Optimizing CRFs for SLU Tasks in Various Languages Using Modified Training Criteria

Stefan Hahn, Patrick Lehnen, Georg Heigold, Hermann Ney

Human Language Technology and Pattern Recognition
Computer Science Department, RWTH Aachen University, Germany.
{hahn, lehnen, heigold, ney}@cs.rwth-aachen.de

Abstract

In this paper, we present improvements of our state-of-the-art concept tagger based on conditional random fields. Statistical models have been optimized for three tasks of varying complexity in three languages (French, Italian, and Polish). Modified training criteria have been investigated leading to small improvements. The respective corpora as well as parameter optimization results for all models are presented in detail. A comparison of the selected features between languages as well as a close look at the tuning of the regularization parameter is given. The experimental results show in what level the optimizations of the single systems are portable between languages.
Index Terms: spoken language understanding, conditional random fields, training criteria, tagging

1. Introduction

In the last years, conditional random fields (CRFs) have become quite popular in the speech processing domain for e.g. transliteration of named entities or Parts-of-Speech tagging [1]. This method has also proven to be an effective approach to solve the task of attribute name extraction or concept tagging [2, 3]. The task of concept tagging is usually one of the first steps when building an SLU system. It can be described as extracting basic semantic chunks out of a given word sequence. Although several languages and tasks are under investigation in this paper, the general idea of attribute name extraction is the same for all of these. Figure 1 shows an example taken from the French MEDIA corpus. The input word sequence is shown in the first line, the resulting attribute names and accompanying values are shown in lines three and four. Line two shows a way of how to model resp. circumvent the alignment problem. Usually, an attribute name may cover more than one word. For the training of CRF models, a 1-to-1 alignment between words and tags is needed. One way to get this alignment is to assign so-called "start" and "continue" *concept tags* to the words. Using this approach results in a 1-to-1 alignment and the original attribute name sequence can be recovered. The disadvantage of this approach is that we now have to train the CRF model on tag-level and there are roughly twice as many tags as attribute names. Since the complexity of CRFs is proportional to the square of the size of the attribute name vocabulary (if only transitions of length two are used on the concept side), the training time will be higher.

Besides the optimization of CRF-based systems for various tasks and languages, one focus of this paper is the investigation of modified training criteria. The experiments also show how the model parameters like regularization vary between tasks/languages.

The following Section 2 gives a short overview of the original linear-chain CRF model and the modifications to the training criterion we applied. Section 3 presents the three tasks and corpora for each of which we trained and optimized a model from scratch. The results and findings of our experiments are presented in Section 4. The paper is completed with a conclusion and an outlook presented in Section 5.

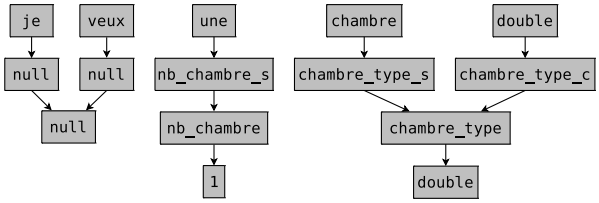

Figure 1: Example illustrating the general idea of concept tagging (French: "I want a double room"). The first line shows the input word sequence, the third and fourth line the appropriate attribute names and values. The second line shows how the 1-to-1 alignment is modelled using start and continue tags.

2. CRFs

Conditional random fields (CRFs) were introduced in [1] as a graphical framework for building probabilistic sequential models. This discriminative approach directly models the posterior probability of sequence c_1^N given the sequence w_1^N

$$p_\Lambda(c_1^N|w_1^N) = \frac{1}{Z}\exp\left(\sum_i \lambda_i f_i(c_1^N, w_1^N)\right) \quad (1)$$

where Z denotes the normalization constant, $Z := \sum_{c_1^N}\exp\left(\sum_i f_i(c_1^N, w_1^N)\right)$. The model parameters are $\Lambda = \{\lambda_i\}$. The feature functions are used to model dependencies between the random variables. In contrast to general log-linear models, CRFs restrict the allowed feature functions to a subset of structured feature functions $f_i(c_1^N, w_1^N) = \sum_{n=1}^N f_{i,n}(c_{n-1}, c_n, w_1^N)$ [1]. We use binary features functions covering lexical and word part information (i.e. prefix, suffix, capitalization) as defined in [3].

Standard Training Criterion Assume labeled training data $\{c_1^{N_r}, w_1^{N_r}\}$. The standard training criterion for CRFs maximizes the entropy

$$\mathcal{F}^{(MMI)}(\Lambda) = \sum_r \log p_\Lambda(c_1^{N_r}|w_1^{N_r}) - C\sum_i |\lambda_i|^p. \quad (2)$$

Typically, some regularization is added for a more stable convergence. We use the Lp-norm for some $p > 0$ in Equation (2)

with some normalization constant $C \geq 0$. In this work, the training criteria are optimized using Rprop [4]. For the default setting $p = 2$ (L2-norm), the result is not expected to be sensitive on the optimization algorithm because the training criterion is convex.

2.1. Modified Training Criteria

Next, different modifications to this standard training criterion are investigated. The proposed training criteria are all instances of the unified training criterion and thus, can be solved in our transducer-based discriminative framework [5].

Power approximation to logarithm For the standard training criterion in Equation (2), small class posterior probabilities are assigned a high loss. This is because the logarithm diverges for zero probabilities, $\log p \stackrel{p \to 0}{\to} \infty$. This means that the standard training criterion in Equation (2) is not robust against outliers, e.g. incorrect transcriptions. To avoid the divergence of the logarithm, the identity

$$\log x \quad = \quad \lim_{r \to 0} \frac{x^r - 1}{r} \tag{3}$$

is used to approximate the logarithm. In contrast to the logarithm, this approximation is bounded below for $r > 0$. This approximation is termed *power approximation* and resembles an error-based training criterion. The effect of this approximation is that bad outliers are assigned zero weight for accumulation. For this reason, this training criterion is expected to perform more robustly than the standard training criterion. Like all bounded/error-based training criteria for log-linear models (without proof), this training criterion has the disadvantage of not being convex. In our transducer-based framework supporting the unified training criterion [5], the smoothing function $\log x$ for the standard training criterion is replaced with $\frac{x^r - 1}{r}$.

Margin-based extension A margin term can be incorporated into the standard training criterion as introduced in [5]. To do this, the posterior in Equation (1) needs to be changed into a margin-posterior

$$p_{\Lambda,\rho}(c_1^N | w_1^N) = \frac{1}{Z} \exp\left(\sum_i \lambda_i f_i(c_1^N, w_1^N) - \rho \mathcal{A}(c_1^N, \tilde{c}_1^N) \right) \tag{4}$$

The normalization constant Z is similarly defined as above. Here, the margin score is set to the word accuracy

$$\mathcal{A}(c_1^N, \tilde{c}_1^N) = \sum_n \delta(c_n, \tilde{c}_n) \tag{5}$$

between the hypothesis c_1^N and the truth $\tilde{c}_1^N$, scaled with the factor $\rho \geq 0$. The margin-based training criteria are obtained by replacing the posterior by the margin-posterior. The such modified training criteria again fit into our transducer-based framework because the margin score can be incorporated by a composition [5].

3. Corpora

In this section, the three tasks resp. corpora from the SLU domain are presented which have been chosen to evaluate the various training criteria. The statistics for all corpora are given in Table 1. Besides the general question which criterion performs best, it is also interesting to see how the optimized parameter settings and feature functions vary between languages/tasks. Concerning the general tagging quality, the amount of training data as well as the size of the concept vocabulary are important figures. They change heavily between the investigated tasks. For all data collections, a table is given with the most interesting statistics. The number of `NULL` tokens on concept level refers to the running number of `NULL` attribute names. On word level, it is the sum of all words tagged with `NULL`. This tag occurs in all corpora and marks words with no semantic meaning for the particular task and is usually the most frequent occurring tag. Concerning evaluation, the `NULL` tag is deleted from reference and hypothesis prior to scoring. Thus, the results better reflect the performance of the models on the attribute names with semantic meaning.

French The so-called MEDIA corpus is a state-of-the-art corpus especially designed for the evaluation of SLU systems [6]. 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, a development set, and an evaluation set. Within this corpus, modes and specifiers are also manually annotated. The experiments carried out in this paper can be directly compared with the so-called "relaxed-simplified" condition within the MEDIA/EVALDA project. Here, some specifiers are dropped and thus the resulting data is not as sparse.

Polish The data for the Polish corpus has been collected at the Warsaw Transportation call-center [7]. Also as part of the LUNA project, the manual annotation of these human-human dialogues has been performed [8]. 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 subsets. 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 concepts is also the largest w.r.t. the three corpora.

Italian The Italian corpus has been collected within the scope of the LUNA project [9]. It covers the domain of software and hardware repairing in the area of an IT help-desk. This corpus is still being collected resp. annotated. So there is only a small amount of data available which does not allow to split the corpus into three sets. Instead, we use the partitioning into two sets as proposed in [10]. It should be noted that the corpus used here consists only of wizard-of-oz dialogues.

4. Experimental Results

Conditional random fields were tuned on all three corpora by first assuming a basic feature set with lexical and bigram transition features to estimate a good regularization constant. Using these models, word part features were added, and the resulting models were used as baseline for experiments testing modified training criteria.

The experiments were evaluated on the respective development and test sets for the three corpora via the NIST scoring toolkit [11]. 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. As already noted, `NULL` tokens are

Table 1: Statistics of the French, Polish and Italian SLU corpora.

		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	1,019		373		–	
	# tokens	8,512	4,742	2,888	1,621	–	–
	# NULL tokens	3,777	1,855	1,294	637	–	–
	vocabulary	1,172	34	636	30	–	–
	# singletons	560	2	313	1	–	–
	# OOV rate [%]	–	–	6.48	0.06	–	–

deleted from hypothesis and reference transcription before scoring. Regularization and feature selection were optimized on the development sets using the standard training criterion (cf. Equation 2). For all reported experiments, only attribute name extraction is considered.

Regularization Two regularization variants *L2*-norm and *L1*-norm are widely used with CRFs. Both were tested on the French MEDIA corpus, resulting in lower error rates for the *L2*-norm regularization (CER of 13.1% on the development set versus a CER of 13.5% for *L1*-norm; see Figure 2).

Based on the results on the French MEDIA corpus, only *L2*-norm regularization was optimized for the Polish and Italian corpora. Variations in the concept error rate are only significant, when changing the regularization parameter C in an exponential manner. Evaluating the range from 2^{-11} to 2^{0} resulted in $C = 2^{-3}$, $C = 2^{-6}$, and $C = 2^{-8}$ with a CER on the development set of 13.1%, 25.7%, and 22.1% for French, Polish, and Italian respectively (cf. Figure 2).

Since our modelling approach relies on a 1-to-1 mapping between word and attribute name sequence, the attribute names are usually broken down in "start" and "continue" tags. In general during search, CRFs permit an attribute name tag sequence `start_A` **`A B`**, which can not be seen in training, since it conflicts with the *start tag* rule. The correct sequence would have been `start_A A start_B`. This problem can be solved by either interpreting a transition `A→B` as `A→start_B` or reducing the search space by all conflicting transitions like `A→B`. On all three corpora better results were obtained for a range of regularization parameters by interpreting a transition `A→B` as `A→start_B`. E.g. for the Italian corpus, the CER on the development set increases from 22.1% to 22.6%, if the search space is reduced.

Feature Selection The feature build up was done in three steps: First, the window size of lexical features was optimized. Second, prefix, suffix and capitalization features were optimized independently in addition to the lexical and transition features. Prefix and Suffix lengths have been tested by incrementally increasing their length and always including smaller

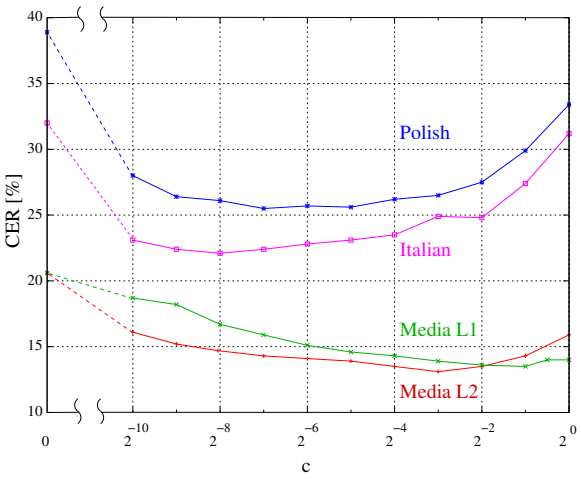

Figure 2: Regularization versus concept error rate for the various corpora. For the French MEDIA corpus, L1-norm regularization is given in addition to L2-norm regularization.

prefixes/suffixes. Finally, the best feature sets were combined in the order of their independent performance. Since extraction of prefixes and suffixes is not meaningful for all languages, it is possible that they do not help to obtain a better performance. In these cases they were not included in the feature build-up.

On all corpora a small window size on the input words was sufficient. Using more than one successor word and more than two predecessors did not result in a better error rate. Using word part features reduced the error rate about 11% in average across languages. An overview of the tagging results for each optimization step is presented in Table 2.

Improved Training Criteria The different variants of the standard training criterion for CRFs ('log') were tested on the three tasks described in Section 3. All setups were optimized from scratch. The experimental results are summarized in Table 3. The experiments based on the power approximation in Equation (3) ('power approx.') suggest that robustness is not an

Table 2: Concept Error Rates (CER) (attribute name extraction) for various feature settings (build-up) on the DEV and EVA corpora for French, Polish and Italian.

		CER [%]	
	features [window]	DEV	EVA
French	lexical [-1..1] + concepts[-1]	13.1	12.3
	+capitalization	13.0	12.0
	+prefixes [1..4]	12.8	11.5
Polish	lexical [-1..1] + concepts[-1]	25.7	26.1
	+prefixes [1..4]	22.8	23.5
	+suffixes [1..4]	22.0	22.7
	+capitalization	21.8	22.6
Italian	lexical [-2..1] + concepts[-1]	21.5	–
	+prefixes [1..6]	19.7	–
	+suffixes [1..5]	18.6	–

Table 3: Concept Error Rates (CER) (attribute name extraction) for various training criteria on the French, Polish and Italian DEV and EVA corpora.

	French		Polish		Italian
training criterion	DEV	EVA	DEV	EVA	DEV
log	12.8	11.5	21.8	22.6	18.6
power approx.	12.8	11.3	21.8	22.5	18.9
margin & log	12.5	10.6	21.1	21.5	17.8
margin & power	12.3	10.7	21.1	21.6	18.3

issue for these three corpora, probably because of the careful transcriptions of the data. The incorporation of a margin term into the standard criterion ('margin & log') leads to consistent improvements, in particular on the independent evaluation corpora. The Polish task benefits least from the margin term. This might be due to the increased confusability caused by the significantly larger vocabulary compared with the other two tasks. For numerical reasons and similar to SVMs, the margin parameter ρ was set to unity and only the regularization constant was tuned. The optimum regularization constant for the margin-based training criteria tended to be smaller than for the corresponding training criterion without margin, for all tasks around 0.1, cf. Figure 2. Combining the power approximation with the margin concept ('margin & power'), again does not help. In some cases, it even leads to worse error rates. An explanation for this observation might be that in contrast to the log-based criteria, the criteria based on the power approximation are non-convex and thus, can get stuck in spurious local optima.

5. Conclusion and Outlook

In this paper, we have presented state-of-the-art concept tagging results on three corpora in the languages French, Italian and Polish using statistical models based on CRFs. Modification to the classical training criterion have been investigated leading to improved concept error rates on two of the three tasks. For the French MEDIA corpus, the best result for the "relaxed-simplified" condition could be improved by 5% relatively. The effect of the regularization parameter on the CER has been shown as well as the differences in the feature selection process leading to optimal results for all languages.

Until now, we apply the modified training criteria after optimizing the feature functions using the classical MMI criterion. It is still an open question if the optimal set of feature functions may vary with the modifications of the training criterion. Currently, the input for our experiments are manual transcriptions of the original recordings. Since in a deployed SLU system, usually an ASR system is applied which introduces a certain kind of errors, it would be interesting to investigate the influence of the erroneous input on the robustness and quality of the tagging systems across languages.

6. Acknowledgements

This work was partly funded by the European Union under the integrated project LUNA - spoken language understanding in multilingual communication systems (FP6-033549).

7. References

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

[2] C. Raymond and G. Riccardi, "Generative and Discriminative Algorithms for Spoken Language Understanding," in *Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 1605–1608.

[3] 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.

[4] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The Rprop algorithm," in *Proc. of the IEEE Int. Conf. on Neural Networks*, 1993.

[5] G. Heigold, R. Schlüter, and H. Ney, "Modified MPE/MMI in a transducer-based framework," in *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009.

[6] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn *et al.*, "Semantic Annotation of the French Media Dialog Corpus," in *Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 3457–3460.

[7] K. Marasek and R. Gubrynowicz, "Design and Data Collection for Spoken Polish Dialogs Database," in *Proc. of the Sixth Int. Conf. on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 2008.

[8] A. Mykowiecka, K. Marasek, M. Marciniak, J. Rabiega-Wiśniewska, and R. Gubrynowicz, "Annotation of Polish spoken dialogs in LUNA project," in *3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC)*, Poznan, Poland, Oct. 2007.

[9] C. Raymond, G. Riccardi, J. Rodrígez, and J. Wisniewska, "The LUNA Corpus: an Annotation Scheme for a Multi-domain Multi-lingual Dialogue Corpus," in *Proc. of the 11th Workshop on the Semantics and Pragmatics of Dialogue (Decalog)*, Trento, Italy, May 2007, pp. 185–186.

[10] M. Dinarelli, A. Moschitti, and G. Riccardi, "Joint Generative and Discriminative Models for Spoken Language Understanding," in *Proc. of the 2008 IEEE Workshop on Spoken Language Technology (SLT)*, Goa, India, Dec. 2008, pp. 61–64.

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