A Comparison of Various Methods for Concept Tagging for Spoken Language Understanding

Stefan Hahn*, **Patrick Lehnen***, **Christian Raymond**[†], **Hermann Ney***

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

[†] LIA/CNRS - University of Avignon
BP1228 84911 Avignon cedex 09, France
christian.raymond@univ-avignon.fr

Abstract

The extraction of flat concepts out of a given word sequence is usually one of the first steps in building a spoken language understanding (SLU) or dialogue system. This paper explores five different modelling approaches for this task and presents results on a French state-of-the-art corpus, MEDIA. Additionally, two log-linear modelling approaches could be further improved by adding morphologic knowledge. This paper goes beyond what has been reported in the literature, e.g. in (Raymond & Riccardi 07). We applied the models on the same training and testing data and used the NIST scoring toolkit to evaluate the experimental results to ensure identical conditions for each of the experiments and the comparability of the results. Using a model based on conditional random fields, we achieve a concept error rate of 11.8% on the MEDIA evaluation corpus.

1. Introduction

The task of concept tagging 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, like the attribute name or the corresponding value. An example from the MEDIA corpus can be represented as:

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

where the attribute values are written in square brackets behind the attribute name. Within this paper we distinguish between two tasks, the extraction of just the attribute name and the extraction of the attribute name and the corresponding attribute value. In the following section, the various methods which are explored in this paper are shortly described. Section 3. introduces the morphologic features which led to an improved performance for the log-linear models. After the presentation of the training and testing data in Section 4., the experimental results are presented in Section 5. A summary is given in Section 6. The paper concludes with an outlook in Section 7.

2. Methods

2.1. 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 (abbreviated with *log-pos*) and the second one on sentence level (conditional random fields, abbreviated with *CRF*). The general representation of these models is described in equation 1 as a conditional probability of a concept sequence $c_1^N = c_1, \ldots, c_N$ given a word sequence $w_1^N = w_1, \ldots, w_N$:

$$p(c_1^N|w_1^N) = \frac{1}{Z} \prod_{n=1}^{N} exp \left(\sum_{m=1}^{M} \lambda_m \cdot h_m(c_{n-1}, c_n, w_{n-2}^{n+2}) \right).$$
(1)

The log-linear models are based on feature functions $h_m(c_{n-1}, c_n, w_{n-2}^{n+2})$ 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 section 2.1.2. and section 2.1.3. respectively for each model.

2.1.1. Feature Functions

In our experiments we use binary feature functions $h_m(c_{n-1}, c_n, w_{n-2}^{n+2})$, i.e. they either return the value "0" or "1". If a pre-defined combination of the values $c_{n-1}, c_n, w_{n-2}, \ldots, w_{n+2}$ is found within the date, the value "1" is returned, otherwise the value "0". E.g. a feature function may fire if and only if the predecessor word w_{n-1} is "the" and the concept c_n is "name". Another example of a feature function would be, if and only if the predecessor concept c_{n-1} is "number" and the concept c_n is "currency". We will call the feature functions based on predecessor, current, and successor word *lexical features* and the features based on the predecessor concept *bigram features*.

For clarity we will abbreviate the term in the numerator of equation 1 by

$$H(c_{n-1}, c_n, w_{n-2}^{n+2}) =$$
$$exp \left(\sum_{m=1}^{M} \lambda_m \cdot h_m(c_{n-1}, c_n, w_{n-2}^{n+2}) \right)$$

resulting in

$$p(c_1^N|w_1^N) = \frac{1}{Z} \prod_{n=1}^{N} H(c_{n-1}, c_n, w_{n-2}^{n+2}).$$
(2)

2.1.2. Log-Linear on position level

One possible normalization of Equation 2 is on a positional level:

$$p(c_1^N|w_1^N) = \prod_{n=1}^{N} \frac{H(c_{n-1}, c_n, w_{n-2}^{n+2})}{\sum_{\tilde{c}} H(c_{n-1}, \tilde{c}, w_{n-2}^{n+2})}.$$

This results in the following normalization term:

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

Using equation 2 with normalization 3 and a given training dataset $\{\{c_1^N\}_t, \{w_1^N\}_t\}_{t=1}^{T}$, the criteria for training and decision making are given by

$$\hat{\lambda}_1^M = \operatorname*{argmax}_{\lambda_1^M} \left\{ \sum_{t=1}^{T} \log p(\{c_1^N\}_t, \{w_1^N\}_t) \right\} \quad (4)$$

and

$$\hat{c}_1^N(w_1^N) = \operatorname*{argmax}_{c_1^N} \left\{ p(c_1^N|w_1^N) \right\} \quad (5)$$

respectively. This modelling approach is usually referred to as Maximum Entropy approach in the literature, e.g. in (Bender & Macherey[+] 03).

2.1.3. Linear Chain Conditional Random Field (CRFs)

Linear Chain Conditional Random Fields (CRFs) as defined in (Lafferty & McCallum[+] 01) could be represented with equation 2 and a normalization Z on sentence level:

$$Z = \sum_{\tilde{c}_1^N} \prod_{n=1}^{N} H(\tilde{c}_{n-1}, \tilde{c}_n, w_{n-2}^{n+2}). \quad (6)$$

resulting in the probability

$$p(c_1^N|w_1^N) = \frac{\prod_{n=1}^{N} H(c_{n-1}, c_n, w_{n-2}^{n+2})}{\sum_{\tilde{c}_1^N} \prod_{n=1}^{N} H(\tilde{c}_{n-1}, \tilde{c}_n, w_{n-2}^{n+2})}. \quad (7)$$

For both log-linear modelling approaches, the same training and decision criterion is applied. For our experiments, we apply the CRF++ toolkit (Kudo 05) used in (Kudo & Yamamoto[+] 04).

2.2. Machine Translation (MT)

We use a standard phrase-based machine translation method, which combines several models: 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 and the respective model weights λ_m are optimized using minimum error training. A detailed description of the single models can be found in (Mauser & Zens[+] 06).

2.3. Support Vector Machines (SVMs)

SVMs realize a standard classifier-based approach to concept tagging. Binary classifiers are trained for each pair of competing classes. For the final classification, the weighted voting of the single classifiers is considered. We apply the open-source toolkit YAMCHA (Kudo & Matsumoto 01).

2.4. Stochastic Final State Transducers (SFSTs)

In the SFST approach, the translation process from word sequences w_1^N to concept sequences c_1^N is implemented by Finite State Machines. The transducer representing the translation process is a composition of

- a transducer λ_{w2c}, which groups transducers translating words to concepts,

- a transducer λ_{SLM}, representing the stochastic conceptual language model

$$P(w_1^N, c_1^N) = \prod_{n=1}^{N} P(w_n c_n | h_n)$$

  with $h_n = \{w_{n-1} c_{n-1}, w_{n-2} c_{n-2}\}$ (3-gram),

- a transducer $\lambda_{w_1^N}$, which is the FSM representation of the sentence w_1^N.

The best translation is the best path in λ_{SLU}:

$$\lambda_{SLU} = \lambda_{w_1^N} \circ \lambda_{w2c} \circ \lambda_{SLM} \quad (8)$$

All operations are done using the AT&T FSM/GRM Library (Mohri & Pereira[+] 02).

3. Morphologic Features

In addition to the lexical and concept bigram features described in Section 2.1.1., we also tested a set of morphological features. E.g. a capitalized word is a hint for the concept "name". We integrated the following features within both log-linear models:

- capitalization: The capitalization feature is true, if a word is capitalized, longer than three letters (to omit abbreviations), and is not after a fullstop (to omit words at the beginning of a sentence).

- prefixes with given length n: The prefix feature is true, if the first n letters of a word are equal to a predefined sequence of letters, e.g. for length 2: "in-formal".

- suffixes with given length: Similar to the prefix feature, but works on the last letters of a word, e.g. for length 2: "current-ly".

Before the model parameters λ_m are estimated, a list containing all features which have been seen within the training corpus at least once is generated.

4. Corpus Description

For the comparison of the various concept tagging methods respectively modelling approaches described in the previous Section 4., we have chosen a state-of-the art corpus from a spoken language understanding task, namely the MEDIA corpus (Devillers & Maynard[+] 04). This corpus was collected within the scope of the French Media/Evalda project and covers the domain of the reservation of hotel rooms and tourist information. It is divided into three parts: a training set (approx. 13k sentences), a development set (approx. 1.3k sentences) and an evaluation set

(approx. 3.5k sentences). Since the corpus has been collected for the Evaluation of Dialogue systems, there are complete dialogues annotated, i.e. the utterances from the user and from the operator. For this paper, we only consider the dialogue turns uttered by the human users of the system. There are 74 different concept tags ranging from simple date and time expressions (annotated as `date` resp. `temps`) to more complex ones like coreferences (annotated as `lienRef-coRef`). So also the attribute names are written in French since they have been developed within the scope of a French project. One example sentence from the Media training corpus would be:

`je veux une chambre double pour deux personnes.`

It translates to "I would like one double room for two persons". The same sentence, annotated on concept level:

`null{je veux} nombre-chambre{une}`
`chambre-type{chambre double}`
`sejour-nbPersonne{pour deux personnes}`

So the annotation on concept level is basically a segmentation of the input sentence into various chunks. Since the `null`-tag mainly tags words without semantic meaning or hesitations etc, the corresponding attribute name and value pairs which have to be extracted by the various algorithms would be

`nombre-chambre[1] chambre-type[double]`
`sejour-nbPersonne[2]`

The statistics of the corpora are presented in Table 1. Within this corpus, there is a much richer annotation used than explored within this paper. Here, we just evaluate the concept tagging performance of the various approaches and drop some specifiers and modal information. I.e., the resulting corpus does not stick completely to the MEDIA evaluation guidelines but fits well for a comparison of the systems. Thus, only the statistics w.r.t. the word and concept level are presented in the aforementioned table.

5.   Experiments and Results

For all experiments in this paper, we use exactly the same evaluation corpus and the same scoring script, based on the NIST evaluation toolkit (NIST). Thus, we ensure, that the results of the different modelling approaches are comparable. As evaluation criteria, we use the well-established *Concept Error Rate (CER)* and *Sentence Error Rate (SER)*. The CER is defined as the ratio of the sum of deleted, inserted and confused concepts w.r.t. a Levenshtein-alignment for a given reference concept string, and the total number of concepts in that reference string. The SER is defined as ratio of the number of wrong tag sequences and the total number of tag sequences w.r.t. the concept level.

In a first experiment, we compare the various models as described in Section 2. w.r.t. tagging performance on the MEDIA eval corpus set (cf. Table 2). The CRF approach outperforms all other models, on both tasks, the attribute name extraction and the additional attribute value extraction. We obtain a CER of 11.8% on the evaluation corpus just considering attribute names and 16.2% also considering attribute values. The log-linear approach on a positional level is second best. Thus, exponential models seem

Table 2: Results on the MEDIA evaluation corpus for various modelling approaches. The error rates are given w.r.t. attribute name extraction only (columns 2,3) and additional attribute value extraction (columns 4,5).

model	attribute		attribute/value	
	CER [%]	SER [%]	CER [%]	SER [%]
CRF	11.8	20.6	16.2	23.0
log-pos	14.9	22.2	19.3	26.4
FST	17.9	24.7	21.9	28.1
SVM	18.5	24.5	22.2	28.5
MT	19.2	24.6	23.3	27.6

Table 3: Effect of using morphologic features for the CRF modelling approach.

corpus	attribute	
MEDIA-NLU eval	CER [%]	SER [%]
lexical [-2..2]	19.5	28.8
+concepts[-1]	12.8	22.3
+capitalization	12.6	22.2
+suffixes [1..5]	12.0	21.3
+prefixes [1..4]	11.8	20.6

to have a better tagging performance than the other three approaches. For all of the five systems, the attribute value extraction is done in the same way using a rule-based approach.

In a second experiment, we explore the effect of morphologic features within log-linear models. Here, we only report results on attribute name extraction. We tried various feature sets and optimized the parameter settings on the development set of the MEDIA corpus. For the CRF model, we get a CER of 12.8% with taking into account only features on word and concept level. Adding morphologic features could reduce the CER by 8% relative from 12.8% CER down to 11.8% CER (cf. Table 3). The gain in SER is also roughly 8% relative.

For the position dependent log-linear modelling approach, the CER drops from 16.0% with just the elementary features down to 14.9% CER, a gain of 7% relative. The SER can be improved by roughly 6% relative. The results are presented in Table 4.

6.   Conclusion

In this paper, we presented a comparison of various models for concept tagging on the MEDIA corpus w.r.t. tagging performance. Two of the models could further be improved

Table 4: Effect of using morphologic features for the log-linear on a positional level modelling approach.

corpus	attribute	
MEDIA-NLU eval	CER [%]	SER [%]
lexical [-2..2]	20.1	26.4
+concepts[-1]	16.0	23.5
+capitalization	15.5	23.2
+suffixes [4..7]	15.3	22.9
+prefixes [1..5]	14.9	22.2

Table 1: Statistics of the MEDIA training, development and evaluation corpora used for all experiments.

corpus MEDIA-NLU	training		development		evaluation	
	words	concepts	words	concepts	words	concepts
# sentences	12,908		1,259		3,518	
# tokens	94,466	43,078	10,849	4,705	26,676	12,022
vocabulary	2,210	74	838	64	1,312	72
# singletons	798	5	338	3	508	4
# OOV rate [%]	–	–	0.01	0.0	0.01	0.0

by adding morphologic knowledge. To ensure the comparability of the models, they were trained and tested on exactly the same data sets and the evaluation of the tagging hypotheses was done using the NIST evaluation toolkit. With the best model, we achieved a CER of 11.8% on the MEDIA evaluation set.

7. Outlook

Additionally to improving the single systems we plan to do experiments on system combination. Also, since there usually is an ASR component involved in an SLU system, we will explore the effect of ASR errors on the tagging performance. It would also be interesting to apply the presented models on lattices and use ASR-based scores, e.g. word posterior confidences, to improve the SLU systems.

Acknowledgements

8. References

O. Bender, K. Macherey, F.-J. Och, H. Ney. Comparison of alignment templates and maximum entropy models for natural language understanding. In *Conference of the European Chapter of the Association for Computational Linguistics*, pp. 11–18, Budapest, Hungary, April 2003.

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 *Proceedings of the Fourth Int. Conf. on Language Resources and Evaluation (LREC)*, pp. 855–858, Lisbon, Portugal, May 2004.

T. Kudo, Y. Matsumoto. Chunking with support vector machines. In *Proceedings of the Meeting of the North American chapter of the Association for Computational Linguistics (NAACL)*, pp. 1–8, Pitsburgh, PA, USA, June 2001.

T. Kudo, K. Yamamoto, Y. Matsumoto. Applying conditional random fields to japanese morphological analysis. In D. Lin, D. Wu, editors, *Proceedings of EMNLP 2004*, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

T. Kudo. Crf++ toolkit. http://crfpp.sourceforge.net/, 2005.

J. Lafferty, A. McCallum, F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pp. 282–289, Williamstown, MA, USA, June 2001.

A. Mauser, R. Zens, E. Matusov, S. Hasan, H. Ney. The rwth statistical machine translation system for the iwslt 2006 evaluation. In *International Workshop on Spoken Language Translation*, pp. 103–110, Kyoto, Japan, Nov. 2006. Best Paper Award.

M. Mohri, F. Pereira, M. Riley. Weighted finite-state transducers in speech recognition. *Computer, Speech and Language*, Vol. 16, No. 1, pp. 69–88, 2002.

NIST. Speech recognition scoring toolkit (SCTK). http://www.nist.gov/speech/tools/.

C. Raymond, G. Riccardi. Generative and discriminative algorithms for spoken language understanding. In *Interspeech*, pp. 1605–1608, Antwerp, Belgium, Aug. 2007.