Comparing stochastic approaches to spoken language understanding in multiple languages

Stefan Hahn, Marco Dinarelli, Christian Raymond, Fabrice Lefèvre, Patrick Lehnen, Renato de Mori, Alessandro Moschitti, Hermann Ney, Giuseppe Riccardi

Comparing Stochastic Approaches to Spoken Language Understanding in Multiple Languages

Stefan Hahn, Marco Dinarelli, Christian Raymond, Fabrice Lefèvre, Patrick Lehnen, Renato De Mori, *Fellow, IEEE*, Alessandro Moschitti, Hermann Ney, *Fellow, IEEE*, and Giuseppe Riccardi, *Fellow, IEEE*

***Abstract*—One of the first steps in building a spoken language understanding (SLU) module for dialogue systems is the extraction of flat concepts out of a given word sequence, usually provided by an automatic speech recognition (ASR) system. In this paper, six different modeling approaches are investigated to tackle the task of concept tagging. These methods include classical, well-known generative and discriminative methods like Finite State Transducers (FSTs), Statistical Machine Translation (SMT), Maximum Entropy Markov Models (MEMMs), or Support Vector Machines (SVMs) as well as techniques recently applied to natural language processing such as Conditional Random Fields (CRFs) or Dynamic Bayesian Networks (DBNs). Following a detailed description of the models, experimental and comparative results are presented on three corpora in different languages and with different complexity. The French MEDIA corpus has already been exploited during an evaluation campaign and so a direct comparison with existing benchmarks is possible. Recently collected Italian and Polish corpora are used to test the robustness and portability of the modeling approaches. For all tasks, manual transcriptions as well as ASR inputs are considered. Additionally to single systems, methods for system combination are investigated. The best performing model on all tasks is based on conditional random fields. On the MEDIA evaluation corpus, a concept error rate of 12.6% could be achieved. Here, additionally to attribute names, attribute values have been extracted using a combination of a rule-based and a statistical approach. Applying system combination using weighted ROVER with all six systems, the concept error rate (CER) drops to 12.0%.**

***Index Terms*—Generative and discriminative models, spoken dialogue systems, system combination.**

Manuscript received June 07, 2010; revised September 15, 2010; accepted October 29, 2010. Date of publication November 18, 2010; date of current version May 25, 2011. This work was supported in part by the European Union under the integrated project LUNA—spoken Language UNderstanding in multilinguAl communication systems (FP6-033549). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gokhan Tur.

S. Hahn, P. Lehnen, and H. Ney are with the Chair of Computer Science 6, Computer Science Department, RWTH Aachen University, D-52056 Aachen, Germany (e-mail: hahn@cs.rwth-aachen.de; lehnen@cs.rwth-aachen.de; ney@cs.rwth-aachen.de).

M. Dinarelli was with the Department of Information Engineering and Computer Science, University of Trento, 38050 Povo—Trento, Italy. He is now with LIMSI-CNRS, 91403 Orsay Cedex, France (e-mail: marcod@limsi.fr).

G. Riccardi and A. Moschitti are with the Department of Information Engineering and Computer Science, University of Trento, 38050 Povo—Trento, Italy (e-mail: riccardi@disi.unitn.it; moschitti@disi.unitn.it).

C. Raymond was with LIA—University of Avignon, BP1228 84911 Avignon Cedex 09, France. He is now with Université Européenne de Bretagne, IRISA-INSA, UMR 6074, Rennes, France (e-mail: christian.raymond@irisa.fr).

F. Lefèvre is with LIA—University of Avignon, BP1228 84911 Avignon Cedex 09, France (e-mail: fabrice.lefevre@univ-avignon.fr).

R. de Mori is with LIA—University of Avignon, BP1228 84911 Avignon Cedex 09, France, also and with the Department of Computer Science, McGill University, Montreal, QC H3A 2T5, Canada (e-mail: renato.demori@univ-avignon.fr; rdemori@cs.mcgill.ca).

Digital Object Identifier 10.1109/TASL.2010.2093520

I. Introduction

COMPUTER interpretation of spoken language is a complex process performed by a spoken language understanding (SLU) system that can be decomposed into various tasks using different methods and models. A possible decomposition includes an automatic speech recognition (ASR) task to obtain a sequence or a lattice of word hypotheses and an interpretation task that transforms word hypotheses into semantic structure hypotheses described by a meaning representation language (MRL). Based on linguistic theories, described for example in [1], semantic structures are obtained by composition of semantic constituents that are fragments of the SLU system application ontology.

Interpretation of automatic transcriptions of speech is difficult because errors are introduced by the ASR process. In order to take into account the effects of errors and imprecision, probabilistic interpretation models have been introduced. Conceptual hidden Markov Models (HMM) are proposed in the Chronus system [2]. The model is based on a view of utterances as generated by a process using a model whose observations are word hypotheses and hidden states correspond to meaning units called *concepts*.

Generative approaches model the joint probability $P(w_1^N, c_1^T)$ of a sequence of words $w_1, \ldots, w_N$ and a sequence of semantic constituents (*concepts*) $c_1, \ldots, c_T$. Thus, they are able to generate samples from the joint distribution. Dynamic Bayesian Networks (DBNS) are proposed and evaluated in this paper as a generative model more powerful than first-order HMMs.

Another possibility is to consider discriminative classification models for computing the conditional probability distribution of $P(c_1^T \mid w_1^N)$. It has been shown that generative models can be converted into discriminative models, at least in principle [3].

Some generative and discriminative approaches have been compared in the literature, e.g., in [4]–[7]. In the first reference, it is concluded that discriminative training is expensive albeit more robust and special knowledge about the true distribution is not needed. In contrast, training of generative approaches is cheap but the models need to fit well to the true distribution.

DBN and discriminative models based on support vector machines (SVMs) are also compared in this paper showing that SVMs outperform DBNs.

Stochastic grammars like the ones described in [8], [9] have been proposed for SLU. It has been observed that spoken language does not always follow the rules of a formal grammar and

that it is difficult to obtain correct parse trees from imprecise ASR hypotheses. For this reason, the possibility of performing partial parsing has been considered. Semantic constituents that are fragments of the application ontology have been introduced. They have been conceived in such a way that it is possible to annotate them by associating each of them with finite length sequences of words. Stochastic finite state transducers (FST) have been obtained from constituent annotations. These generative models describe local syntactic structures with a sequence of words like noun phrases with a variable and possibly long sequence of words.

In the attempts to combine features from generative and classification models, exponential models have also been considered and evaluated. Some of them are used in stochastic machine translation (SMT) processes from natural language to the constituent MRL improving early approaches proposed in [10], while some others are based on maximum entropy Markov models (MEMM) and conditional random fields (CRF).

Experiments reported in this paper show that CRFs systematically outperform all the other methods even using fairly simple functions in the model exponents. The proposed CRFs seem to model the overall expression of a concept better than the other considered models when this semantic information is conveyed by word sequences. This does not appear to be the case for spoken opinion analysis performed on arbitrarily long telephone messages and dialogs as described in [11].

This paper describes and compares the use of the just introduced generative, discriminative, and exponential models on the French MEDIA corpus [12].

Two more corpora comparable to the MEDIA corpus in size and detail of annotation have been collected within the EU FP6 LUNA project: the Polish Warsaw transportation corpus [13] and the Italian help-desk corpus [14]. The considered corpora have ontologies of different types and complexity that can be represented in a frame language described in [7]. In tasks like MEDIA, there are frames describing properties of objects in application domains and frames describing dialog acts (DAs). These frames have some properties whose values are instances of other frames resulting in fairly complex semantic structures. Attribute value logical predicates can be obtained from these frames, an attribute being a frame property. When the value of a property is a frame structure, this structure can be represented by a semantic class name. For example, the request for a reservation is represented by a frame REQUEST that has a property with name `request_object`. Value types for an object representing this property are listed in the slot facet of the property. The facet of `request_object` contains a structure type represented by a semantic class whose name is RESERVATION. In the MEDIA annotation a name corresponding to the property `request_object` of the frame REQUEST will have values corresponding to the elements of the property slot facet. References are also examples of other elements in the facet. In case of ambiguities, disambiguation is performed by constituent composition, a process that is not described in this paper.

A distinction is made between two tasks: extraction of only attribute name and extraction of attribute name with corresponding attribute value.

Additionally, two conditions are considered, namely, manual transcriptions of word hypotheses as input, which can be considered more or less flawless, and automatically generated transcriptions using an ASR system. Corpora in the three languages, namely French, Italian, and Polish are used for training and testing of the methods. The methods are not new, but some of them are applied for the first time to SLU.

A review of motivations and solutions using some of these models can be found, for example, in [7]. In addition to the extensive and consistent experimental comparison of six different statistical methods on MEDIA as well as on two newly collected corpora in different languages, this paper describes an original application of DBNs to SLU, improved CRFs by introducing margin posteriors leading to best published results on the MEDIA corpus in relaxed-simplified condition, ROVER system combination, a new FST-based re-ranking method using six systems all carefully tuned on exactly the same data and statistical improved attribute value extraction using CRFs in combination with rule-based attribute value extraction.

There are certain similarities between tasks such as part-of-speech (POS) tagging [15], name transliteration [16], or grapheme-to-phoneme conversion [17] and concept tagging suggesting that some findings described in this paper may be helpful also for these tasks.

Methods and models are reviewed in Section II. Section III describes methods for attribute value extraction, namely rule-based and statistical. After the presentation of the training and testing data in Section IV, the experimental results for the single systems are presented in Section V. The possibility of reducing interpretation errors by combining some of the proposed methods is discussed in Section VI.

II. Modeling Approaches

In this section, six different approaches to the task of concept tagging are presented. They include classical, well-known methods based on finite state transducers (FSTs) or support vector machines (SVMs) as well as techniques recently applied to natural language processing such as conditional random fields (CRFs) or dynamic Bayesian networks (DBNs). Since two approaches use log-linear models, one subsection will give an overview of these techniques. It is followed by a presentation of the theoretical background of each method.

For consistency, the following naming scheme will be used throughout this paper.

- **Concept**: a set of attributes which is assigned to a sequence of words. This set contains up to two elements: the attribute name and the attribute value.
- **Attribute name**: the tag representing the semantic meaning of the word sequence. The attribute name is required for each concept.
- **Attribute value**: depending on the attribute name, there may be an associated normalized value which has to be extracted additionally from the word sequence.

A. Alignment

Within all approaches, except for the DBN approach, the probability of the *concept* sequence $P(c_1^T \mid w_1^N)$ is projected

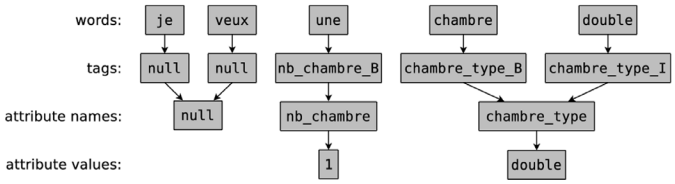

Fig. 1. Example illustrating the general idea of concept tagging (French: "I would like a double-bed room"). The first line shows the input word sequence, the third and fourth line the appropriate attribute names and values. The second line shows how the 1-to-1 alignment is modelled using "begin" (B) and "inside" (I) tags.

to the probability of the *concept tag* sequence $P(t_1^N \mid w_1^N)$ by assigning "begin" (B) and "inside" (I) markers to concepts, so as to model a 1-to-1 alignment. Here, the so-called *BIO scheme*, proposed in [18], has been adopted, e.g., the utterance part "chambre double" in Fig. 1 is mapped to

$$\underbrace{\text{``chambre''} : \text{chambre} - \text{type\_begin}}_{w_4 : t_4}, \times \underbrace{\text{``double''} : \text{chambre} - \text{type\_inside}}_{w_5 : t_5}.$$

Using this approach results in a 1-to-1 alignment and the original attribute name sequence can be recovered. It should be noted that the concept tags are just introduced for modeling reasons and do not appear in the final output of the systems. Fig. 1 gives an example from the French MEDIA corpus [19]. The input word sequence is shown in the first line, the resulting attribute names and accompanying values are shown in lines 3 and 4. Concept tags, including BIO notation, are given in line 2.

B. Log-Linear Models

We are using two log-linear models, which only differ in the normalization term. The first one is normalized on a positional level (maximum entropy Markov models [20], MEMM) and the second one on sentence level (conditional random fields [21], *CRF*). The general representation of these models is described in (1) as a conditional probability of a concept tag sequence $t_1^N = t_1, \ldots, t_N$ given a word sequence $w_1^N = w_1, \ldots, w_N$

$$p(t_1^N \mid w_1^N) = \frac{1}{Z} \prod_{n=1}^{N} \exp\left(\sum_{m=1}^{M} \lambda_m \cdot h_m\left(t_{n-1}, t_n, w_1^N\right) \right). \tag{1}$$

The log-linear models are based on feature functions $h_m(t_{n-1}, t_n, w_1^N)$ representing the information extracted from the given utterance, the corresponding parameters λ_m which are estimated in a training process, and a normalization term Z discussed in Sections II-B3 and II-B2, respectively, for each model.

1) Feature Functions: In our experiments, we use binary feature functions h_m. If a predefined combination of the values $t_{n-1}, t_n, w_1, \ldots, w_N$ is found in the data, the value "1" is returned, otherwise the value is "0." For instance a feature function may fire if and only if:

- the predecessor word w_{n-1} is "the" and the concept t_n is "name";
- the predecessor concept t_{n-1} is "number" and the concept t_n is "currency"'
- the prefix of a word w_n = "euros" (resp. word stem) of length $\delta = 4$ is "euro" and the concept t_n is "currency".

We will call the feature functions based on predecessor, current, and successor words *lexical features* and the features based on the predecessor concept *bigram features*. Features based on word parts (e.g., prefixes, suffixes, capitalization) are referred to as *word part features*.

Feature cutoffs are not applied. Thus, a feature h_m is used if it is seen with any combination of t_n, t_{n-1}, and w_1^N from the training corpus.

For clarity, we will abbreviate the term in the numerator of (1) by

$$H\left(t_{n-1}, t_n, w_1^N\right) = \exp\left(\sum_{m=1}^{M} \lambda_m \cdot h_m(t_{n-1}, t_n, w_1^N) \right)$$

resulting in

$$p\left(t_1^N \mid w_1^N\right) = \frac{1}{Z} \prod_{n=1}^{N} H\left(t_{n-1}, t_n, w_1^N\right) \tag{2}$$

2) Maximum Entropy Markov Models (MEMM): A possible normalization of (2) is on a positional level:

$$Z = \prod_{n=1}^{N} \sum_{\tilde{t}} H\left(t_{n-1}, \tilde{t}, w_1^N\right). \tag{3}$$

Here, $\tilde{t}$ stands for all possible concept tags. Using (2) with normalization (3) and a given training dataset $\{\{\bar{t}_1^N\}_s, \{w_1^N\}_s\}_{s=1}^S$, the criteria for training and decision making are given by

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \left\{ \sum_{s=1}^{S} \log p\left(\{\bar{t}_1^N\}_s \,\middle|\, \{w_1^N\}_s \right) - c \left\| \lambda_1^M \right\|^2 \right\} \tag{4}$$

using a L2-regularization constant c, and

$$\hat{t}_1^N = \underset{t_1^N}{\operatorname{argmax}} \left\{ p\left(t_1^N \mid w_1^N\right) \right\} \tag{5}$$

respectively. Here, $\bar{t}_1^N$ denotes the reference tag sequence. This modeling approach is also referred to as maximum entropy Markov models [20], maximum entropy [22] approach, or log-linear on position level [23] in the literature.

3) Linear Chain Conditional Random Fields (CRFs): Linear Chain Conditional Random Fields as defined in [21] could be represented with (2) and a normalization Z at sentence level

$$Z = \sum_{\tilde{t}_1^N} \prod_{n=1}^{N} H\left(\tilde{t}_{n-1}, \tilde{t}_n, w_1^N\right) \tag{6}$$

Here, $\tilde{t}_1^N$ represents all possible concept tag sequences.

For CRFs, the same training and decision criteria as for MEMMs are used [cf. (4) and (5)].

In [24], the idea of merging the optimization of feature weights (training) based on SVMs and CRFs, called MMI there, is described. The authors start from an SVM training process described by

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \left\{ -\frac{1}{S} \sum_{s=1}^{S} l\left(\{\bar{t}_1^N\}_s, d, \rho \right) - c \left\| \lambda_1^M \right\|^2 \right\} \tag{7}$$

with the distance

$$d = \sum_{m=1}^{M} \lambda_m \left(h_m \left(\{\bar{t}_1^N\}_s, w_1^N \right) - h_m \left(t_1^N, w_1^N \right) \right) \quad (8)$$

and the hinge loss

$$l\left(\bar{t}_1^N, d, \rho\right) = \max_{t_1^N \neq \bar{t}_1^N} \left\{ \max\left\{ \rho - d\left(\bar{t}_1^N, t_1^N\right), 0 \right\} \right\} \quad (9)$$

Equations (4) and (7) differ mainly in the loss function. They smoothed the loss function, used the accuracy instead of the 0/1 loss, and added it to the loss function of MMI resulting in a modified posterior defined as

$$p_{\Lambda,\rho}\left(t_1^N \mid w_1^N\right) = \frac{1}{Z'} \exp\left(\sum_{i=1}^{I} \lambda_i f_i\left(t_1^N, w_1^N\right) - \rho \mathcal{A}\left(t_1^N, \bar{t}_1^N\right) \right) \quad (10)$$

The normalization constant Z' is similarly defined as above:

$$Z' = \sum_{\tilde{t}_1^N} \exp\left(\sum_{i=1}^{I} \lambda_i f_i\left(\tilde{t}_1^N, w_1^N\right) - \rho \mathcal{A}\left(\tilde{t}_1^N, \bar{t}_1^N\right) \right). \quad (11)$$

Here, the margin score is set to the word accuracy

$$\mathcal{A}\left(t_1^N, \bar{t}_1^N\right) = \sum_{n=1}^{N} \delta(t_n, \bar{t}_n) \quad (12)$$

between the hypothesis t_1^N and the reference $\bar{t}_1^N$, scaled by $\rho \geq 0$. The margin-based training criteria are obtained by replacing the posterior. Note that only the training and not the decision process is changed. Further extensions of CRFs for SLU are possible, e.g., Triangular-CRFs as suggested in [25], taking dialog manager states into account.

Both MEMM and CRF are realized using an inhouse software. If not stated otherwise, the presented results using the CRF approach always include the margin term. A detailed comparison of CRFs with and without margin term can be found in [26].

C. Statistical Machine Translation (SMT)

A standard phrase-based machine translation method which combines several models is used. The incorporated models include phrase-based models in source-to-target and target-to-source direction, IBM-1 like scores at phrase level, again in source-to-target and target-to-source direction, a target language model, and additional word and phrase penalties. These models are log-linearly combined [27]:

$$\hat{t}_1^N = \underset{t_1^N}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m \log\left(p_m\left(t_1^N, w_1^N\right)\right) \right\}. \quad (13)$$

Here, $\log(p_m(\cdot))$ represents feature functions (which are the aforementioned statistical models) and λ_m the corresponding scaling factors. These factors are optimized using some numerical algorithm in order to maximize translation performance on a development corpus. In this case, optimization of the scaling factors is done with respect to the CER score using the downhill simplex algorithm. In contrast to general translation models, reordering of the target phrases composing the translation is not needed for NLU.

There is a certain relation between SMT and the log-linear models presented in the previous section. The feature functions in this case are statistical models which return float values, i.e., the features are no more binary. Merely seven parameters for the combination of the models are tuned in contrast to the millions of parameters used within CRFs.

D. Support Vector Machines (SVM)

SVMs are used to implement a local classifier-based approach to concept tagging, where the labeling problem is seen as a sequence of classification problems (one for each of the labels in the sequence). The algorithm handles correlated and non-local features, but unlike generative models it cannot trade off decisions at different positions against each other. YAMCHA, one system based on this approach, performed best in the CoNLL2000 Shared Task on chunking and BaseNP chunking [28]. It uses heuristic combinations of forward-moving and backward-moving sequential SVMs classifiers taking the previous decisions it made as features. Since SVMs are binary classifiers, the system extends SVMs to K-class classifiers using pairwise classifications. Therefore, $(K(K-1)/2)$ classifiers are built considering all pairs of classes. The final decision is given by their weighted voting.

The open-source toolkit YAMCHA is applied in the experiments [29].

E. Stochastic Finite State Transducers (FST)

The FST approach is a stochastic generative approach which computes the joint probability between the word sequence and the concept tag sequence. This approach is well suited to process speech since it is based on the paradigm generally used for the automatic speech recognition process.

The decoding process is to find the concept tag sequence maximizing $p(t_1^N \mid A)$, A being the acoustic observation of the user's speech. Finding the best concept tag sequence $\hat{t}_1^N$ given the acoustic observations A is formulated as

$$\begin{aligned} \hat{t}_1^N &\approx \underset{t_1^N}{\operatorname{argmax}} \sum_{w_1^N} p\left(A \mid w_1^N, t_1^N\right) p\left(w_1^N, t_1^N\right) \\ &\approx \underset{t_1^N, w_1^N}{\operatorname{argmax}} p\left(A \mid w_1^N\right) p\left(w_1^N, t_1^N\right). \end{aligned} \quad (14)$$

The probability $p(A \mid w_1^N)$ is estimated by the acoustic models of the speech recognition system. $p(w_1^N, t_1^N)$ is the joint probability between a sequence of words and concept tags. The joint probability is estimated as a *tri*-gram:

$$\begin{aligned} p\left(w_1^N, t_1^N\right) &= \prod_{n=1}^{N} p(w_n, t_n \mid h_n) \\ &\text{with } h_n = (w_{n-1}, t_{n-1}), (w_{n-2}, t_{n-2}) \end{aligned} \quad (15)$$

This decoding process is usually done sequentially with w_1^N fixed (i.e., w_1^N is the best transcription hypothesis made by the ASR). The strength of this approach is to make an "integrated" decoding able to process word graphs performing operations

with finite state transducers using the AT&T FSM/GRM Library [30]. The best sequence of couples word/concept is the best path of the transducer λ_{SLU} resulting in the composition of five transducers:

$$\lambda_{\mathrm{SLU}} = \lambda_G \circ \lambda_{\mathrm{gen}} \circ \lambda_{w2c} \circ \lambda_{\mathrm{SLM}}[\circ\lambda_v].$$

The five finite state transducers are defined as follows.

1) λ_G is a finite state machine representing a word graph generated by the ASR with the acoustic scores [$p(A \mid w_1^N)$ in (14)]. In the following experiments, in order to be comparable with other methods, λ_G encodes the ASR one-best hypothesis only.
2) λ_{gen} converts words to categories (e.g., CITIES, MONTH,...), it represents the *a priori* knowledge of the task and allows for a better generalization on the training data.
3) λ_{w2c} translates phrases to concepts. It is induced from the training data and/or with handwritten grammars (e.g., dates or prices). It contains the possible sequences of words (and/or categories) which could emit a concept but also a filler model able to accept the word sequences that do not support any concepts (i.e., words associated with the dummy concept).
4) λ_{SLM} encodes the stochastic conceptual language model computing the joint probability of (15). Notice that in order to compute this probability a word w is replaced by its class in order to have a better estimation.
5) λ_v converts chunks (i.e., phrases associated with an attribute name) to a normalized attribute value with a procedure analogous to the use of rule-based methods described in Section III–1.

F. Dynamic Bayesian Networks (DBNs)

DBNs provide a great flexibility for complex stochastic system representation. In the last years, they have been used for many sequential data modeling tasks such as speech recognition, part-of-speech tagging, dialog-act tagging [31], and DNA sequence analysis. Their application to SLU is described here following preliminary presentations in [32] and [33]. With regard to their underlying probabilities, FST and DBN approaches are pretty comparable. However the two frameworks differ in the way they perform computation of inference, decoding and the relative probabilities. Interpretation in the proposed DBN framework is carried out by a purely stochastic process including the value normalization step (see Section III). It is based on the following decision rule:

$$\begin{aligned}\hat{c}_1^N, \hat{v}_1^N &= \underset{c_1^N, v_1^N}{\operatorname{argmax}} p\left(c_1^N, v_1^N \mid w_1^T\right) \\ &= \underset{c_1^N, v_1^N}{\operatorname{argmax}} p\left(w_1^T \mid c_1^N, v_1^N\right) p\left(v_1^N \mid c_1^N\right) p\left(c_1^N\right).\end{aligned} \tag{16}$$

Marginalization of (16) is required to derive concept hypotheses

$$\hat{c}_1^N = \underset{c_1^N}{\operatorname{argmax}} \sum_{v_1^N} p\left(w_1^T \mid c_1^N, v_1^N\right) p\left(v_1^N \mid c_1^N\right) p\left(c_1^N\right). \tag{17}$$

A peculiar aspect of the approach is that decoding is performed at the segmental level, i.e., models have an inner mechanism to deal with transitions from a concept segment to another. Hypothesization of values is performed under the same scheme and is detailed in Section III.

In our context, all variables are observed during training. The conditional probability tables can be directly derived avoiding EM iterations from observation counts. To improve their estimates, factored language models (FLMs) have been used, along with generalized parallel backoff (GPB) [34]. FLMs are an extension of standard n-gram LMs. They are based on a set of features not limited to previous word occurrences. Furthermore, GPB extends the standard backoff procedures to handle heterogeneous feature types not necessarily in a rigid temporal order. Unlike classical LM features, FLM features may appear at any time, including the time of prediction. Several FLM implementations are used in the SLU models:

- $p(c_1^N) \simeq \prod p(c_i \mid c_{i-h}^{i-1})$: attribute name sequences;
- $p(v_1^N \mid c_1^N) \simeq \prod p(v_i \mid c_i)$: attribute normalized values conditioned on attribute names;
- $p(w_1^T \mid c_1^N) \simeq \prod p(w_i \mid w_{i-h}^{i-1}, c_i)$: word sequences conditioned on attribute names (GPB works with order w_{i-h}^{i-1}, c_i);
- $p(w_1^T \mid v_1^N, c_1^N) \simeq \prod p(w_i \mid w_{i-h}^{i-1}, v_i, c_i)$: word sequences conditioned on attribute names and values (GPB works with order w_{i-h}^{i-1}, c_i, v_i);

where h represents a history which could vary according to the length of the model used (either 2-grams or 3-grams in our system). GPB uses the modified Kneser–Ney discounting technique [35] in all conditions. In the actual DBN models used in our system, the concept and value decoding steps are decoupled. The conceptual decoding process generates concept and transition sequences that become observed variables for the value decoding (see Section III).

III. Attribute Value Extraction

It is possible to hypothesize attributes represented by concept tags and their values in a single computational process. Except for simple application domains, values are characterized by different model types. For example, dates are well represented by regular expressions, while city names, even in the case of compound names, are better represented as single lexical items. For this reason, it may be practically useful to hypothesize attribute names first and use these hypotheses to constrain the type of knowledge used for hypothesizing values. Furthermore, in certain cases, values are normalized, especially when they can be expressed with synonyms.

For example, in the sentence of the MEDIA corpus "*I'd like a room for no more than fifty euros*," normalization translates the sequence "*no more than*" expressing the attribute name `comparative-payment-room` to the normalized value "*less than*."

The attribute values in the MEDIA corpus are numeric units, proper names or semantic classes. Thus, for the MEDIA task, there are three different value extraction models, namely:

- a value enumeration (e.g., "comparative" with possible values "around", "less-than", "maximum", "minimum" and "more-than");

- regular expressions (as for dates); or
- open values (i.e., no restrictions, as for client's names).

This normalization step is commonly based on deterministic rules, but can also be introduced in a global stochastic model by introducing an additional level. Several approaches have been considered based on rules and stochastic models.

The experimental results summarized in Table VI refer to three different tasks for which a rule-based approach requiring a human effort outperforms a statistical approaches based on CRFs.

1) Approaches Based on Deterministic Rules: The normalization step based on manual rules is exploited within the finite state transducer approach and a script language based approach. Both use simple concept attribute dependent expressions to convert phrases supporting an attribute name into a normalized attribute value.

Hand-crafted rules for value hypothesization are obtained with the training data and can be encoded into FSTs or procedurally into script languages. When word graphs are available, FSTs can be composed with them to obtain graphs of concept hypotheses. The result can be kept for further processing [36], [37]. As for the experiments described in the following only the most likely ASR hypothesis has been used, rules have been encoded in scripts.

2) Stochastic Approaches: Despite a general formulation, stochastic approaches for value normalization in the SLU context are very tied to the model into which they are integrated. In principle, a stochastic approach for attribute value extraction can be realized for each method presented in the paper. In this paper, DBN and CRF approaches for value normalization have been integrated into the models used for concept hypothesization.

a) DBN: The decoding process hypothesizes a combined sequence of concepts with their values according to (16). As a consequence, the word sequence probabilities become conditioned both on the concepts and their normalized values. The complexity of the conceptual model is then greatly increased making it necessary to simplify the decoding process. Traditional suboptimal decoding strategies such as beam search do not perform well in these tasks [32]. Under the assumption that normalized values have a slight or no influence on the segmentation process, a better setup is to first marginalize v [see (16)] then hypothesize v given $\hat{c}$, the hypothesis from the former level:

$$\hat{v}_1^N = \underset{v_1^N}{\operatorname{argmax}}\, p\left(w_1^T \mid \hat{c}_1^N, v_1^N\right) p\left(v_1^N \mid \hat{c}_1^N\right) p\left(\hat{c}_1^N\right). \quad (18)$$

Even if the design of rules is costly because it requires the intervention of human experts, the use of rules outperforms the results obtained with just stochastic approaches. In the experiments described in this paper, only results obtained with deterministic rules for normalization are reported for DBNs. Nevertheless, some improvements are obtained with a stochastic approach including handcrafted rules, as it is presented with the CRF model below.

b) CRF: Knowing the location and the attribute name of content words given by the attribute name extraction, normalized values are hypothesized in a successive step for most of the attribute names as in the following example from the Polish corpus concerning a *Request* about bus *151*:

@Action[Request]{chciałam} @BUS[151]{linię sto pięćdziesiąt jeden}...

@Action[Request]{I would like} @BUS[151]{line one hundred fifty one}...

A 1-to-1 mapping like in attribute name extraction is not used, instead exactly one value is hypothesized per attribute name. Therefore, a CRF model is trained on the word/attribute name pairs on source side and the attribute values on target side. Thus, search is constrained to the set of seen attribute values for each attribute name. Additionally, mixing of attribute values is not allowed. Lexical features on the predecessor, the current, and the successor word were used. For attribute names with a huge number of values, it is possible to reduce the search space only to a *null* value, leaving the attribute value extraction to a rule-based approach in a possible postprocessing step.

In the reported experiments, CRFs for attribute value extraction have only been included in the CRF approach and always in combination with rules. One reason supporting this combination is that the number of possible values varies highly between attribute names. For example, always in the Polish corpus, the attribute name *Reaction* can take either the value "Confirmation" or "Negation" and is triggered by only few content words. In contrast, the value of *STREET_NUMB* can at least theoretically be any number. It is likely that not all these numbers appear in the training corpus, which is the only information source available for training models in purely data driven approaches.

IV. Corpora Description

The evaluation of the just introduced methods was carried out using three different telephone speech corpora in three different languages. The evaluation made it possible to compare performances on the manually annotated data with the annotations obtained with ASR hypotheses and to establish and to observe some trends consistent across the corpora. The French MEDIA corpus was publicly available with manual transcriptions and annotations in terms of concept tags and values. It consists of human–machine dialogues collected with a wizard of Oz procedure involving selected speakers. The other two corpora were specifically acquired, manually transcribed and annotated with semantic information for the experiments described in this paper. The Polish corpus consists of human–human conversations recorded in the call center of the Polish Warsaw transportation system while the Italian corpus consists of dialogues of a help-desk application in which the employees of the Consorzio per il Sistema Informativo Piemonte, a public regional institution, seek advice on problems related to their computers. The characteristics of the three corpora are summarized in Table I. Since we only want to perform concept tagging on word sequences uttered by (human) users, only these user turns have been used. No filtering of turns has been carried out. Thus, some turns may contain just the NULL tag indicating chunks that do not convey a meaning relevant for the application domain.

A. French MEDIA Corpus

This corpus was collected in the French Media/Evalda project in the domain of negotiation of tourist services [19]. It is divided

into three parts: a training set (approx. 13 k sentences), a development set (approx. 1.3 k sentences) and an evaluation set (approx. 3.5 k sentences).

There are 99 different attribute name tags ranging from simple date and time expressions to more complex ones like coreferences. One example sentence from the MEDIA training corpus would be:
"`je veux une chambre double pour deux personnes`" (I would like a double-bed room for two persons). The sentence is annotated in terms of concepts and the words expressing them as follows:

```
null{je veux (I would like)}
nombre-chambre{une (a)}
chambre-type{chambre double (double-bed
room)}    sejour-nbPersonne{pour deux per-
sonnes (for two persons)}
```

This annotation essentially segments the input sentence into chunks. The attribute-value pairs for the above example that the SLU system is expected to hypothesize by placing the values between brackets are:

```
nombre-chambre[1] chambre-type[double] se-
jour-nbPersonne[2]
```

The MEDIA corpus also includes annotations, called *specifiers*, about certain relations between concept names and values that are semantic structures. Furthermore, other annotations are included to represent the major speech act of a sentence, like assertion, negation and request [12]. They define the so called *mode* of a sentence. These annotations refer to more complex semantic relations than attribute name/value pairs and are not considered in the experiments described in this paper. Not considering them corresponds to operate in the *relaxed simplified* condition defined in the MEDIA project. Within this condition, only two modes have to be distinguished.

B. Polish LUNA Corpus

The data for the Polish corpus are human–human dialogues collected at the Warsaw Transportation call-center [38], [13]. This corpus covers the domain of transportation information like, e.g., transportation routes, itinerary, stops, or fare reductions. Three subsets have been created using the available data: a training set comprising approx. 8 k sentences, a development and an evaluation set containing roughly 2 k sentences each. It is the first SLU database for Polish and from the three corpora presented in this paper the most complex one. The number of different annotated concepts is close to 200, the largest in the three corpora. Furthermore, many concepts are closely related. The SLU task is particularly difficult in this case because Polish is an inflectional language with a relatively free word order as shown by the following examples of different types of inflection for Polish location names:

- (jestem) na Polnej$_{adj,fem,loc}$/Dąbrowskiego$_{adj,masc,gen}$ *(I am) on Polna Street / Dąbrowskiego Street*
- (jadę) z Polnej$_{adj,fem,loc}$ /Dąbrowskiego$_{adj,masc,gen}$ *(I am coming) from Polna Street / Dąbrowskiego Street*
- (jadę) na Polną$_{adj,fem,acc}$ / Dąbrowskiego$_{adj,masc,gen}$ *(I am going) to Polna Street / Dąbrowskiego Street*

In theses phrases, there are three different concepts describing places: LOCATION_STR, SOURCE_STR and GOAL_STR (STR is an abbreviations for street).

C. Italian LUNA Corpus

The application domain of the Italian corpus [14] is software and hardware repairing in the area of an IT help-desk. It consists of human–machine dialogs acquired with a Wizard-of-Oz approach. The data, containing approximately 40 different concepts, are split into training, development and test sets made of respectively 3 k, 400 and 640 sentences.

Given the sentence: "*Buongiorno io ho un problema con la stampante da questa mattina non riesco piu' a stampare*" (Good morning I have a problem with the printer since this morning I cannot print any more), the corresponding semantic annotation is: **null**{Buongiorno io ho} **HardwareProblem.type**{un problema} **Peripheral.type**{con la stampante} **Time.relative**{da questa mattina} **HardwareOperation.negate**{non riesco} **null**{piu'} **HardwareOperation.operationType**{a stampare}.

The corresponding attribute-value annotation is: **HardwareProblem.type** [general_problem] **Peripheral.type** [printer] **Time.relative** [morning] **HardwareOperation.negate** [non] **HardwareOperation.operationType** [to_print].

The semantic annotation is context dependent at turn level, meaning that the same words can be associated with different concepts depending on the object they refer to (for example "*it is not working*" can be "*SoftwareProblem*" or "*HardwareProblem*"). This, together with the very spontaneous form of user turns, makes the task rather complex despite the relatively small number of concepts to be recognized.

V. Single Systems Results

The results for all the systems presented in this section were obtained with the same data for training and testing. Scoring of the hypotheses was done using the NIST scoring toolkit [39]. *Concept Error Rate (CER)* was used as error criterion. It is defined as the percentage obtained with the ratio of the sum of deleted, inserted and confused concepts (not concept *tags*) hypothesized in the test set, and the total number of manually annotated concepts used as reference. The *sentence error rate (SER)* is also used. It is defined as the percentage of sentences whose complete semantic annotation is equal to the one in the corresponding reference. Substitutions, deletions, and insertions are calculated using a Levenshtein-alignment between a hypothesis and a given reference concept sequence. The *NULL* concept, representing out of domain groups, is removed from reference and hypothesis prior scoring.

As a first step, the systems are optimized on the development (DEV) set. Since the choice of feature functions is essential for the performance of log-linear models, the training process of the CRF system will now be shortly described as an example. The basic features have already been introduced in Section II-B1. Since it is not feasible to test all possible combinations of features and window sizes, we stick to the following selection process: first, the regularization term is tuned with a basic feature set consisting of lexical features in a window of

TABLE I
STATISTICS OF THE TRAINING, DEVELOPMENT AND EVALUATION SLU CORPORA AS USED FOR ALL EXPERIMENTS

		training		development		evaluation	
		words	concepts	words	concepts	words	concepts
French	# sentences	**12,908**		**1,259**		**3,005**	
	# tokens	94,466	43,078	10,849	4,705	25,606	11,383
	# NULL tokens	32,580	11,442	4,157	1,372	9,040	2,999
	vocabulary	2,210	99	838	66	1,276	78
	# singletons	798	16	338	4	494	10
	OOV rate [%]	–	–	1.33	0.02	1.39	0.04
Polish	# sentences	**8,341**		**2,053**		**2,081**	
	# tokens	53,418	28,157	13,405	7,160	13,806	7,490
	# NULL tokens	21,973	9,811	5,680	2,384	5,743	2,486
	vocabulary	4,081	195	2,028	157	2,057	159
	# singletons	1,818	19	1,119	23	1,113	28
	OOV rate [%]	–	–	4.95	0.13	4.96	0.11
Italian	# sentences	**3,171**		**387**		**634**	
	# tokens	30,470	14,683	3,764	1,818	6,436	3,057
	# NULL tokens	15,233	5,872	1,893	723	3,287	1,242
	vocabulary	2,386	43	777	39	1,059	39
	# singletons	1,140	0	417	4	537	3
	OOV rate [%]	–	–	4.22	0.06	3.68	0.00

TABLE II
FEATURE BUILD-UP OF THE CRF SYSTEM ON THE FRENCH MEDIA CORPUS INCLUDING THE NUMBER OF ACTIVE FEATURES

features	number of features	CER [%] Train	DEV	EVA
(t_n, w_n)	419,900	82.6	91.6	88.9
+ (t_{n-1}, t_n)	456,190	9.7	15.3	14.6
+ $(t_n, w_{n-1}) + (t_n, w_{n+1})$	1,247,730	3.9	13.1	12.4
+ capitalization, prefixes	1,683,210	3.5	12.8	11.5
+ margin-posterior	1,683,210	10.0	12.3	10.6

TABLE III
FEATURES USED WITH CRFS ON THE VARIOUS CORPORA ("CAP." DENOTES THE CAPITALIZATION FEATURE)

corpus	lexical	cap.	prefix	suffix	# features
French	$w_{n-1}, \ldots, w_{n+1}$	✓	$1 \ldots 4$	-	1,683,210
Polish	$w_{n-1}, \ldots, w_{n+1}$	✓	$1 \ldots 4$	$1 \ldots 4$	6,926,983
Italian	$w_{n-3}, \ldots, w_{n+1}$	-	$1 \ldots 6$	$1 \ldots 6$	1,424,291

$[-1, \ldots, 1]$ around the current word and the bigram feature. Afterwards, enlarged windows for lexical features are tested. With the optimal lexical window w.r.t. CER on the DEV set, the gain of word part features is determined in a similar manner: For pre- and suffix features, the length is successively enlarged and the best performing length for pre- and suffixes is determined. The capitalization feature is simply enabled in one experiment. Finally, the word part features are combined according to their gain. In a last step, the margin-posterior is used for the training of the final CRF system. An exemplary feature buildup for the French MEDIA corpus is presented in Table II. Note that the margin-posterior leads to a better generalization as the training error rates rise while the CER on DEV and EVA decreases.

The optimization process depends on the task and the language. Table III shows the different setups for the final CRF systems for the three tasks presented in this paper.

We produced single best results using all six presented approaches. The results on manual annotations indicated as text and on ASR hypotheses indicated as speech input are given in Table IV. As contrastive results, the table also contains system combination results which will be discussed in detail in Section VI. The numbers in brackets refer to results obtained with a combination of rule-based and statistical attribute value extraction. All other figures are obtained using only rule-based attribute value extraction.

For each system, results for the development (DEV) and the test (EVA) sets have been produced with and without attribute value extraction. The systems are ranked according to their performance in attribute name/value extraction on the evaluation set on text input.

The CRF model leads to the best tagging performance on the MEDIA evaluation corpus with 10.6% CER considering only attribute names. If attribute values are additionally extracted (via a combination of rule-based and stochastic approaches; details are given below), a CER of 12.6% is achieved. Compared to the best result submitted to the MEDIA evaluation campaign in 2005 (19.6% CER, attribute name/value extraction, relaxed-simplified condition, cf. [12]), this is a relative reduction of roughly 35%. A first comparison of SVM, FST and CRF for SLU on French and English corpora has been published in [6] and a detailed comparison of five of the six techniques described in this paper in [40].

Within the latter publication, all methods except DBN have been tuned and applied to a former version of the MEDIA corpus (ASR and manual transcriptions as inputs). The best single system (CRF) performed slightly worse with 16.2% CER (compared to 12.6%). Using ASR input, the respective numbers are 28.9% for the combination of the five systems and 29.8% for the CRF system alone.

Compared to the results presented in [40], improvements in CER have been achieved for all systems, e.g., due to the introduction of categorization as an additional feature for the FST system. The categorization is realized by the use of 18 generalization classes, e.g., numbers, weekdays, country names, hotel names, etc. A detailed error analysis on concept level has shown that four concepts are tagged (slightly) better by competing systems: `object` (e.g., hotel) and `date` by the FST system, `connectProp` by the SVM system and `payment` by the MEMM model.

With a closer look at the different kinds of errors produced by the systems (cf. Table V), we observe an imbalance between

TABLE IV
TAGGING RESULTS ON FRENCH MEDIA, POLISH, AND ITALIAN LUNA. SINGLE SYSTEM AND SYSTEM COMBINATION RESULTS (CER [%]) ON THE MANUALLY (TEXT INPUT) AND AUTOMATICALLY (SPEECH INPUT) TRANSCRIBED DEV AND EVA CORPORA. THE WER FOR SPEECH INPUT FOR FRENCH IS 30.3% ON DEV AND 31.4% ON EVA, FOR POLISH 39.5% ON DEV AND 38.9% ON EVA AND FOR ITALIAN 28.5% ON DEV AND 27.0% ON EVA. NUMBERS IN BRACKETS REFER TO A COMBINATION OF STATISTICAL AND RULE-BASED ATTRIBUTE VALUE EXTRACTION USED FOR THE CRF APPROACH. ALL OTHER FIGURES USE THE SAME RULE-BASED APPROACH

		text input				speech input			
		a. name		a. name & value		a. name		a. name & value	
	model	DEV	EVA	DEV	EVA	DEV	EVA	DEV	EVA
French	CRF	12.3	10.6	15.2 (14.5)	13.5 (12.6)	24.0	23.8	29.0 (28.6)	28.2 (27.3)
	SVM	14.2	13.4	17.2	15.9	27.1	25.8	31.5	29.7
	MEMM	15.8	13.7	18.2	16.3	26.6	26.4	31.4	30.7
	FST	16.1	14.1	18.3	16.6	28.3	27.5	32.5	31.3
	DBN	17.0	15.5	19.3	17.4	29.5	29.1	34.6	32.8
	SMT	16.0	15.1	18.8	17.8	28.4	29.0	33.3	33.5
	weighted ROVER	11.6	**10.2**	13.8 (13.6)	12.0 (**12.0**)	23.4	**23.1**	27.8 (27.5)	27.0 (**26.0**)
	FST Re-ranking	10.7	11.3	13.6	13.3	24.5	24.3	29.1	27.8
Polish	CRF	21.0	21.5	26.4 (24.5)	26.3 (24.7)	53.6	**51.7**	59.7 (59.1)	57.3 (**56.7**)
	FST	20.5	21.9	26.1	27.1	58.3	57.9	65.3	64.0
	MEMM	24.0	25.1	29.1	30.0	58.0	57.0	63.1	61.7
	SVM	26.2	27.3	30.3	31.2	59.1	58.1	63.3	61.5
	DBN	27.5	26.6	33.2	31.4	58.9	57.7	64.8	63.1
	SMT	27.2	27.7	33.6	33.6	60.3	59.0	66.2	64.4
	weighted ROVER	18.7	**18.9**	23.7 (23.2)	24.4 (**23.7**)	53.5	52.9	60.4 (58.6)	58.6 (57.2)
	FST Re-ranking	17.4	19.5	22.6	24.1	57.4	56.5	62.5	61.3
Italian	CRF	20.6	20.0	22.2 (21.7)	22.4 (21.8)	30.0	28.4	33.1 (32.5)	32.1 (**31.3**)
	FST	22.1	20.1	24.2	23.1	35.6	33.3	39.4	37.2
	SVM	24.6	25.3	25.8	27.1	36.3	34.0	39.7	36.7
	DBN	24.3	25.7	26.2	28.9	33.6	32.1	37.2	36.3
	SMT	25.0	25.0	27.4	27.9	35.0	33.7	38.8	37.5
	MEMM	24.6	27.3	26.3	29.3	33.2	33.3	36.9	37.0
	weighted ROVER	19.5	19.8	20.8 (20.3)	21.4 (21.6)	29.3	**27.5**	32.2 (32.3)	31.3 (31.6)
	FST Re-ranking	19.3	**18.3**	21.2	**20.9**	31.3	29.2	34.8	32.6

TABLE V
ATTRIBUTE/VALUE CER FOR THE SIX DESCRIBED SYSTEMS ON THE MEDIA EVALUATION CORPUS (TEXT INPUT). THE CER IS ALSO PRESENTED BROKEN DOWN IN SUBSTITUTION, INSERTION, AND DELETION ERRORS

	attribute/value CER [%]				
model	Sub	Del	Ins	CER	SER [%]
CRF	5.1	4.8	2.8	12.6	21.0
SVM	5.9	6.8	3.2	15.9	24.7
MEMM	6.5	7.0	2.8	16.3	25.4
FST	6.6	4.8	5.1	16.6	25.8
SMT	6.5	6.1	5.3	17.8	26.6
DBN	5.7	6.1	5.6	17.4	26.9

the different kinds of errors across the various systems. This is an indication that system combination may help to reduce the overall error rate.

In any deployed dialog system, a speech recognition system is used to provide the input word sequence for the concept tagging module. Since ASR is always error prone, it is necessary to analyze the effect of ASR errors on the tagging performance. Therefore, we use an automatic transcription of the development and the evaluation corpora. For MEDIA, the ASR word error rate is 30.3% for DEV and 31.4% for EVA. The corresponding tagging results of all six systems are given in Table IV. The performance is measured w.r.t. the attribute name/value sequence for the manually transcribed corpora. Concerning the different kinds of errors produced by the systems, there is roughly the same trend as for the manual transcriptions.

The CER raises by a factor of approx. 1.7–2.3 for speech input compared to text input. An error analysis revealed that for two concepts the tagging performance degenerates heavily due to the introduced recognition errors:

- the concept `answer` is relatively short covering mainly the key words "oui" (yes), "non" (no) and "d'accord" (agreed) which have often been deleted by the ASR system;
- `payment` often corresponds to the currency word "euro" which is also often deleted or confused by non-content words;
- there are also concepts for which the tagging performance is comparatively stable, e.g., `object` which is often found next to a co-reference tag `coRef`.

For the Polish task, the experimental results are also given in Table IV. The overall trend is similar as for the MEDIA task: the CRF model outperforms all other models with a CER of 24.7% for attribute/value extraction on text input. The second best performing system, FST, has a relative loss in performance of 10% w.r.t. CRFs. It seems to tend to over-fitting, since it is much better on the DEV sets than on the EVA sets. Since the Polish task is more complex and there is less training material available, the overall CER is worse than for the MEDIA task. Additionally, the corpus consists of human–human dialogs (annotated by linguists) which are in general more natural and thus complex to learn for statistical approaches.

The results on ASR input are also given in Table IV. Due to the pretty high word error rate (WER) (roughly 40%), even the CRF system gets a CER of 56.7% on the evaluation set considering attribute names and values. These results on ASR input show that the CRF approach is quite robust, since the second best performing system scores 64.0%, which is a relative drop in performance of approx. 13%.

The results for the Italian task are given in the same Table IV. Again, the whole picture is similar to French and Polish. CRFs lead to the best result for text and ASR input (21.8% resp. 31.3%

CER), followed by the other systems with a clear gap of several percents. Another interesting point is the ranking of the systems across languages. CRF seems to be the method of choice, since it always outperforms the five other methods. SMT seems to be the weakest modeling approach. Altogether, the gap between the various models is pretty big: the drop in performance between the best and the weakest model on text input is roughly 38% for French, 36% for Polish, and 28% for Italian. On speech input, the corresponding figures are 20% for French and Italian, and 14% for Polish (note that the error rates for Polish speech input are pretty high in general).

Except for the CRF system, the attribute value extraction is performed in the same way for all systems using a rule-based approach. For CRFs, the procedure has been the following: on the development set, the stochastic and rule-based attribute value extraction is performed in parallel on the reference text input. The errors of both processes are compared and, for each attribute name, the extraction method with less errors is chosen, e.g., for the MEDIA corpus, 16 out of 99 attribute names are covered by rules, namely date and time expressions. For Polish, 94 out of 195 attribute names are covered using rules. Here, the overall confusion is higher due to the high number of attribute names within the corpus. Mostly date/time expressions, bus numbers and locations/places are extracted using rules. For the much smaller Italian task, only 10 out of 43 rules are used, which cover user data like names or surnames, problem types or cardinal numbers. In general, rule-based approaches work better for enumerable types like numbers or for items which can be listed and put into a category like names or places.

A comparison of rule-based and statistical attribute value extraction and their combination on all of the three tasks is given in Table VI for the CRF approach. For all languages, the rule-based approach outperforms the stochastic approach, at least if reference or test input is considered. For speech input, the gap between rules and the statistical approach is pretty small, due to the fact that the rules fail to correctly process erroneous input. For Italian, the statistical approach appears to perform slightly better than the one using rules, even if the advantage is statistically insignificant. For almost all input conditions and tasks, the combination of both approaches gives a gain in performance. Thus, the combination of stochastic approach and rules has been used for the CRF approach for all tasks/languages.

All the presented results show that there is a need for further error reduction. Even if it is difficult to make an assessment without building a real system, it is very likely that any dialogue manager will have difficulties in deciding erroneous inputs. While the best available sequence classifiers have been tested individually, system combination is now conceivable to take the best advantage of them all.

VI. Methods For Dealing With Multiple Hypotheses

In this section, two approaches to combine systems for dealing with multiple hypotheses are described and evaluated. First, the well-known recognizer output voting error reduction (ROVER) is evaluated on the MEDIA corpora. Afterwards, a re-ranking approach combining discriminative and generative methods is presented.

A. ROVER

Motivated by the differences in tagging performance on some individual concepts for the six systems, we performed light-weighted system combination experiments using (weighted) ROVER, which is known to work well for speech recognition [41]. Since we currently only consider the single best output of each system, ROVER performs majority voting after alignment based on the Levenshtein edit distance of the sequences of concept hypotheses generated by all the systems. The reference for the alignment is the most likely sequence according to the CRF system. Additionally, the system weights are optimized on the DEV set using Powell's method (*multi-start*) [42]. The results are presented in Table IV for text and speech input for all three tasks. Using all six systems on the MEDIA corpus, there is a relative gain of approx. 5% for text and speech input on the EVA corpora (considering name and value pairs). We also tried to estimate system weights using the downhill simplex algorithm, but there is no significant difference compared to Powell's method.

It should be noted that ROVER is rather robust as it improves the single-best system in all input conditions and improvements on the DEV corpora always correspond to improvements on the EVA corpora.

For Polish, ROVER gives comparatively good results for text input. The relative improvements over the CRF system are roughly 12% for attribute names only and 4% for attribute name and value extraction on the EVA corpora. Again, also the results on the DEV corpora are better than the single-best system. On speech input, the results on the DEV corpora are slightly better than single-best, but this does not carry over to the EVA sets, presumably due to the overall high error rates. Additionally, the gap between the best and the second best system is also pretty big. In fact, also re-ranking (described in Section VI-B), does not produce a gain over the CRF approach.

ROVER applied to the Italian task only produces statistically insignificant improvements (approx. 1% of relative improvement) on text input. On speech input, the picture is similar to Polish: the CER of the second best system is roughly 20% relatively worse than the CRF system. However, if only attribute name extraction is considered, ROVER leads to a small improvement of approx. 3% relative over the single-best system. If additionally attribute value extraction is performed, the ROVER result is comparable to the CRF result.

ROVER seems to be a good choice for robust system combination, since it is very easy and cheap to compute once the single-system outputs are available and leads to improvements in most cases. For the tasks, where the results are worse than single-best (Polish and Italian ASR input), however the loss in performance is not statistically significant.

To analyze how much gain would be theoretically possible using system combination techniques, we computed the oracle error rates for text inputs (cf. Table VII) for all corpora. Concerning MEDIA, the oracle CER for the name and value condition is roughly half of the system combination result. This indicates that considering all system outputs provides a very high recall that can be exploited by a dialogue manager with potential improvements over the results obtained by just using system weights. For Polish and Italian, the figures are similar. For speech input, the oracle error rates only drop by 20%–30%

TABLE VI
COMPARISON OF RULE-BASED AND STATISTICAL ATTRIBUTE VALUE EXTRACTION AND THEIR COMBINATION FOR THE CRF APPROACH ON ALL TASKS COVERED IN THIS PAPER (CER[%])

		reference input		text input		speech input	
	extraction method	DEV	EVA	DEV	EVA	DEV	EVA
French	rule-based	4.3	4.8	15.2	13.5	29.0	28.2
	statistical	5.3	5.2	16.4	14.0	29.5	28.0
	combination	2.6	3.5	**14.5**	**12.6**	**28.6**	**27.3**
Polish	rule-based	6.8	7.2	26.4	26.3	59.7	57.3
	statistical	13.9	14.3	29.2	29.8	61.8	59.9
	combination	4.8	5.3	**24.5**	**24.7**	**59.1**	**56.7**
Italian	rule-based	3.2	2.9	22.2	22.4	33.1	32.1
	statistical	4.8	4.6	23.0	22.5	**32.4**	**31.1**
	combination	2.1	3.4	**21.7**	**21.8**	32.5	31.3

TABLE VII
ADDITIVE ORACLE ERROR RATES (CER [%]) ON THE MANUALLY TRANSCRIBED (TEXT INPUT) CORPORA FOR THE SIX SYSTEMS ORDERED BY DECREASING PERFORMANCE

		a. name		a. name & value	
	model	DEV	EVA	DEV	EVA
French	CRF	12.3	10.6	14.5	12.6
	+SVM	9.4	7.7	11.1	9.2
	+MEMM	8.6	6.9	10.2	8.3
	+FST	6.8	5.5	10.1	8.3
	+SMT	6.1	4.9	9.1	7.4
	+DBN	5.0	4.3	7.2	6.4
Polish	CRF	21.0	21.5	24.5	24.7
	+FST	13.2	14.2	17.4	18.1
	+MEMM	11.8	12.8	16.0	16.6
	+SVM	10.6	11.6	14.7	15.4
	+DBN	9.4	10.3	13.7	14.2
	+SMT	8.7	9.5	13.0	13.5
Italian	CRF	20.6	20.0	21.7	21.8
	+FST	14.7	12.8	16.2	14.7
	+SVM	12.5	11.4	14.2	13.4
	+DBN	10.9	10.1	12.4	12.2
	+MEMM	10.1	9.5	11.7	11.6
	+SMT	10.1	9.1	11.6	11.2

w.r.t. the single best system. This indicates that all systems have problems with erroneous input and to merely apply system combination techniques is not enough to improve performance.

B. Combination of Discriminative and Generative Algorithms (Re-Ranking)

The models used in the single systems described so far have very different characteristics. In the context of SLU, generative models learn a joint probability of words and concepts in order to estimate probabilities for all the possible events. Discriminative models learn a conditional probability of concepts given words. The former are more robust to over-fitting on training data, the latter can take into account many complex features and use the most relevant to learn word and concept dependencies.

Given so different characteristics, it is expected that integration of generative and discriminative models could bring improvements for the SLU task mixing characteristics of both models. Following this intuition we applied the combination of generative and discriminative models described in [43]: the FST-based model produces the *n*-best interpretation hypotheses ranked by the joint probabilities of the Stochastic Conceptual Language Model (see Section II-E). An SVM model, using particular kernels for text processing, provides an alternative ranking of the *n*-best interpretation list.

Discriminative re-ranking is based on a binary classification model, which, given a pair of hypotheses, detects the *most correct*. The pairs needed for re-ranking are built from the *n*-best interpretation list generated by the FST-based model. At training time, the best hypothesis in the n-best list is selected (computing the edit distance of each hypothesis from the manual reference annotation), positive instances for the classifier are then built comparing the best hypothesis with all the others. Since the model is symmetric, negative instances are built inverting elements in positive ones. This hypothesis organization allows the SVM classifier to learn which annotation in each pair has an error rate lower than the others so that the n-best annotations can be sorted based on their correctness (see [43] for more details). At classification time, all the possible pairs are built from the *n*-best list.

The kernel that we used to evaluate pair similarity in the re-ranking model is the Partial Tree Kernel (PTK) [44] applied to the semantic tree called FEATURES [45]. This captures different important aspects of an SLU hypothesis: concepts annotated by the baseline model, concept segmentation and surface form of the input sentence together with some word features. For the MEDIA corpus, the word categories introduced in the previous section were used as features in the tree. For the Italian corpus, similar generalization features were used, comprising some domain independent categories (e.g., Months, Numbers, Dates, etc.) and syntactic categories for articles, prepositions, adjectives and some adverbs, useful to generalize semantic head prefixes (e.g., "with my printer" becomes "PREP ADJ printer"). For the Polish corpus, no additional features were used, only the surface form was represented in the tree structure.

As shown in [45], starting from pairs of annotated sentences produced by the FST-model, we build pairs of trees. Let us suppose that ten-best interpretations are kept from FST model output, s_i is the ith interpretation for $i \in [1, \ldots, 10]$ and s_j is the best interpretation among them. Positive instances for training are built as pairs $e_k = \langle s_j, s_i \rangle$ for $i \in [1, \ldots, 10]$ with $i \neq j$, whereas the negative instances will be $e_k = \langle s_i, s_j \rangle$. Instances for classification are then built with all possible combinations of the n-best list $e_k = \langle s_m, s_n \rangle$ for $m, n \in [1, \ldots, 10]$ with $m \neq n$. With an abuse of notation, let s_i denote also the tree built from the corresponding interpretation, the pairs of trees built from the n-best list are used to train the re-ranker using the following re-ranking kernel:

$$K_R(e_1, e_2) = \mathrm{PTK}\left(s_1^1, s_2^1\right) + \mathrm{PTK}\left(s_1^2, s_2^2\right) - \mathrm{PTK}\left(s_1^1, s_2^2\right) - \mathrm{PTK}\left(s_1^2, s_2^1\right) \quad (19)$$

where s_k^i is the ith tree of the kth pair e_k and e_1 and e_2 are two pairs in the set of training instances.

This schema, consisting in combining the results of four kernels, was used for the first time for SLU in [43] and refined in [45], but it has been applied before in [46] for semantic role labeling re-ranking, in [47] and [48] for parse re-ranking and in [49] for machine translation.

At classification time, the re-ranking kernel is computed on classification instances to get the score used to re-sort the n-best hypotheses list and to take the new best interpretation.

Results are in general close to the best model, except on Polish speech input, where all models are affected by the high WER of the speech recognizer. Re-ranking is particularly effective on text input, where in most cases the other methods are outperformed. On speech input this approach is penalized by the lack of robustness of the FST-based model. Discriminative models in general show better performances on speech input (see Table IV). This last point is more evident for the Polish task, where the WER of the ASR is particularly high and so no model can go below 50% CER for attribute name extraction, since the re-ranking approach has the lowest improvement with respect to the FST model baseline on speech input: 1.5% and 2.4% relative improvement on Polish DEV and EVA sets, respectively, against 15.8% and 12.6% on MEDIA, 12.1% and 12.3% on Italian.

VII. Conclusion

In this paper, we have presented six state-of-the-art models for concept tagging applied to three tasks of different complexity in different languages. Additionally, comparative results as well as results for system combination methods have been presented. The models have been applied in two conditions: manual transcriptions and automatic transcriptions provided by an ASR system. CRF has turned out to be the best performing single-system on all tasks.

On the well-known MEDIA corpus, a CER of 10.6% resp. 12.6%, if attribute value extraction is considered, could be achieved on EVA using manual transcriptions as input. This corresponds to a relative reduction of approx. 35% w.r.t. literature results. With automatic transcriptions, the comparable figures are 23.8% and 27.3%. Thus, when attribute values are additionally extracted, the CER relatively raises by approx. 17–27%. For Polish and Italian, there are no comparable figures available by other groups yet, since the corpora have been collected only recently, but a CER of 24.7% on EVA for Polish text input, attribute name and value extraction, and 21.8% CER for the comparable figure in Italian seem to be a good start.

Applying ROVER system combination of all six models could further reduce the CER on most tasks. On French MEDIA, a 3%–5% relative improvement could be achieved depending on the input condition. For Polish and Italian, ROVER could outperform the single-best system on text input whereas on speech input the performance is slightly worse. Overall, ROVER seems to be a quite robust approach to system combination.

The re-ranking approach, combining an FST model with an SVM model, shows significant improvements on text input with respect to ROVER, which yields results combining six models. Re-ranking is less robust on speech, but this is mostly due to characteristics of the FST model, that shows less robustness on speech input with respect to discriminative models. This approach can be improved in the future, especially on speech input, by re-ranking 1) hypotheses coming from models more robust to noise (e.g., CRF) and 2) more than ten-best interpretations.

Some general considerations can be made based on the results obtained with the proposed approaches and their comparisons.

Interpretation can be seen as a special form of translation from natural language into a meaning representation language. The best results have been obtained with CRFs, probably because the approach handles in an effective way the context of an entire dialog turn in the input data.

Handcrafted knowledge based on rules can be effectively combined with knowledge acquired with automatic methods. Building a knowledge source of this type requires a considerable effort. Nevertheless, general purpose semantic knowledge properly representing, for example, space and time entities and relations can be reused in many applications. This knowledge can also be used in the functions of exponential models.

Different approaches often produce different types of errors. By pooling the results of many systems, high recall can be obtained at the expense of precision that can probably be increased by imposing additional constraints from a conversation context.

Word errors introduced by ASR systems are particularly high with telephone applications involving real-world users. As expected, they induce a large number of errors in concept hypothesization because they affect semantically relevant words and phrases. Many of them are due to background noise and multiple voices, failure in end-point detection, mispronunciation of words, difficulty in recognizing a large variety of proper names and other causes.

Interpretation errors also appear in manual transcriptions of conversations, especially those between real-world users, primarily because spoken language often does not follow the structure of the written-style text which is used for designing the models.

In spite of all these problems, partial automation is possible for certain types of applications by transferring to a human operator sentences interpreted with low confidence measured, for example, by the posterior probability. Confidence measures should not only depend on the ASR results, but also on the coherence of the interpretations with the conversation history and, in the case of dialogues, with system prompts.

References

[1] R. Jackendoff, *Semantic Structures*. Cambridge, MA: MIT Press, 1990.

[2] R. Pieraccini, E. Levin, and C. H. Lee, "Stochastic representation of conceptual structure in the ATIS task," in *Proc. Speech and Natural Lang. Workshop*, Los Altos, CA, 1991, pp. 121–124.

[3] G. Heigold, P. Lehnen, R. Schlüter, and H. Ney, "On the equivalence of Gaussian and log-linear HMMs," in *Proc. ISCA Interspeech*, Brisbane, Australia, Sept. 2008, iSCA best student paper award.

[4] Y. Rubinstein and T. Hastie, "Discriminative vs informative learning," in *Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining. AAAI Press*, 1997, pp. 49–53.

[5] G. Santafé, J. Lozano, and P. Larranaga, "Discriminative vs. generative learning of bayesian network classifiers," *Lecture Notes in Computer Science*, vol. 4724, p. 453, 2007.

[6] C. Raymond and G. Riccardi, "Generative and discriminative algorithms for spoken language understanding," in *Proc. ISCA Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 1605–1608.

[7] R. De Mori, D. Hakkani-Tur, M. McTear, G. Riccardi, and G. Tur, "Spoken language understanding," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 50–58, May 2008.

[8] S. Seneff, "TINA: A Probabilistic Syntactic Parser for Speech Understanding Systems," in *Proc. IEEE ICASSP*, Glasgow, U.K., 1989, pp. 711–714.

[9] S. Miller, R. Schwartz, R. Bobrow, and R. Ingria, "Statistical language processing using hidden understanding models," in *HLT'94: Proc. Workshop Human Lang. Technol.*, Morristown, NJ, 1994, pp. 278–282, Association for Computational Linguistics.

[10] K. A. Papieni, S. Roukos, and R. T. Ward, "Maximum likelihood and discriminative training of direct translation models," in *Proc. IEEE ICASSP*, Seattle, WA, 1998, pp. 189–192.
[11] N. Camelin, F. Béchet, G. Damnati, and R. De Mori, "Detection and interpretation of opinion expressions in spoken surveys," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 369–381, Mar. 2010.
[12] H. Bonneau-Maynard, C. Ayache, F. Béchet, A. Denis, A. Kuhn, F. Lefèvre, D. Mostefa, M. Qugnard, S. Rosset, J. Servan, and S. Vilaneau, "Results of the french evalda-media evaluation campaign for literal understanding," in *Proc. 5th Int. Conf. Lang. Resources Eval. (LREC)*, Genoa, Italy, May 2006, pp. 2054–2059.
[13] A. Mykowiecka, K. Marasek, M. Marciniak, J. Rabiega-WiÅŻniewska, and R. Gubrynowicz, "Annotated corpus of polish spoken dialogues," in *Human Language Technology. Challenges of the Inf. Soc.. Proc. 3rd Lang. Technol. Conf., LTC 2007, Poznan, Poland, October 5–7, 2007, Revised Selected Papers, LNCS 5603*, 2009, pp. 50–62, Springer.
[14] M. Dinarelli, S. Quarteroni, S. Tonelli, A. Moschitti, and G. Riccardi, "Annotating spoken dialogs: From speech segments to dialog acts and frame semantics," in *Proc. SRSL Workshop EACL*, Athens, Greece, 2009.
[15] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proc. Int. Conf. New Methods Lang. Process.*, Sep. 1994.
[16] T. Deselaers, S. Hasan, O. Bender, and H. Ney, "A deep learning approach to machine transliteration," in *Proc. EACL Workshop Statist. Mach. Translation*, Athens, Greece, Mar. 2009, pp. 233–241.
[17] S. Jiampojamarn and G. Kondrak, "Online discriminative training for grapheme-to-phoneme conversion," in *Proc. ISCA Interspeech*, Brighton, U.K., Sep. 2009, pp. 1303–1306.
[18] L. Ramshaw and M. Marcus, "Text chunking using transformation-based learning," in *Proc. 3rd Workshop Very Large Corpora*, Cambridge, MA, June 1995, pp. 84–94.
[19] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa, "Semantic annotation of the French MEDIA dialog corpus," in *Proc. ISCA Eurospeech*, Lisboa, Portugal, 2005.
[20] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 591–598, Citeseer.
[21] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Eighteenth Int. Conf. Mach. Learn. (ICML)*, Williamstown, MA, Jun. 2001, pp. 282–289.
[22] O. Bender, K. Macherey, F.-J. Och, and H. Ney, "Comparison of alignment templates and maximum entropy models for natural language understanding," in *Proc. Conf. Eur. Chap. Assoc. Comput. Linguist.*, Budapest, Hungary, Apr. 2003, pp. 11–18.
[23] S. Hahn, P. Lehnen, and H. Ney, "System combination for spoken language understanding," in *Proc. ISCA Interspeech*, Brisbane, Australia, Sep. 2008, pp. 236–239.
[24] G. Heigold, R. Schlüter, and H. Ney, "Modified MPE/MMI in a transducer-based framework," in *Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 3749–3752.
[25] M. Jeong and G. G. Lee, "Triangular-chain conditional random fields," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 7, pp. 1287–1302, Sep. 2008.
[26] S. Hahn, P. Lehnen, G. Heigold, and H. Ney, "Optimizing CRFs for SLU tasks in various languages using modified training criteria," in *Proc. ISCA Interspeech*, Brighton, U.K., Sep. 2009, pp. 2727–2730.
[27] A. Mauser, R. Zens, E. Matusov, S. Hasan, and H. Ney, "The RWTH statistical machine translation system for the IWSLT 2006 evaluation," in *Proc. Int. Workshop Spoken Lang. Translation (IWSLT)*, Kyoto, Japan, Nov. 2006, pp. 103–110.
[28] CoNLL-2000, Results of the CoNLL-2000 Shared Task on Chunking. [Online]. Available: http://www.cnts.ua.ac.be/conll2000/chunking/
[29] T. Kudo and Y. Matsumoto, "Chunking with support vector machines," in *Proc. NAACL'01: 2nd Meeting North Amer. Chap. Assoc. Comput. Linguist. Lang. Technol.*, Morristown, NJ, 2001, pp. 1–8, Association for Computational Linguistics.
[30] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Comput., Speech Lang.*, vol. 16, no. 1, pp. 69–88, 2002.
[31] G. Ji and J. Bilmes, "Backoff model training using partially observed data: Application to dialog act tagging," in *Proc. Human Lang. Technol. Conf. NAACL, Main Conf.*, New York, Jun. 2006, pp. 280–287 [Online]. Available: http://www.aclweb.org/anthology/N/N06/N06-1036, Association for Computational Linguistics
[32] F. Lefèvre, "A DBN-based multi-level stochastic spoken language understanding system," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Aruba, Dec. 2006, pp. 82–85.
[33] F. Lefèvre, "Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Honolulu, HI, Apr. 2007, vol. 4, pp. 13–16.
[34] J. A. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *Proc. NAACL'03: Conf. North Amer. Chap. Assoc. Comput. Linguist. Human Lang. Technol.*, Morristown, NJ, 2003, pp. 4–6, Association for Computational Linguistics.
[35] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," Center for Research in Computing Technology, Harvard Univ., Cambridge, MA, Tech. Rep. 10, Aug. 1998.
[36] C. Raymond, F. Béchet, R. De Mori, and G. Damnati, "On the use of finite state transducers for semantic interpretation," *Speech Commun.*, vol. 48, no. 3–4, pp. 288–304, Mar.–Apr. 2006.
[37] C. Servan, C. Raymond, F. Béchet, and P. Nocéra, "Conceptual decoding from word lattices: Application to the spoken dialogue corpus MEDIA," in *Proc. Int. Conf. Spoken Lang. Process.*, Pittsburgh, PA, 2006.
[38] K. Marasek and R. Gubrynowicz, "Design and data collection for spoken polish dialogs database," in *Proc. 6th Int. Conf. Lang. Resources Eval. (LREC)*, Marrakech, Morocco, May 2008.
[39] NIST, Speech Recognition Scoring Toolkit (SCTK). [Online]. Available: http://www.nist.gov/speech/tools/
[40] S. Hahn, P. Lehnen, C. Raymond, and H. Ney, "A comparison of various methods for concept tagging for spoken language understanding," in *Proc. 6th Int. Conf. Lang. Resources Eval. (LREC)*, Marrakech, Morocco, May 2008.
[41] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser Output Voting Error Reduction (ROVER)," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding (ASRU)*, Santa Barbara, CA, Dec. 1997, pp. 347–352.
[42] M. Powell, "A fast algorithm for nonlinearly constrained optimization calculations," in *Proc. Biennial Conf. Nume. Anal.*, G. Watson, Ed., Dundee, U.K., Jun. 1977, pp. 144–157, Lecture Notes in Mathematics, Vol. 630. Berlin, Heidelberg, New York: Springer 1978.
[43] M. Dinarelli, A. Moschitti, and G. Riccardi, "Re-ranking models for spoken language understanding," in *Proc. Conf. Eur. Chap. Assoc. Comput. Linguist. (EACL)*, Athens, Greece, Apr. 2009, pp. 202–210.
[44] A. Moschitti, "Efficient convolution kernels for dependency and constituent syntactic trees," in *Proc. ECML*, Berlin, Germany, 2006, pp. 318–329.
[45] M. Dinarelli, A. Moschitti, and G. Riccardi, "Re-ranking models based on small training data for spoken language understanding," in *Proc. Conf. Empirical Methods for Natural Lang. Process. (EMNLP)*, Singapore, Aug. 2009, pp. 11–18.
[46] A. Moschitti, D. Pighin, and R. Basili, "Semantic role labeling via tree kernel joint inference," in *Proc. CoNLL-X*, New York, 2006.
[47] M. Collins and N. Duffy, "New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron," in *Proc. ACL*, 2002, pp. 263–270.
[48] L. Shen, A. Sarkar, and A. K. Joshi, "Using LTAG based features in parse reranking," in *Proc. EMNLP*, 2003.
[49] L. Shen, A. Sarkar, and F. J. Och, "Discriminative reranking for machine translation," in *Proc. HLT-NAACL*, 2004, pp. 177–184.

Stefan Hahn studied computer science at RWTH Aachen University, Aachen, Germany. He received the Diplom degree in computer science from RWTH Aachen University in 2006. He is currently pursuing the Ph.D. degree at the Computer Science Department, RWTH Aachen University.

He joined the Human Language Technology and Pattern Recognition Group headed by Prof. Dr-Ing. Hermann Ney in 2004. He is currently working at the Computer Science Department, RWTH Aachen University, as Ph.D. Research Assistant. His research interests include automatic speech recognition, log-linear modeling, spoken language understanding, and monotone string-to-string translation.

Marco Dinarelli received the M.Sc. and B.Sc. degrees from University of Pisa, Pisa, Italy, in 2006 and 2003, respectively, and the Ph.D. degree in information and communication technology from International Doctoral School, University of Trento, Trento, Italy, in March 2010, under the supervision of Prof. Ing. Giuseppe Riccardi. The main topic of his Ph.D. dissertation was spoken language understanding for spoken dialog systems, with particular focus on model integration via discriminative reranking.

During the Ph.D. degree, he was involved in the European Project LUNA. His main research interests are automatic speech recognition and understanding, machine learning and kernel methods for natural language processing, data-driven stochastic approaches.

Christian Raymond received the M.S. and Ph.D. degrees in computer science from the University of Avignon, Avignon, France, in 2000 and 2005, respectively.

From 2006 to 2009, he worked on the European-funded LUNA project as a post-doc for one year at the university of Trento, Trento, Italy, and two years at the University of Avignon. He was appointed in September 2009 as an Associate Professor at the "Institut National des Sciences Appliquées" (INSA) in Rennes, France, and joint the TexMex team, devoted to multimedia document analysis, at the IRISA research unit.

Dr. Raymond is a member of the International Speech Communication Association, and he has served on the scientific committees of several international conferences (ICASSP, LREC, EACL). His research activities focus mainly on speech understanding, machine learning for natural language processing, and data driven stochastic approaches.

Fabrice Lefèvre received the degree in electrical engineering from ENSEA-Cergy, Cergy-Pontoise, France, and the Ph.D. degree in computer science from the University Paris VI, Paris, France, in 2000.

He was appointed an Assistant Professor position at the University of Orsay, Paris XI, in 2001 where he worked in the Spoken Language Processing Group, LIMSI-CNRS. He joined the University of Avignon, Avignon, France, in 2005, where he works in the Dialog Team at LIA. He was an Academic Visitor in the Engineering Department, Cambridge University, Cambridge, MA, in 2009. His primary research activities include automatic speech recognition, speech understanding, and spoken dialog systems. He was involved in several European and U.S. projects (CORETEX, AMITIES, EARS, and LUNA). He participated in several international (NIST) and French (AUPELF, Technolangue) spoken language recognition and understanding system evaluation campaigns.

Dr. Lefevre is a member of the International Speech Communication Association, cofounder of the French Spoken Communication Association (AFCP), and a member of its administrative board since 2001.

Patrick Lehnen received the Diploma degree in physics at RWTH Aachen University, Aachen, Germany, in 2007. He is currently pursuing the Ph.D. degree at the Computer Science Department, RWTH Aachen University.

Afterwards, he joined the Human Language Technology and Pattern Recognition Group headed by Prof. Dr-Ing Hermann Ney. He is currently working at the Computer Science Department, RWTH Aachen University as Ph.D. Research Assistant. His research interests include log-linear modeling, spoken language understanding, and monotone string-to-string translation.

Renato De Mori (M'83–SM'89–F'94) received the doctorate degree in electronic engineering from Politecnico di Torino, Turin, Italy.

He has been a Professor and Chairman at the University of Turin and at the School of Computer Science, McGill University, Montreal, QC, Canada, and a Professor at the University of Avignon, Avignon, France. He is currently an Emeritus Professor at McGill University and at the University of Avignon. He is the author or editor of four books and has published more than 100 scientific papers in many international journals. His major contributions have been in the area of automatic speech recognition and understanding, signal processing, computer arithmetic, software engineering and human–machine interfaces. He has been a member of the Executive Advisory Board at the IBM Toronto Lab, Scientific Advisor at France Télécom R&D, Chairman of the Computer and Information Systems Committee, Natural Sciences and Engineering Council of Canada, and Vice-President R&D, Centre de Recherche en Informatique de Montréal.

Prof. De Mori is a Fellow of the Computer Society and has been distinguished lecturer of the IEEE Signal Processing Society. He is an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, has been Chief Editor of *Speech Communication* (2003–2005) (member of the editorial board since 1982), and an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (1998–1992). He is in the Editorial board of *Computer Speech and Language* (since 1988). He has been in the Editorial board of *Computational Intelligence* (1990–2002), *Pattern Recognition Letters* (1980–2004), and *Signal Processing* (1979–1989). He has been a member of the IEEE Speech Technical Committee (1984–1987, 2003–2006), the Interdisciplinary Board, Canadian Foundation for Innovation, and Interdisciplinary Committee for Canadian chairs. He has been involved in many Canadian and European projects and has been scientific leader of the LUNA European project (2006–2009).

Alessandro Moschitti received the Ph.D. degree in computer science from the University of Rome "Tor Vergata," Rome, Italy, in 2003.

He is an Assistant Professor on the Computer Science Department, University of Trento, Trento, Italy, and he worked for two years as an Associate Researcher at the University of Texas at Dallas (between 2002 and 2004). His expertise concerns machine learning approaches to natural language processing, information retrieval and data mining. He has devised innovative kernels within support vector and other kernel-based machines for advanced syntactic/semantic processing documented by more than a 100 scientific articles. These have been published in the major conferences of several research communities, e.g., ACL, ICML, ECML-PKDD, CIKM, ECIR, ICDM, for which he is also an active PC member.

Dr. Moschitti has been member of the guest editorial board of the Computational Linguistics Journal and guest editor of JNLE and reviewer for several journals of different research communities: CLJ, JNLE, JMLR, MLJ, AIJ, AIJR, IEEE-PAMI, IEEE-TKEJ, ACL-TALIP, and IEEE-IIS. He has been chair/organizer of ACL workshops and Special Tracks and he is currently an area chair of ACL 2011. He has participated in six projects of the European Community (EC) and in three U.S. projects: MTBF with Con-Edison, IQAS for ARDA AQUAINT PROGRAM and Deep QA (the Jeopardy! challenge) with IBM of NY. He is currently the project coordinator of the EC Coordinate Action, EternalS, and the local coordinator of the Italian Project PARLI. He has received several best paper and other awards, e.g., the IBM Faculty award.

Hermann Ney (M'86–SM'07–F'11) is a Full Professor of computer science at RWTH Aachen University, Aachen, Germany. Prior to that, he headed the Speech Recognition Group at Philips Research. His main research interests lie in the area of statistical methods for pattern recognition and human language technology and their specific applications to speech recognition, machine translation, and image object recognition. In particular, he has worked on dynamic programming for continuous speech recognition, language modeling, and phrase-based approaches to ma-

chine translation. He has authored and coauthored more than 350 papers in journals, books, conferences, and workshops.

Prof. Ney was a member of the Speech Technical Committee of the IEEE Signal Processing Society from 1997 to 2000. In 2006, he was the recipient of the Technical Achievement Award of the IEEE Signal Processing Society.

Giuseppe Riccardi (M'96–SM'04–F'10) received the Laurea degree in electrical engineering from the University of Padua, Padua, Italy, the M.S. degree in information technology from CEFRIEL Research Center, Milan, Italy, and the Ph.D. degree in electrical engineering from the Department of Electrical Engineering, University of Padua.

From 1990 to 1993, he worked with Alcatel-Telettra Research Laboratories on designing algorithms for audio and speech coding. In 1993, he joined AT&T Bell Laboratories and then AT&T Labs-Research where he worked in the Speech and Language Processing Lab. In 2005, he joined the faculty of Engineering at the University of Trento, Trento, Italy, where he is affiliated with the Electrical Engineering and Computer Science Department and Center for Mind/Brain Sciences. He is the founder and head of the Signals and Interactive Systems Lab. His research on stochastic models for speech and language processing has been applied to a wide range of speech and language tasks. He and his colleagues designed the state-of-the-art AT&T spoken language system ranked first in the 1994 DARPA ATIS evaluation. He co-pioneered the speech and language research in spontaneous speech for the "How May I Help You?" research program which led to breakthrough speech services. His research on learning finite state automata and transducers has lead to the creation of the first large-scale finite state chain decoding for machine translation (Anuvaad). He has coauthored more than 100 papers and 30 patents in the field of speech processing, spoken interfaces, speech understanding, machine learning, and machine translation. His current research interests are language modeling and acquisition, language understanding, spoken/multimodal dialog, affective interfaces, machine learning, and machine translation.

Prof. Riccardi has been on the scientific committee of EUROSPEECH, INTERSPEECH, ICASSP, NAACL, ACL, and EACL. He has co-organized the IEEE ASRU Workshop in 1993, 1999, 2001 and General Chair in 2009. He has been the Guest Editor of the IEEE Special Issue on Speech-to-Speech Machine Translation. He has been founder and Editorial Board member of the ACM Transactions of Speech and Language. He has been elected member of the IEEE SPS Speech Technical Committee (2005–2008). He is a member of ACL, ISCA, and ACM. He has received many national and international awards and more recently the Marie Curie Research Excellence grant by the European Commission, the IEEE SPS Best Paper Award (2009), and the IBM Faculty award (2010).