models, they accommodate many statistically correlated features of the inputs and train them discriminatively. Like generative models, they trade off decisions at different sequence positions, however it is not so easy to integrate them in the early stages of the speech recognition or with word lattices.

A conditional random field is defined by a dependency graph G and a set of features f_k to which are associated weights λ_k. The conditional probability of an annotation given an observation is given by:

$$p(y|x) = \frac{1}{Z(x)} exp(\sum_{c \in C} \sum_{k} \lambda_k f_k(y_c, x, c))$$
$$with$$
$$Z(x) = \sum_{y} exp(\sum_{c \in C} \sum_{k} \lambda_k f_k(y_c, x, c))$$

Semantic features, (*a-priori* defined) concept relations, *etc.*, are encoded in the model using these functions. In most case, the features are binary functions returning 1 if there is a match, 0 if not. These features take in parameter the values taken by the random variables (y_c) of the clique (c) to which they apply, and also the *whole* observation x. The weights λ_k associated to each features are the parameters of the model. Learning a CRF is to compute the weights λ_k.
We used in these experiments CRF++ [10] an open source implementation of Conditional Random Fields.

3. Datasets

In our experiments we used two datasets. ATIS is a publicly available corpus used in the early nineties for SLU evaluation. MEDIA has been recently collected and will be made available through ELRA.

3.1. ATIS

The Air Travel Information System (ATIS) task [11] is dedicated to provide flight information. The semantic representation used is frame based. The SLU goal is to find the good frame and fill the corresponding slots.

words: flights from boston to philadelphia		
FRAME:	FLIGHT	
	DEPARTURE.CITY = boston	
	ARRIVAL.CITY = philadelphia	

We start from the same dataset as [4]: the training set consists of 4978 utterances selected from the Class A (context independent) training data in the ATIS-2 and ATIS-3 corpora whilst the ATIS test set contains both the ATIS-3 NOV93 and DEC94 datasets. Each training utterance is annotated with an abstract semantic annotation which provides the hierarchical semantic relationships between semantic constituents where terminal are part of domain specific lexical classes (city, airline_name, *etc.*) which can be extracted for the application domain database:
flights from boston to philadelphia → FLIGHT(FROMLOC(CITY) TOLOC(CITY))
In order to learn directly from data, we fully annotate automatically the train set to get the words/concept pairs:

- replace all class members by their corresponding class names (boston, philadelphia → city)
- extract from the abstract semantic annotation the sequence of concepts (fromloc.city and toloc.city)
- build a regular expression to find concepts in words sequence (.*city.*city.*→ null fromloc.city null toloc.city)

3.2. MEDIA

The research project MEDIA [12] evaluates different SLU models of spoken dialogue systems dedicated to provide tourist in-

Table 1: Example of message with concept+value information. The original French transcription is: "*oui l'hôtel dont le prix est inférieur à cinquante cinq euros*"

n	W^c	c	**value**
1	yes	answer	yes
2	the	RefLink	singular
3	hotel	BDObject	hotel
4	which	null	
5	price	object	payment-amount
6	is below	comparative-payment	below
7	fifty five	payment-amount-int	55
8	euros	payment-currency	euro

Table 2: *MEDIA and ATIS test sets results given in term of precision, recall, F-measure and CER for all models FST, Yamcha, CRF. (cor) corresponds to the results on disambiguated data and (+f) to the results adding* a-priori *knowledge*

	ATIS Nov93+Dec94 test set results			
Model	Precision	Recall	F-measure(cor)(+f)	SER(cor)(+f)
FST	91.61%	91.87%	91.73%(94.11)	9.10(6.7)
Yamcha	90.54%	89.00%	89.76%(95.00)(95.74)	11.6(6.4)(5.5)
CRF	89.41%	89.25%	89.32%(95.00)(95.58)	11.6(6.0)(5.3)

	MEDIA test set results		
Precision	Recall	F-measure(+f)	CER(+f)
92.51%	91.31%	91.90%	11.8
93.02%	88.20%	90.54%(91.47)	14.5(13.4)
94.55%	90.75%	92.61%(92.96)	11.2(10.9)

formation. A 1250 French dialogue corpus has been recorded by ELDA following a Wizard of Oz protocol: 250 speakers have followed each 5 hotel reservation scenarios. This corpus has been manually transcribed, then conceptually annotated according to a semantic representation defined within the project. This representation is based on the definition of concepts that can be associated to 3 kinds of information. First a concept is defined by a label and a value; for example to the concept date can be associated the value 2006/04/02. Second a specifier can be attached to a concept in order to link the concept together in order to go from a flat concept/value representation to a hierarchical one; for example, to the concept date can be associated the specifiers *reservation* and *begin* to specify that this date is the beginning date of an hotel reservation. Third modal information is attached to each concept (positive, affirmative, interrogative or optional). Table 1 shows an example of message from the MEDIA corpus with the concept-value information only. The first column contains the segment identifier in the message, the second column shows the chunks W^c supporting the concept c of the third column. In the fourth column is displayed the value of the concept c in the chunk W^c. The semantic dictionary MEDIA contains 83 concept labels, 19 specifiers and 4 modal information. In this study we will focus on the concept extraction only. No specifiers, values or modal information are considered. So the tagset considered is made of 83 labels. The MEDIA corpus is split into 3 parts. The first part (720 dialogues, 12K messages) is used for training the models, the second (79 dialogues, 1.3K message) and the third part (200 dialogues, 3.4K message) are used as test.

4. Experiments

We evaluate the three methods on the ATIS and MEDIA manually transcribed utterances. We learn the three models (FST,Yamcha,CRF) using the same training corpus. FST is learn as 3-gram model since it's the parameter which give the best result, while discriminative methods, Yamcha and CRF in order to be comparable, use the same set of features. Features are the indicators for specific words (and their corresponding lexical class) in a window $[-4, 2]$ around the decision state. Since dependency graph of CRFs is a first order Markov chain, Yamcha uses as feature the previous label it chose. The evaluation of the concepts sequence decoded is given in term of F-measure (F) and concept error rate (CER). CER computation is computed like word error rate at the concept level, for ATIS it corresponds to the slot error rate.

$$F = \frac{2 \times precision \times recall}{precision + recall} CER = \frac{\#Ins + \#Subs + \#del}{\#concepts\_ref}$$

4.1. ATIS experiments

Results are reported in table 2. Figure 1 reports the accuracy with different size of training data. Surprisingly, FST

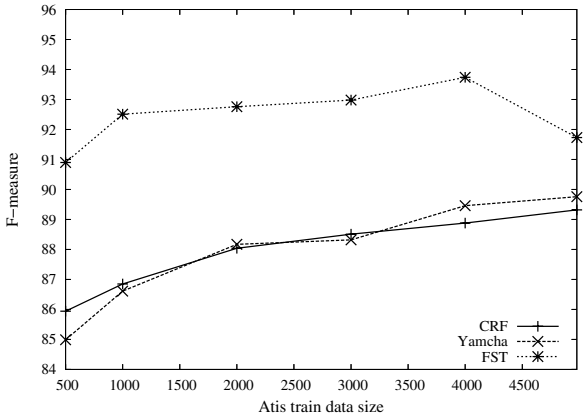

Figure 1: *F-measure on ATIS raw data learning curves*

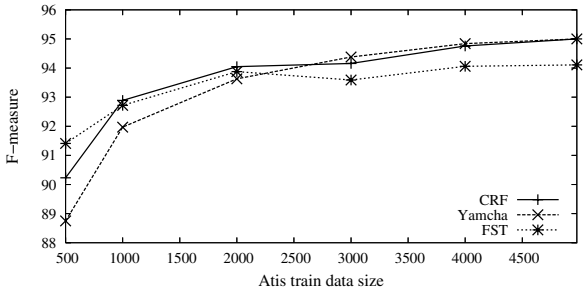

Figure 2: *F-measure on ATIS disambiguated data learning curves*

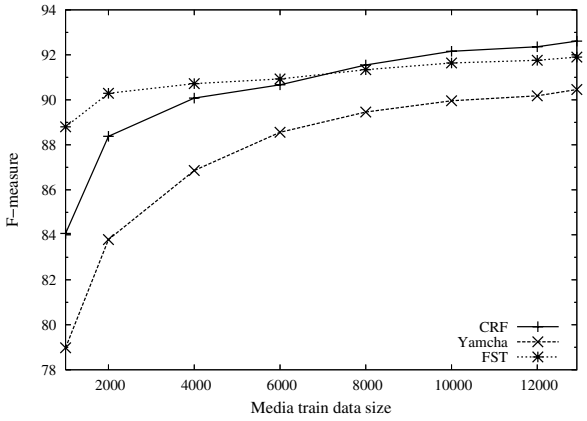

Figure 3: *F-measure on MEDIA data learning curves*

performs better than discriminative models by a large margin. Moreover, there is a drop on the FST curve when adding the last 1000 training examples. By manual inspection, we found ambiguous annotations such as: "CITY(BOSTON) NULL(to) TOLOC.CITY(DENVER)". Despite the fact that there is no explicit lexical information which indicates BOSTON is the departure city, the semantic is not ambiguous: BOSTON *is* a departure city. To correct most of these problems, we chose, when there is no explicit information of departure or arrival (for place or date) to use by default departure. This correction affects 448 examples in the training corpus. New results are reported in column (cor) in table 2 and in figure 2. With these corrections, all models benefit from the disambiguation, especially the discriminative which now outperform FST. In fact, discriminative models, both CRF and Yamcha, tend to fit the train data (100% accuracy on the train corpus) while FST stay at about 93%. Discriminative methods are really good to model rules from data, but in presence of annotation errors or ambiguities in the training data, they learn overfit false rules while FST have a good generalization.

4.2. MEDIA experiments

We did the same experiments on the MEDIA data, with same set of features and same parameters for all the models. Results are reported in table 2 and figure 3. In this case, CRFs outperforms both FST and Yamcha. As we saw in the ATIS experiments FST has very good generalization properties especially when small amount of data is available, CRFs outperforms FST after introducing about 7000 examples in the training corpus.

4.3. Adding *a-priori* knowledge in discriminative methods

The discriminative algorithms have been shown very powerful to integrate *a-priori* knowledge or long features dependency. Most of the ambiguity in the ATIS task is to determine if a city or a date is an arrival or departure information, long dependencies are often needed to solve this problem [6]. Taking a look at the train corpus, many examples containing arrival information follow the verb "to arrive". We add this feature in discriminative models as "do the verb *to arrive* is present in the utterance?". While this feature is simple, it copes with many ambiguities in the corpus. Both discriminative models benefit in term of model accuracy, see table 2, columns (+f).
We did a similar experiment on MEDIA. We computed the concepts confusion pairs obtained using the models on a development set and we chose the two first most confusable concept labels. For instance, the concepts 'TEMPS-DATE' and 'TEMPS-JOUR-MOIS' that could appear in the following situation:

...TEMPS-DATE-DEBUT(from the 15) TEMPS-DATE-FIN(to the 17 february)...
...TEMPS-JOUR-MOIS-DEBUT(from the 15) TEMPS-JOUR-MOIS-FIN(to the 17)...

in the first example, the concept is 'TEMPS-DATE-DEBUT' because the date is fully expressed (15 of february), in the second example, we have to find that 'from the 15' is a month day 'TEMPS-JOUR-MOIS-DEBUT' but we don't know this month. Intuitively, the ambiguity relies to a long dependency with the presence or not of a month name in the utterance. A new simple feature has been added to both discriminative model as "do a month name is specified in the utterance?". Like the ATIS experiments, while this simple and intuitive feature is not incorporable in FST, it is really benefit for both discriminative model and solves many of these ambiguities, see columns (+f) in table 2.

5. Conclusion

We compared generative FST and two discriminative, CRFs and classifier based, algorithms to perform concepts extraction and segmentation in the context of Spoken Language Understanding. We compared them in term of accuracy and robustness to data sparseness on two different tasks with different complexity. FST is both more robust to data sparseness and inconsistent training corpus than discriminative models, nevertheless with a sufficient amount of consistent training data, discriminative models CRFs outperform FST. Discriminative models showed the ability to incorporate easily *a-priori* knowledge and long dependencies. While it is difficult to introduce such information in generative models, introducing very intuitive and simple features in discriminative models permit to improve the accuracy consistently across SLU tasks.

6. Acknowledgements

We would like to thank Yulan He for sharing with us her ATIS annotated dataset. This work was founded by the European Commission - LUNA project contract N° 33549.

7. References

[1] A. L. Gorin, G. Riccardi, and J. H. Wright, "How may i help you ?" *Speech Communication*, vol. 23, no. 1-2, pp. 113–127, 1997.

[2] E. Levin and R. Pieraccini, "Concept-based spontaneous speech understanding system," in *EUROSPEECH*, Madrid, Spain, 1995, pp. 555–558.

[3] C. Servan, C. Raymond, F. Béchet, and P. Nocéra, "Conceptual decoding from word lattices: application to the spoken dialogue corpus media," in *ICSLP*, 2006.

[4] Y. He and S. Young, "Semantic processing using the hidden vector state model," *Computer Speech and Language*, vol. 19, no. 1, pp. 85–106, 2005.

[5] T. Kudo and Y. Matsumoto, "Chunking with support vector machines," in *NAACL*, 2001, pp. 1–8. [Online]. Available: http://chasen.org/~taku/software/yamcha

[6] M. Jeong and G. G. Lee, "Exploiting non-local features for spoken language understanding." in *ACL*, 2006.

[7] Y.-Y. Wang and A. Acero, "Discriminative models for spoken language understanding," in *ICSLP*, 2006.

[8] J. D. Lafferty, A. Mccallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001, pp. 282–289.

[9] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer, Speech and Language*, vol. 16, no. 1, pp. 69–88, 2002.

[10] T. Kudo, "Crf++." [Online]. Available: http://chasen.org/~taku/software/CRF++/

[11] D. A. Dahl, M. Bates, M. Brown, W. Fisher, K. Hunicke-Smith, D. Pallett, C. Pao, A. Rudnicky, and E. Shriberg, "Expanding the scope of the atis task: the atis-3 corpus," in *HLT '94*, 1994, pp. 43–48.

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