System Combination for Spoken Language Understanding

Stefan Hahn, Patrick Lehnen, Hermann Ney

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

Abstract

One of the first steps in an SLU system usually is the extraction of flat concepts. Within this paper, we present five methods for concept tagging and give experimental results on the state-of-the-art MEDIA corpus for both, manual transcriptions (REF) and ASR input (ASR). Compared to previous publications, some single systems could be improved and the ASR results are presented for the first time. We could improve the tagging performance of the best known result on this task by approx. 7% relatively from 16.2% to 15.0% CER for REF using light-weight system combination (ROVER). For the ASR task, we achieve improvements by approx. 3% relatively from 29.8% to 28.9% CER. An analysis of the differences in performance on both tasks is also given.

Index Terms: Spoken dialogue systems, system combination

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 French MEDIA corpus can be represented as:

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

This sentence part roughly translates into "...on the seventh of April in this hotel". The tagging of a sentence with concepts can be interpreted as a segmentation of a word sequence in semantical chunks. The chunks are represented with curly brackets, with attribute values written below. In square brackets behind the name follows the attribute value. Since the modelling approaches rely on a 1-to-1 mapping between word and concept sequence, usually *concept tags* are introduced. Thus, it is ensured that the word sequence has the same length as the concept sequence. For the first part of the example from above, this would look like:

$$\texttt{...}\quad \underbrace{\texttt{au}}_{\texttt{temps-date\_start}}\quad \underbrace{\texttt{sept}}_{\texttt{temps-date\_cont}}\quad \underbrace{\texttt{avril}}_{\texttt{temps-date\_cont}}\quad \texttt{...}$$

For ease of terminology, we will just speak of concepts rather than concept tags throughout this paper. It should be noted that the concept tags are just introduced for modelling reasons and do not appear in the final output of the systems. We explore various methods for concept tagging, which are shortly described in the following section. After the presentation of our training and testing data, the state-of-the-art MEDIA corpus, in Section 3, the experimental results are presented in Section 4. We present improved single-system results for REF and ASR as well as first system combination results including an error analysis on concept level. A summary of the paper is given in Section 5 and it concludes with an outlook in Section 6.

2. Methods and Models

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 $c_1^N = c_1, \dots, c_N$ given a word sequence $w_1^N = w_1, \dots, w_N$:

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

using

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

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 parameters λ_m which are calculated 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 $h_m(c_{n-1}, c_n, w_{n-2}^{n+2}) \in \{0, 1\}$. If a pre-defined combination of the values $c_{n-1}, c_n, w_{n-2}, \dots, w_{n+2}$ 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". We apply feature functions based on predecessor, the current, and successor words (*lexical features*), features based on the predecessor concept (*bigram features*) and *morphological features* capturing pre- and suffixes as well as capitalization.

2.1.2. Log-Linear on position level

One possible normalization of Equation 1 is on a positional level:

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

Using equation 1 with normalization 2 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

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \left\{ \sum_{t=1}^{T} \log p(\{c_1^N\}_t, \{w_1^N\}_t) \right\} \qquad (3)$$

and

$$\hat{c}_1^N(w_1^N) = \underset{c_1^N}{\operatorname{argmax}} \left\{ p(c_1^N | w_1^N) \right\} \qquad (4)$$

respectively.

2.1.3. *Linear Chain Conditional Random Field (CRFs)*

Linear Chain Conditional Random Fields (CRFs) as defined in [1] could be represented with equation 1 and a normalization Z on sentence level:

$$Z = \sum_{\tilde{c}_1^N} \prod_{n=1}^{N} H(\tilde{c}_{n-1}, \tilde{c}_n, w_{n-2}^{n+2}). \qquad (5)$$

For both log-linear modelling approaches, the same training and decision criterion is applied. For our CRF experiments, we apply the CRF++ toolkit [2], while we use an in-house software (including [3]) for the log-pos model.

2.2. 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}:

$$\lambda_{SLU} = \lambda_{w_1^N} \circ \lambda_{w2c} \circ \lambda_{SLM} \qquad (6)$$

All operations are done using the AT&T FSM/GRM Library [4].

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 [5].

2.4. 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. For a more detailed description, see [6].

3. Corpus Description

For the comparison of the various concept tagging methods resp. modelling approaches described in the previous section, we have chosen a state-of-the-art corpus from a spoken language understanding task, namely the MEDIA corpus [7]. It covers the domain of the reservation of hotel rooms and tourist information and the incorporated concepts have been designed to match this task. There is e.g. a concept for hotel name or room type. The corpus 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). The statistics of the training, development and evaluation corpora are presented in Table 1. Within this corpus, there is a much richer annotation used than explored within this paper. 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.

4. Experimental Results

The results for all systems presented in this section where produced using the same data for training and testing. Scoring of the hypotheses was done using the NIST scoring toolkit [8]. As error criterion we use the well-known *Concept Error Rate (CER)*, which is defined as the ratio of the sum of deleted, inserted and confused concepts (not concept *tags*), and the total number of concepts in all reference strings. Substitutions, deletions and insertions are calculated using a Levenshtein-alignment between a hypothesis and a given reference concept string.

4.1. Single Systems - REF task (manual transcriptions)

Compared to the results presented in [9], improvements in CER have been achieved for the FST and SVM approach (approx. 21% resp. 14% relatively) due to the introduction of categorization as an additional feature. The categorization is realized by the use of 18 generalization classes, e.g. numbers, weekdays, country names, hotel names, etc. Results for all of the five systems are given in Table 2. The systems are ranked by performance w.r.t. CER. The CRF approach clearly outperforms all other systems. A detailed error analysis on concept level has shown that four concepts are tagged (slightly) better by competing systems: `objet` (e.g. hotel) and `temps` (time data) by the FST system, `connectprop` (conjunction) by the SVM system and `paiement` (payment) by the log-pos model.

If we take a closer look at the different kinds of errors produced by the systems, we observe that substitution errors are responsible for approx. 25-30% of the total errors of the systems. For Deletion and Insertion errors, there is a much higher variability. E.g., for CRF and SVM, more than 50% of the errors are deletions, approx. 20% insertions whereas for the MT system, approx. 26% of the errors are deletions and approx. 49% insertions. Due to this imbalance between the different kinds of errors across the various systems, this is an indication that system combination may help to reduce the overall error rate.

Since concept tagging is only the first step in an SLU pipeline, we mainly discuss the system performance w.r.t. attribute name and value pairs, since usually both are passed to the next module within the pipeline, which deals with interpretation. The attribute value extraction is performed in the same way for all systems using a rule-based approach.

Table 1: Statistics of the MEDIA training, development and evaluation corpora.

corpus	training		development		evaluation	
MEDIA-SLU	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

Table 2: Attribute and Attribute/Value CER for the five described systems on the MEDIA evaluation corpus. The attribute CER is also presented broken down in substitution, insertion and deletion errors.

	attribute CER [%]				attr./value
model	Sub	Del	Ins	CER	CER [%]
CRF	3.2	6.4	2.3	11.8	16.2
FST	4.2	5.0	4.9	14.1	18.1
log-pos	4.2	6.1	4.8	15.0	19.3
SVM	4.5	8.2	3.2	15.9	19.7
MT	4.8	5.0	9.4	19.2	23.3

Table 3: Attribute and Attribute/Value CER for the five described systems on an automatic transcription of the MEDIA evaluation corpus (WER: 31.4%). The attribute CER is also presented broken down in substitution, insertion and deletion errors.

	attribute CER [%]				attr./value
model	Sub	Del	Ins	CER	CER [%]
CRF	6.5	12.3	5.7	24.6	29.8
FST	8.4	9.4	9.6	27.5	32.5
log-pos	8.4	10.4	8.9	27.8	33.5
SVM	8.1	14.3	6.2	28.6	33.5
MT	8.2	11.1	9.9	29.2	35.2

Table 4: System combination results on the REF corpora (CER [%]).

	attribute		attr./value	
model	DEV	EVA	DEV	EVA
CRF	13.0	11.8	16.5	16.2
FST	15.9	14.1	19.2	18.1
log-pos	17.5	15.0	21.7	19.3
SVM	17.9	15.9	21.1	19.7
MT	20.3	19.2	24.2	23.3
ROVER	12.9	11.1	16.5	15.1
weighted ROVER	11.7	11.0	15.2	15.0

4.2. Single Systems - ASR task

In any deployed dialogue 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. Therefor, we use an automatic transcription of the MEDIA development and the evaluation corpus. The ASR word error rate is 30.3% for DEV and 31.4% for EVA. The corresponding tagging results of all five systems on the evaluation corpus are given in Table 3. 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 REF task. The substitution errors are around 25-30% of the total errors for all systems whereas there is a higher variance w.r.t. deletion and insertion errors across systems. The main difference to the REF task is that the MT system does not produce so many insertions anymore: approx. 34% of the errors are deletions and 38% are insertion errors.

As Table 3 compared to Table 2 shows, the CER raises by approx. 150-180% relatively for ASR compared to REF. An error analysis revealed that for two concepts the tagging performance degenerates heavily due to introduced recognition errors: The concept `response` (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. `paiement` (payment) often corresponds to the currency word “euro” which is as well often deleted or confused by non-content words. There are also concepts where the tagging performance is comparatively stable, e.g. `objet` which is often found next to a co-reference tag `lienref`.

4.3. System Combination

Motivated by the differences in tagging performance of some concepts for the five systems, we performed light-weighted system combination experiments using ROVER, which is known to work well for speech recognition [10]. Since we currently just consider the single best output of each system, ROVER is just a majority voting on concept level after a Levenshtein alignment of all systems has been performed. The results are presented in Table 4 for REF and Table 5 for ASR. Using all five systems, there is a gain of approx. 7% relatively for REF (considering attribute/value pairs) and 3% relatively for ASR. We also tried to estimate system weights using Powell’s method, but there is only little improvement.

To analyze how much gain would be theoretically possible, we computed the oracle error rates (cp. Tables 6 and 7). For REF, the oracle CER for the attribute/value condition is 33% lower than the system combination result. Due to the comparatively high WER of the ASR system, the difference in CER between oracle and ASR system combination is lower as in the REF case, namely approx. 20%.

Another interesting aspect is the comparison of the CRF and the log-pos model, since they only differ in the normalization term (position-wise vs. sentence-wise, cf. sections 2.1.2 and 2.1.2) but the CRF approach has a significantly better performance. A comparison of errors on concept level for the REF task is given in Table 8. Especially the concept `command` is tagged more accurately by the CRF system. Here, we did not perform a Levenshtein alignment to count the errors, but compared the concept tag sequences position-wise.

Table 5: System combination results on the ASR corpora (CER [%]).

	attribute		attr./value	
model	DEV	EVA	DEV	EVA
CRF	24.9	24.6	30.2	29.8
FST	28.3	27.5	33.4	32.5
log-pos	28.8	27.8	34.5	33.5
SVM	28.3	28.6	33.3	33.5
MT	29.7	29.2	36.5	35.2
ROVER	24.5	24.0	29.8	29.1
weighted ROVER	23.9	23.8	28.9	28.9

Table 6: Additive oracle error rates (CER [%]) on the NLU corpora for the five systems ordered by decreasing performance.

	attribute		attr./value	
model	DEV	EVA	DEV	EVA
CRF	13.0	11.8	16.5	16.2
+FST	8.3	7.4	12.3	12.1
+log-pos	7.5	6.4	11.6	11.2
+SVM	6.8	5.7	10.8	10.6
+MT	6.0	5.0	10.1	10.0

5. Conclusion

In this paper, we presented tagging performance results for five systems on manual transcriptions (REF) as well as ASR output (ASR) for the state-of-the-art MEDIA corpus. We performed light-weight system combination for both conditions and could thus reduce the CER by 7% relatively for the REF condition and 3% relatively for the ASR condition. Additionally, an error analysis partly explaining the gap in performance for both conditions has been performed.

6. Outlook

Motivated by the error analysis in section 4.1, we expect to improve system combination results by optimizing weights on concept level for each system. These weights could be trained using CRFs. We plan to switch from single-best input to word lattices, at least for the ASR task. This step will hopefully reduce the oracle error rates and we can compute confidence measures which can be used to improve system combination . Preliminary experiments have shown that there is a gain in performance on the ASR set for the MT system, if the parameters are optimized on ASR input. Currently, all systems are trained and optimized on manual transcriptions.

Table 7: Additive oracle error rates (CER [%]) on the SLU corpora for the five systems ordered by decreasing performance.

	attribute		attr./value	
model	DEV	EVA	DEV	EVA
CRF	24.9	24.6	30.2	29.8
+FST	20.7	20.3	26.2	25.9
+log-pos	19.0	18.8	25.0	24.8
+SVM	17.6	17.5	23.5	23.6
+MT	16.3	16.6	22.7	23.1

Table 8: Comparison of the error rate of the log-pos and CRF models on the MEDIA evaluation corpus. Here, no Levenshtein alignment has been performed.

tag	events	log-pos [%]	CRF [%]
null	9534	9.2	8.6 (-7%)
command	2160	32.6	21.5 (-34%)
other	14982	18.2	15.5 (-15%)
Sum	26676	16.2	13.5 (-16%)

7. Acknowledgements

The authors like to thank David Torres Vilar for the MT system and Björn Hoffmeister for his help with the Rover system. Christian Raymond thankfully provided the SVM and FST system as well as the attribute value extraction. This work was partly funded by the European Union under the integrated project LUNA - spoken language understanding in multilingual communication systems (FP6-033549).

8. References

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

[2] T. Kudo, “Crf++ toolkit,” 2005, http://crfpp.sourceforge.net/.

[3] F. Och, “Yet another small maxent toolkit,” 2002, http://www-i6.informatik.rwth-aachen.de/web/Software/YASMET.html.

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

[5] T. Kudo and Y. Matsumoto, “Chunking with support vector machines,” in *NAACL ’01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*. Morristown, NJ, USA: Association for Computational Linguistics, 2001, pp. 1–8.

[6] A. Mauser, R. Zens, E. Matusov, S. Hasan, and H. Ney, “The RWTH statistical machine translation system for the IWSLT 2006 evaluation,” in *International Workshop on Spoken Language Translation*, Kyoto, Japan, Nov. 2006, pp. 103–110, best Paper Award.

[7] 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)*, Lisbon, Portugal, May 2004, pp. 855–858.

[8] NIST, “Speech recognition scoring toolkit (SCTK),” http://www.nist.gov/speech/tools/.

[9] S. Hahn, P. Lehnen, C. Raymond, and H. Ney, “A comparison of various methods for concept tagging for spoken language understanding,” in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 2008, p. accepted for publication.

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