ON THE EQUIVALENCE OF GAUSSIAN AND LOG-LINEAR HMMS

Georg Heigold, Patrick Lehnen, Ralf Schlüter, Hermann Ney

RWTH Aachen University
Lehrstuhl für Informatik 6 – Computer Science Department
D-52056 Aachen, Germany
{heigold,lehnen,schlueter,ney}@cs.rwth-aachen.de

Abstract

The acoustic models of conventional state-of-the-art speech recognition systems use generative Gaussian HMMs. In the past few years, discriminative models like for example Conditional Random Fields (CRFs) have been proposed to refine the acoustic models. CRFs directly model the class posteriors, the quantities of interest in recognition. CRFs are undirected models, and CRFs do not assume local normalization constraints as HMMs do. This paper addresses the issue to what extent such less restricted models add flexiblity to the model compared with the generative counterpart. This work extends our previous work in that it provides the technical details used for showing the equivalence of Gaussian and log-linear HMMs. The correctness of the proposed equivalence transformation for conditional probabilities is demonstrated on a simple concept tagging task.

Index Terms: Gaussian HMMs, language models, CRFs

1. Introduction

Conditional Random Fields (CRFs) [1, 2] and Hidden CRFs (HCRFs) [3] are discriminative approaches which were introduced in pattern recognition only a few years ago. These discriminative models are considered to be superior to the conventional Gaussian mixture models (GMMs) and Gaussian HMMs (GHMMs) because of the more direct and flexible modeling of CRFs [1, 4, 5] and HCRFs [3]. Some authors have reported on experimental results supporting this claim [4, 3]. Only little work, however, has been done so far to compare HCRFs and GHMMs on a theoretical level. It has been shown, for instance, that (discriminatively estimated) GMMs and log-linear mixture models (LMMs) are equivalent from a functional point of view [6]. Here, LMMs refer to mixture models with log-linear rather than Gaussian components. In [6], it was also claimed that this result extends to HMMs and language models. Here, we provide the technical details to prove this claim.

Assuming a class of log-linear HCRFs $p_{\text{CRF},\Lambda}(c|x)$ parameterized with Λ and a class of generative distributions $p_{\text{Gen},\theta}(x,c)$ parameterized with θ, we use the following notion of equivalence.

Definition 1 *The two probabilistic models $p_{CRF,\Lambda}(c|x)$ and $p_{Gen,\theta}(c|x)$ induced by $p_{Gen,\theta}(x,c)$ via Bayes rule, are called* equivalent *if for each Λ, there exists some θ such that $p_{CRF,\Lambda}(c|x) = p_{Gen,\theta}(c|x)$ $(\forall c, x)$, and vice versa.*

A direct consequence of such an equivalence is that all posterior-based algorithms would be equivalent, e.g. decoding or discriminative training (MMI, MCE, MPE). Note that it is not difficult to establish equivalence in the general case where there are no (or only little) restrictions to the models. It is, however, not always obvious how additional parameter constraints like for example the local normalization constraints of HMMs can be imposed to the model without changing the induced class posteriors. Keep in mind that the equivalence requires the transformation in either direction, i.e., from the discriminative to the generative model and from the generative to the discriminative model. The latter transformation is well-known and rather straightforward [3, 6]. For this reason, we shall focus on the opposite direction.

In contrast to [6] where the basics can be found, this work focuses on the problem and the technical details of transforming CRF parameters associated with conditional probabilities. Besides the usual constraints of probabilities

$$p(c|c') \geq 0, \sum_c p(c|c') = 1, \ \forall c' \qquad \text{(normalization)}, \quad (1)$$

conditional probabilities (cf. Markov models) often have a couple of additional restrictions on the structure

$$p(c_n|c_1^{n-1}) \equiv p(c_n|c_{n-1}), \ \forall n \qquad \text{(dependence)} \quad (2)$$

$$p_m(c_{m+n}|c_{m+n-1}) \equiv p(c_n|c_{n-1}), \ \forall m \geq 0, n \qquad \text{(stationarity)}. \quad (3)$$

The basic idea consists of first normalizing the joint probabilities $p(x,c)$, which define the complete probabilistic model. Conditional probabilities, for instance, can be derived from the joint probabilities by marginalization and applying Bayes rule. The normalization constant cancels in the resulting ratios of pseudo probabilities, e.g. $p_{\text{CRF}}(c|x)$ without normalization constant. So, the tricky thing about normalizing conditional probabilities is to make sure that the structure (e.g. dependence and stationarity assumptions) is preserved.

The remainder of this paper is organized as follows. In Sec. 2, we discuss the conditional probabilities by means of a simple tagging problem using a bigram model. This result is used to show the equivalence of GHMMs and LHMMs in Automatic Speech Recognition (ASR) in Sec. 3. The theoretical result is experimentally verified in Sec. 4 on a simple real-world task.

2. Tagging Problem

The construction of conditional probabilities from a log-linear CRF model is illustrated by means of a simple, yet non-trivial model: concept tagging with a bigram model. Unlike ASR, the tagging problem assumes a one-to-one mapping from the words x_1^N to the concepts c_1^N, i.e., the alignment problem is deferred until Sec. 3. For the time being, consider the CRF (without normalization constant)

$$p_{\text{CRF}}(c_1^N|x_1^N) \propto \exp(\alpha(c_N, \$)) \prod_{n=1}^{N} \exp(\alpha(c_{n-1}, c_n) + \beta(c_n, x_n)) \quad (4)$$

In addition to the regular concepts $c \in \Sigma$, we use the special concept \$ indicating the sentence end. Assume that this boundary concept is also part of the bigram model and that the sequences c_1^N start and end with this boundary concept, i.e., $c_0 = c_{N+1} = \$$. This model serves as preparation for the transition and language models in ASR, which typically include such information (entry/exit states for HMMs, sentence boundary symbol for language models), see Sec. 3.

As outlined in the introduction, the goal is to find a generative model $p_{\text{Gen}}(x_1^N, c_1^N)$ that is equivalent to the CRF in Eq. (4) in the sense of Def. 1. The generative probabilities can be decomposed into the emission probabilities $p(x|c)$ associated with $\beta(c, x)$ and the bigram probabilities $p(c|c')$ associated with $\alpha(c', c)$

$$p_{\text{Gen}}(x_1^N, c_1^N) \quad = \quad p(\$|c_N) \prod_{n=1}^{N} p(c_n|c_{n-1}) p(x_n|c_n). \tag{5}$$

It is assumed that the emisssion probabilities are subject to the constraints in Eq. (1) and the bigram probabilities $p(c|c')$ are subject to the constraints in Eq. (1-3).

The pseudo emission probabilities $\exp(\beta(c, x))$ can be normalized positionwise

$$p(x|c) \quad = \quad \frac{\exp(\beta(c, x))}{Z(c)}. \tag{6}$$

The normalization constant $Z(c) = \sum_x \exp(\beta(c, x))$ carries over to the bigram parameters, i.e.,

$$\begin{aligned} \alpha(c', c) + \beta(c, x) &= (\alpha(c', c) + \log Z(c)) + (\beta(c, x) - \log Z(c)) \\ &= \qquad \tilde{\alpha}(c', c) \qquad + \qquad \tilde{\beta}(c, x) \end{aligned}$$

with $\tilde{\alpha}(c', c) = \alpha(c', c) + \log Z(c)$ and $\tilde{\beta}(c, x) = \beta(c, x) - \log Z(c)$ such that the CRF remains unchanged. The normalization of the bigram probabilities is based on these modified pseudo probabilities, $\exp(\tilde{\alpha}(c', c))$ and $\exp(\tilde{\beta}(c, x))$.

The bigram probabilities can be constructed in a similar way as in [7]. To avoid lengthy calculations, we state the solution and verify that this solution satisfies the properties in Eq. (1-3). In contrast to [7], we do not only assume that a solution exists but also provide an existence proof. Furthermore, our result does not only apply in the limit of infinite sequences as in [7] but it also applies to finite sequences.

The result is based on the matrix notation of the bigram probabilities. The transition matrix Q is defined to hold the unnormalized bigram probabilities, i.e., $Q = [\exp(\tilde{\alpha}(c', c))]$ where $c' \in \Sigma \cup \{\$\}$ and $c \in \Sigma \cup \{\$\}$ denote the previous and the current concepts, respectively. Our "guess" is

$$p(c|c') \quad = \quad \frac{Q_{c'c} v_c}{\lambda v_{c'}}. \tag{7}$$

As will become clear in the proof of the next lemma, λ is the largest eigenvalue of the transition matrix Q and v is the right eigenvector of Q associated with λ, v_c are the components of v.

Lemma 1 *Assume the emission probabilities $p(x|c)$ from Eq. (6) and the bigram probabilities $p(c|c')$ from Eq. (7). Then, the CRF $p_{CRF}(c_1^N|x_1^N)$ in Eq. (4) and the generative model $p_{Gen}(x_1^N, c_1^N)$ in Eq. (5) with these emission and bigram probabilities are equivalent in the sense of Def. 1 subject to the constraints in Eq. (1-3).*

Proof First, the equivalence of the two models defined in Eq. (4) and Eq. (5) can be verified by plugging the definitions for $p(x|c)$ in Eq. (6) and $p(c|c')$ in Eq. (7) into Eq. (5)

$$p_{\text{Gen}}(x_1^N, c_1^N) = \frac{1}{\lambda^{N+1}} \prod_{n=1}^{N+1} \frac{v_{c_n}}{v_{c_{n-1}}} \times Q_{c_N \$} \prod_{n=1}^{N} Q_{c_{n-1} c_n} \exp(\tilde{\beta}(c_n, x_n)). \tag{8}$$

These joint probabilities induce the posteriors $p_{\text{Gen}}(c_1^N|x_1^N) = \frac{p_{\text{Gen}}(x_1^N, c_1^N)}{\sum_{\tilde{c}_1^N} p_{\text{Gen}}(x_1^N, \tilde{c}_1^N)}$. So, the term $\frac{1}{\lambda^{N+1}}$ cancels because it appears both in the numerator and the denominator of the posterior. The (telescope) product over $\frac{v_c}{v_{c'}}$ is 1 by our model assumption that $v_{c_0} = v_{c_{N+1}} = \$$. Hence, equivalence holds.

Second, we check that $p(c|c')$ is well-defined and satisfies the properties in Eq. (1-3). The properties in Eq. (2-3) are satisfied by definition. All coefficients of the transition matrix Q are positive. According to the *Perron-Frobenius Theorem* [8, p.473], the largest eigenvalue λ of Q is positive and unique. Moreover, the eigenvector v associated with λ has only positive coefficients. Hence, the bigram probabilities in Eq. (7) are non-singular (no division by zero) and positive. These quantities are normalized because v is an eigenvector of Q, i.e., $\sum_c Q_{c'c} v_c = \lambda v_{c'}$, $\forall c'$. This identity is equivalent to the normalization constraint in Eq. (1). ■

In general, the transition matrix describes the transition probabilities between two states where the states encode the contexts. In the case of m-gram models, the contexts consist of the previous $m - 1$ words. For a vocabulary of size C, the transition matrix is approximately a $C^{m-1} \times C^{m-1}$ matrix. In particular, higher order m-gram models can be tackled in the same way as bigram models. This is in contrast to the belief in [7] that higher order m-gram models require tensors of rank more than two which would go beyond the standard matrix formalism.

This is the key result used in the next section where the equivalence of GHMMs and LHMMs in ASR is proven.

3. Automatic Speech Recognition (ASR)

Assume a feature vector x_1^T, HMM state sequences s_1^T, and W which denotes either a single word or a word sequence. Consider the log-linear HCRF with the state sequences as the hidden variables

$$\begin{aligned} &p(W|x_1^T) \propto \exp(\alpha(W)) \times \\ &\qquad \sum_{s_1^T} \prod_{t=1}^{T} \exp(\alpha(s_{t-1}, s_t, W)) \ \exp(\lambda(s_t, W)^\top x_t + \alpha(s_t, W)). \end{aligned}$$

The goal of this section is to construct the generative models, i.e., the emission probabilities $p(x|s, W)$, the transition probabilities $p(s|s', W)$, and the language model probabilities $p(W)$, corresponding to the HCRF parameters $\lambda(s, W), \alpha(s, W)$, $\alpha(s', s, W)$, and $\alpha(W)$, respectively. This is accomplished in different steps. First, equivalence is shown for the simpler task of single word recognition (HMMs). Then, this result is extended to the task of continuous speech recognition where a non-trivial language model is considered in addition. Finally, the effect of scaling factors like for example the language model scale is discussed.

Emission probabilities. The reader refers to [6] for an in-depth discussion concerning the equivalence of LMMs and GMMs. The same approach can be used here. Similar to Sec. 2, modified $\tilde{\alpha}(s', s, W)$ compensate for the additional normalization factors in the emission probabilities.

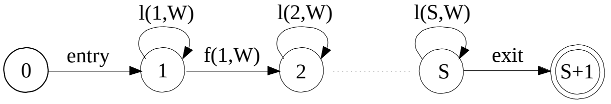

Figure 1: *Finite state automaton representing the valid state sequences for word-based transition probabilities, only loop* $l(s, W)$ *and forward* $f(s, W)$ *transitions,* 0 *and* $S + 1$ *are the entry and exit states.*

Word-based transition probabilities. The left-right topology of the transition probabilities leads to an upper triangular band matrix, see Fig. 1. In contrast to the bigram matrix Q in Sec. 2, this transition matrix is not strictly positive and is reducible (a state cannot be reached by one of its subsequent states). Hence, the algorithm in Sec. 2 is not guaranteed to work. Instead, we employ the general approach via the marginalization of the pseudo joint probabilities as described in the introduction. To guarantee convergence in the marginalization step (summation over state sequences), the loop transitions (the only cycles in the automaton) need to have costs less than 1. Observe that this is no real restriction because all sequences under consideration have the same length and thus, all transition probabilities can be multiplied by the same constant factor.

If only loop and forward transitions with pseudo probabilities $l(s, W) = \exp(\tilde{\alpha}(s, s, W))$, $f(s, W) = \exp(\tilde{\alpha}(s, s + 1, W))$, $f(0, W) = \mathit{entry}$, and $f(S, W) = \mathit{exit}$ are allowed (cf. Fig. 1), the transition probabilities can be calculated explicitly from the partial sums $Z_s(W) := \sum_{T>0} \sum_{s_1^T : s_1 = s, s_T = S+1} \prod_{t=1}^{T} p(s_t | s_{t-1}, W)$ which is the sum over all sequences of different length starting with state s and ending in the final state $S + 1$, i.e., $Z_S(W) = \mathit{exit}, Z_s(W) = \frac{f(s,W) Z_{s+1}(W)}{1-l(s,W)}, Z_0(W) = \mathit{entry} \cdot Z_1(W)$, for $s = S - 1, \ldots, 1$ and for all W. The factor $\frac{1}{1-l(s,W)}$ arises from the infinite sum accounting for the loop transitions and can be interpreted as geometric series. Applying Bayes rule to the marginalized probabilities based on these partial sums, results in the transition probabilities $p(s|s, W) = l(s, W)$, $p(s+1|s, W) = 1 - l(s, W)$, and $p(1|0, W) = p(S|S - 1, W) = 1$. The transition probabilities do not depend on $f(s, W)$ because the contribution of the forward transitions is the same for all state sequences and is absorbed by the language model, $\tilde{\alpha}(W) = \alpha(W) + \log Z_0(W)$. The same approach can be used for more complex topologies (e.g. skips). In general, however, the solution is not so simple and compact.

This word-based transition model is reasonable for phone recognition [3] whereas in ASR, a phone-based transition model is preferred. Such models are discussed in the next paragraph.

Phone-based transition probabilities. In ASR, typically (allo)phone-based rather than word-based transition probabilities are used. These transition probabilities can be represented by a finite state automaton which is similar to that in Fig. 1. The main differences are that now there is a sub-automaton for each phone ϕ and that the exit arcs go back to the initial state. This model can be described by a non-negative and irreducible transition matrix Q. Hence, the approach of Sec.2 applies, using the extended *Perron Frobenius Theorem* for non-negative matrices [8, p.475]. The normalization constant does not depend on W and thus, cancels. The equivalence holds as long as the state tying (e.g. CART) for the transition probabilities is no coarser than the state tying used for the emission probabilities. Otherwise (and as long as the tying scheme is reasonable), the equivalence holds only approximately.

M-gram language model. In continuous speech recognition, W stands for the word sequence $w_1^N = (w_1, \ldots)$. Instead of the simple priors $p(W)$, an m-gram language model is typically used, adding some additional structure. Note that the factor $\frac{1}{\lambda^{N+1}}$ in Eq. (8) does not cancel in ASR because of variable N. However, this is not an issue for $\lambda = 1$. An eigenvalue of 1 is achieved by taking advantage of the ambiguity of the discriminative formulation, or more precisely, by choosing a suitable constant offset for all $\alpha(s', s, w)$ subject to the constraint that the maximum loop pseudo probability is less than 1, e.g. $l_{max} = \max_{s,\phi}\{\exp(\tilde{\alpha}(s, s, \phi))\}$ (see paragraph on phone-based transition probabilities). This is always possible because $\lambda : (0, 1) \mapsto (0, \infty)$ is a continuous function of l_{max} and thus, according to the *Intermediate Value Theorem*, some l_{max} (or more precisely an offset) exists such that $\lambda(l_{max}) = 1$. Then, the algorithm from Sec. 2 applies to the transformation matrix Q induced by $\tilde{\alpha}(w) = \alpha(w) + \log \zeta(w)$ where $\zeta(w)$ accounts for the normalization of the transition probabilities. The equivalence extends to across word models under the common assumption that the (unique) final HMM state of a word cannot be skipped. Note that due to the added dependence of across word models, equivalence requires at last a bigram language model.

Scaling factors. Typically, the different submodels are scaled independently. Obviously, these scaling factors do not break the equivalence of GHMMs and LHMMs. Moreover and unlike for the maximum likelihood estimation, these additional scaling factors do not add any flexibility in the log-linear approach because the log-linear parameters can be redefined to include these scaling factors. Transforming the log-linear model back results in an equivalent GHMM *without* scaling factors. Keep in mind that these scaling factors might indirectly have impact on the results in practice because of spurious local optima of HCRFs. However, these rather heuristic parameters are redundant in the discriminative framework, i.e., they do not need to be tuned or justified.

4. Experimental Verification

In this section, we check the correctness of the theoretical results experimentally. Different testing scenarios are reasonable. An equivalent CRF/generative pair can be optimized separately and then, the performance of the two classifiers can be compared. This was done in the past, e.g. [3, 6]. This indirect approach has the disadvantage that the resulting performance of the two classifiers usually differs in practice. This might be due to numerical issues, local optima etc. For this reason, we decided to pursue a more direct approach to avoid unwanted effects. Here, we estimate a CRF, transform the CRF into an equivalent generative model, and show that this generative model produces the same posteriors and decisions as the CRF. For the experiment, we used the CRF in Eq. (4) which shall serve as a prototype for conditional probabilities. With this choice, the computational complexity can be kept low (cf. "M-gram language model" in Sec. 3, for instance) while avoiding artificial data.

Semantic concept tagging is a comparatively straightforward application domain of CRFs [9]. It is usually defined as the extraction of a sequence of concepts out of a given word sequence. A concept represents the smallest unit of meaning that is relevant for a specific task. A concept may contain various information, e.g the attribute name or the corresponding value. An example from the French Media corpus [10] is given by

$$\underbrace{\texttt{...au sept avril}}_{\texttt{temps-date[07/04]}}\ \underbrace{\texttt{dans cet hôtel...}}_{\texttt{objetBB[hotel]}}$$

Table 1: *Concept Error Rate (CER) for different setups on the Media Corpus eval set (*not *used for verification of equivalence).*

Setup	Simple w/o $	Simple w/ $	Standard w/o $
CER [%]	15.0	14.7	11.5

where the attribute values are written in square brackets behind the attribute name.

For our experiment we used the French Media corpus [10], which deals with hotel reservation and tourist information. We only used the attribute name. All additional information given by the Media corpus was omitted. We tag an attribute name for every source word to get a one-to-one alignment and use a prefix "start_" indicating a new sentence. The feature functions of the CRF use lexical features considering the current word only and transition features similar to a concept bigram model as in Eq.(4). This CRF was estimated on the training part of the Media corpus. This corpus consists of 13k sentences, 94k running word, and 43k running concept tags. The vocabulary comprises 2,210 words, and 146 concepts tags. The resulting CRF was transformed according to the rules in Sec. 2 into an equivalent generative model as given in Eq. (5). The tagging of the training corpus using this generative model leads to exactly the same number of errors as using the original CRF, 9.3% concept error rate. The (differences of the) logarithmic probabilities of both models are illustrated in Fig. 2. They can be considered identical within the numerical precision as the large peak at zero in Fig. 2 clearly shows.

Tab. 1 provides a few error rates on the Media Corpus task to give the interested reader an idea of the relative importance of the different feature functions. Like for speech recognition, the additional boundary symbol $ has some effect. Our best standard setup differs from the simple setup mainly by using lexical features that consider the previous and subsequent words in addition to the current word. As already mentioned, the corpus does not fully comply with the Media evaluation guidelines but fits well for a comparison of the systems.

5. Conclusions

We have shown that the common GHMMs and LHMMs used in ASR are equivalent. This result is surprising and counter-intuitive because this means that the parameter constraints of GHMMs (e.g. local HMM normalization constraints, positivity of variances) do not restrict the model flexibility compared with the related LHMMs. This equivalence does not necessarily imply identical performance of GHMMs and LHMMs in practice. Potential differences might be due to numerical issues (e.g. inversion of covariance matrices for GHMMs), local optima (HCRFs do not guarantee a global optimum), or different optimization criteria (e.g. generative vs. discriminative estimation). Keep in mind that in general, it is essential to consider the complete optimization problem and not only parts of it (e.g. not only the acoustic model) to establish exact equivalence of GHMMs and LHMMs. This equivalence might be useful for the refined analysis and design of algorithms: for instance, why is it hard to outperform state-of-the-art conventional GHMMs with conceptionally much more refined approaches? Finally, the correctness of the presented equivalence transformation for conditional probabilities has been verified experimentally on a simple real-world task.

6. Acknowledgements

This material is partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001-06-C-0023, and was partly funded by the European Union under the integrated project LUNA (FP6-033549). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA.

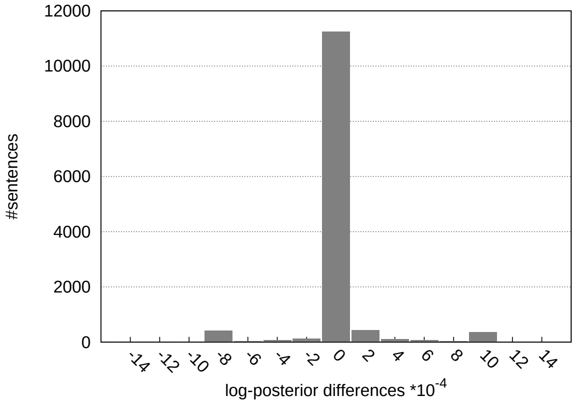

Figure 2: *Distribution of log-posterior differences, zero difference means that the two log-posteriors are identical.*

7. References

[1] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Cong. Machine Learning*, San Francisco, CA, 2001.

[2] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*. MIT Press, 2006.

[3] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *Proc. of the Int. Conf. on Spoken Language Processing (Interspeech)*, Lisbon, Portugal, Sept. 2005.

[4] L. Saul and D. Lee, "Multiplicative updates for classification by mixture models," in *Advances in Neural and Information Processing Systems*, T.G. Dietterich, S. Becker, and Z. Ghahramani, Ed. MIT Press, 2002.

[5] T. Cohn, *Scaling conditional random fields for natural language processing*, Ph.D. thesis, Department of Computer Science and Software Engineering, University of Melbourne, 2007.

[6] G. Heigold, R. Schlüter, and H. Ney, "On the equivalence of Gaussian HMM and Gaussian HMM-like hidden conditional random fields," in *Proc. of the Int. Conf. on Spoken Language Processing (Interspeech)*, Antwerp, Belgium, Aug. 2007.

[7] E.T. Jaynes, *Probability theory: the logic of science*, Cambridge, 2003.

[8] C.R. Rao and M.B. Rao, *Matrix algebra and its applications to statistics and econometrics*, Word Scientific, 1998.

[9] S. Hahn, P. Lehnen, C. Raymond, and H. Ney, "A comparison of various methods for concept tagging for spoken language understanding," in *Proc. of the Sixth Int. Conf. on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 2008.

[10] L. Devillers, H. Maynard, S. Rosset, et al., "The French Media/Evalda project: the evaluation of the understanding capability of spoken language dialog systems," in *Proc. of the Fourth Int. Conf. on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, May 2004.