(Hidden) Conditional Random Fields Using Intermediate Classes for Statistical Machine Translation

Patrick Lehnen, Jan-Thorsten Peter, Joern Wuebker, Stephan Peitz, Hermann Ney
Human Language Technology and Pattern Recognition,
Computer Science Department, RWTH Aachen University, Aachen, Germany
{lehnen,peter,wuebker,peitz,ney}@cs.rwth-aachen.de

Abstract

One of the major components of Statistical Machine Translation (SMT) are generative translation models. As in other fields, where the transition from generative to discriminative training resulted in higher performance, it seems likely that translation models should be trained in a discriminative way. But due to the nature of SMT with large vocabularies, hidden alignments, reordering, and large training corpora, the application of discriminative methods is only feasible when using effective speed up techniques. We will show that translation models trained with Conditional Random Fields (CRFs) using classes are useful in translation, even in addition to a strong baseline. Results with an independent CRF translation system and n-best list rescoring will be presented. To design the tandem of CRF translation model and a phrase based baseline we will evaluate two different ways of n-best list integrations.

1 Introduction

Over the last decade a variant of Maximum Entropy (ME) models for sequences, Conditional Random Fields (CRFs) (Lafferty et al., 2001) and Hidden CRFs (HCRFs) (Quattoni et al., 2007) have shown high accuracies in various fields including part-of-speech tagging (Lafferty et al., 2001), semantic tagging (Hahn et al., 2010), chunking (Sha and Pereira, 2003), speech recognition (Zweig and Nguyen, 2009), and language modeling (Roark et al., 2004). But designing (H)CRFs for Statistical Machine Translation (SMT) seems to be a serious challenge. In SMT a sequence $x_1^J = x_1, \ldots, x_J$ composed of symbols from a large vocabulary $\mathbb{X}$ (of size $10k$-$100k$) is mapped to a sequence $y_1^I = y_1, \ldots, y_I$ composed of symbols which are from a large vocabulary $\mathbb{Y}$ ($10k$-$100k$). The critical points are that the vocabularies are both very large, and an alignment is seldom provided with the corpora. As CRFs include a summation over all possible target sequences (equation 1), the computational complexity of CRFs can be expressed by a polynomial of the target vocabulary size $|\mathbb{Y}|$ with a degree equal to the size of the feature describing the largest tuple in the target sequence (n-gram in language modeling (LM)).

Authors have published approaches to move computation time to the lower degree parts of the polynomial, e.g. (Lavergne et al., 2010). However, this only changes constants in the complexity, not the overall complexity. Blunsom and Cohn (2006) and Niehues and Vogel (2008) avoid this problem by improving the alignment A used for the phrase extraction of a phrase based translation (PBT) system. In this case the source and target sequences are given, and the effective target vocabulary are either active or non-active alignments points $p(A|y_1^I, x_1^J)$, which is faster to compute than a sequence from $\mathbb{Y}$, but reference alignments are needed, which are usually not provided with a machine translation corpus. Another approach is to manually constrain the summation to a reduced set of target sequences. Blunsom et al. (2008) propose to extend a hierarchical machine translation system and constrain the summation to the derivations given by this system, while (Lavergne et al., 2011) use a PBT system and use the phrase table to

constrain the summation. Unfortunately, both approaches could only improve constrained systems. In (He and Deng, 2012) the authors propose to constrain the summation to 100-best lists produced by a generative machine translation system for the training corpus. Up to our knowledge He and Deng (2012) reports the first improvements over a strong baseline with a system similar to CRFs.

Over the last years at the same time a different approach to speed up ME model estimation became popular for neural network (NN) language modeling (LM). In (Goodman, 2001) the usage of an intermediate variable c is proposed, called *classes*. They model the translation problem as a product of first translating the source sequence into the classes and then the classes into the target sequence. This approach was adopted to NN and improved training time significantly (Morin and Bengio, 2005).

The main contribution of this work is to transfer the idea to use intermediate classes (Goodman, 2001) together with n-best rescoring to SMT.

In section 2 CRFs are summarized together with an extension for hidden variables and a possible implementation, followed by the concept of intermediate word classes in section 3, while section 4 focuses on the necessary extensions of CRFs to be used in SMT, we discuss the experimental results in section 5, and conclude with section 6.

2 Conditional Random Fields (CRF)

Linear Chain Conditional Random Field (LC-CRF) introduced by (Lafferty et al., 2001) is a maximum entropy approach modeling the conditional probability of a target sequence y_1^I with respect to a source sequence x_1^J using a cumulative feature function $H(y_{i-1}, y_i, x_1^J) = \sum_{l=1}^{L} \lambda_l h_l(y_{i-1}, y_i, x_1^J)$:

$$p(y_1^I|x_1^J) = \frac{e^{\sum_{i=1}^{I} H(y_{i-1},y_i,x_1^J)}}{\sum_{\tilde{y}_1^I} e^{\sum_{i=1}^{I} H(\tilde{y}_{i-1}^i,x_1^J)}} \quad (1)$$

The feature weights λ_1^L are estimated by maximization of the conditional log-likelihood $\mathcal{L}$, which is commonly extended by prior distributions $p_n(\lambda_1^L) \propto e^{-\frac{c_n}{n}|\lambda_1^L|_n^n}$, e.g. L1 c_1 and L2 c_2 regularization:

$$\mathcal{L} = \sum_{k=1}^{K} \log p(\{\overline{y}_1^I\}_k|\{x_1^J\}_k) - \sum_{n=1}^{2} c_n||\lambda_1^L||_n^n \quad (2)$$

with $k = 1, \ldots, K$ summing over the training corpus and $\{\overline{y}_1^I\}_k$ the k-th reference translation.

To support a latent variable, e.g. an alignment A, various authors (Quattoni et al., 2007; Koo and Collins, 2005; Yu and Lam, 2008) suggested a summation in the numerator and the denominator of equation 1:

$$p(y_1^I|x_1^J) = \frac{\sum_A e^{\sum_{i=1}^{I} H(A,y_{i-1},y_i,x_1^J)}}{\sum_{\tilde{A}} \sum_{\tilde{y}_1^I,I(A)} e^{\sum_{i=1}^{I} H(\tilde{A},\tilde{y}_{i-1},\tilde{y}_i,x_1^J)}} \quad (3)$$

Three types of features $h_l(y_{i-1}, y_i, x_1^J)$ were used to support the conditional probability, first *source-n-gram* features depending only on one target symbol y_i and a combination of source symbols $x_{A(j)+\gamma_1}^{A(j)+\gamma_2}$ relative to the currently aligned source word $x_{A(j)}$ (with $\gamma_1 \leq \gamma_2$), with $\gamma_1, \gamma_2 = -5, \ldots, 5$, $\gamma_1 + \gamma_2 + 1 \leq 3$, second *target-n-gram* features describing the relation of a consecutive set of target symbols y_i, y_{i-1}, and third *word stem* features, including prefixes and suffixes up to length 4 and capitalization.

Lehnen et al. (2012) described the use of *begin* (b), *continue* (c), and *skip* (s) labels (inspired by (Ramshaw and Marcus, 1995)) for each target vocabulary word to support monotonous HCRFs (equation 3). At each source symbol the last target symbol is continued, a new one begins, or two target symbols begin. Each target symbol is labeled to make this mapping bijective. At each source symbol all aligned target words are known and at each target symbol all aligned source symbols are known. Features are applied to all combinations of source and target symbols. This modeling restricts the summation over I to $I < 2J$.

The applied CRF software was realized with weighted finite state transducers (Mohri, 2009). A chain represents the source sentence, which is augmented by the target vocabulary, and composed with a LM like automaton. In augmentation, all features were applied without a dependency on the target context (y_n, x_1^N), while the LM like automaton was weighted with all features depending only on the target context (y_{n-1}, y_n). From the final

number of classes	average # of elements per class	average # of elements per pos.
250	223	162
500	111	66

Table 1: Two sets of unsupervised word classes estimated by the method described in (Och, 1995).

automaton, the best path can be selected by utilization of a single source shortest distance (SSSD) operation on a tropical semi-ring, or the posterior weights in the log-likelihood (equation 2) can be calculated by a posterior operation with respect to a log-semi-ring (Lehnen et al., 2011).

3 Word Classes

Maximum Entropy approaches in general and CRFs in particular have an unfavorable computational complexity with respect to the size of the used target vocabulary $|\mathbb{Y}|$. In general, at least a linear dependency could be expected due to the sums in the denominator in equation 3. In (Goodman, 2001) the probability $p(y_1^I|x_1^J)$ was factorized with the help of a clustering function $\gamma : y \mapsto c$, clustering target words y to classes c:

$$\begin{aligned} p(y_1^I|x_1^J) &= \sum_{c_1^I} p(y_1^I|c_1^I, x_1^J)\, p(c_1^I|x_1^J) \qquad (4) \\ &\approx \max_{c_1^I} \left\{ p(y_1^I|c_1^I, x_1^J)\, p(c_1^I|x_1^J) \right\} \end{aligned}$$

If γ is a strict partitioning clustering, the $\sum_{c_1^I}$ and respectively $\max_{c_1^I}$ can be removed, because $p(y_1^I|c_1^I, x_1^J) = 0$ for all $\gamma(y_i) \neq c_i$. This concept greatly reduces the computational complexity as the effective target vocabulary in $p(c_1^I|x_1^J)$ is the class vocabulary $|C| \ll |\mathbb{Y}|$ and the computation in $p(y_1^I|c_1^I, x_1^J)$ can be restricted to only those words e which are part of the already selected class $|\gamma^{-1}(c)| \ll |\mathbb{Y}|$. In this publication we used the unsupervised maximum entropy word class estimation described in (Och, 1995) and (Kneser and Ney, 1993) already used in the preparation of Giza++ estimations, resulting in two sets of word classes in table 1.

4 CRF models for SMT

Training the CRFs introduced in section 2 directly for $p(y_1^I|x_1^J)$ with a full translation vocabulary with a size of $55k$ was infeasible. Thus we applied word classes (section 3) and split the translation process into three steps. The first step models $p(c_1^I|x_1^J)$ with a HCRF (equation 3). This trained model is used to maximize the alignment A with respect to the reference source x_1^J and target sequences y_1^I, and finally a model $p(y_1^I|c_1^I, A, x_1^J)$ with a LCCRF (equation 1) using the given alignment A is estimated (see figure 1). Alignments A in the context of HCRF (dashed lines in figure 1) are only needed to define the position $A(j)$ of source-to-target features $x_{A(j)+\gamma_1}^{A(j)+\gamma_2}$. Features are not restricted to the aligned words and may include surrounding words.

As described in section 2 search can be realized by utilization of a Single Source Shortest Distance (SSSD) operation on the final automaton. With the decomposition into two steps $p(c_1^I|x_1^J)$ and $p(y_1^I|c_1^I, A, x_1^J)$ the search is realized by consecutive application of SSSD in both automata. We added a LM score and a word penalty by composition to the final automaton

$$-c_{LM} \cdot \log(p_{LM}) + c_{WP},$$

to compensate that the CRF models have only been trained on the bilingual part of the corpus, and some of the CRF models lack target-n-gram features (y_{n-1}, y_n). Prior to composition with the LM the word labels *begin* (b), *continue* (c), and *skip* (s) where replaced with the pure word in case of *begin* (B) and *skip* (s) and by an ϵ-label in case of *continue* (c). Composing the LM automaton with the full search space of the (H)CRF automata is computationally infeasible, raising the need of posterior pruning before LM composition with a beam pruning threshold of 5. Contrary to phrase based systems, the resulting models are able to estimate a score $\sum_{i=1}^{I} H(A, y_{i-1}, y_i, x_1^J)$ for any sequence of words, including Out of Vocabulary words (OOVs). In the case of an OOV, the features of the CRF with respect to the current word are not activated, but the features from the surrounding region are still activated. However, in SMT it is often desirable to leave OOVs, which are often named entities, untranslated. So the used software detects OOVs and translates them with an OOV-label. There is no further processing of the OOV, all OOVs are translated to the same OOV label.

N-best list rescoring is realized by adding scores in addition to the model scores provided

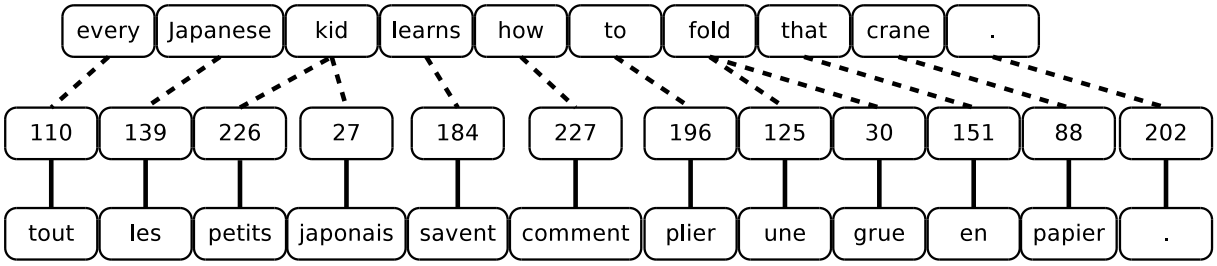

Figure 1: Example of decomposition $p(y_1^I|c_1^I, x_1^J)$ / $p(c_1^I|x_1^J)$. First the English source sentence is translated to a class sequence by a HCRF and finally to the French target sequence by a LCCRF. The dashed lines mark the alignment with the maximum score of the HCRF, while the LCCRF has to use exactly one alignment marked with solid lines. Features are not restricted to the aligned source words and are permitted to use surrounding words as well.

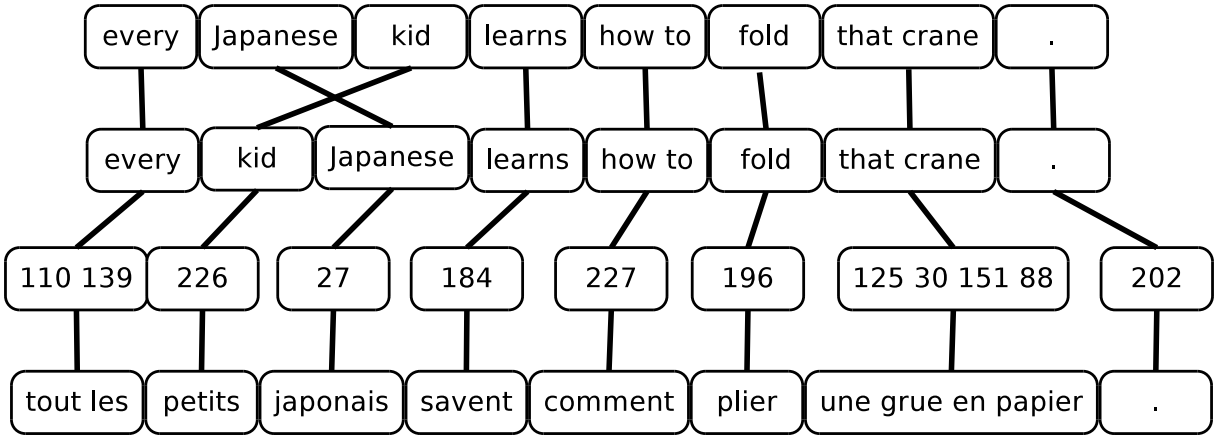

Figure 2: Example of the reordering strategy. A phrase is marked with round boxes. The phrase alignment is used to reorder the English source sequence, afterwards the same approach as in figure 1 is applied. In parameter estimation and n-best rescoring the number of target words in a target phrase is restricted to the number of target words in the target phrase as expected by the phrase alignment.

by the translation system. An (H)CRF score $H(a_1^I, c_1^I, x_1^J) = \sum_{i=1}^{I} H(a_i, c_{i-1}^i, x_1^J)$ can be added in multiple ways. As fully normalized log-probability

$$\begin{aligned} &-\log(p(c_1^I|x_1^J)) = \\ &\quad -\log\left(\sum_{a_1^I} e^{H(a_1^I, c_1^I, x_1^J)}\right) \\ &\quad +\log\left(\sum_{\tilde{a}_1^I}\sum_{\tilde{c}_1^I} e^{H(a_1^I, c_1^I, x_1^J)}\right), \end{aligned}$$

only the numerator of the probability

$$N_{\text{sum}} = -\log\left(\sum_{a_1^I} e^{H(a_1^I, c_1^I, x_1^J)}\right), \qquad (5)$$

or the maximum of the numerator

$$\begin{aligned} N_{\max} &= -\log\left(\max_{a_1^I} e^{H(a_1^I, c_1^I, x_1^J)}\right) \\ &= -\max_{a_1^I} H(a_1^I, c_1^I, x_1^J), \end{aligned} \qquad (6)$$

which equates to the maximum approximation in the alignment. In early experiments it turned out that the full normalization did not improve the translation quality, and as the calculation of the full normalization is much more computational demanding, we did not use it.

(H)CRFs we used do not support reordering directly, only the features support crossing features (e.g. (x_{i-1}, y_i) and (x_i, y_{i-1}) could be used at the same time). To support reordering in parameter estimation we apply Forced Alignments (FA) (Wue-

	training Ted part		dev 2010		test 2010	
	En	Fr	En	Fr	En	Fr
sent.	**107k**		**934**		**1.664**	
words	2M	2M	21k	20k	32k	34k
voc.	44k	56k	3.3k	3.8k	3.8k	4.7k
OOVs	-	-	316	392	322	401

Table 2: IWSLT 2011 evaluation data en→fr

bker et al., 2010) to generate a reference phrase alignment between source and target in the training data (see figure 2). Afterwards the source sequence was reordered with respect to this phrase alignment (between line one and two in figure 2). Additionally, the phrase alignment fixed the number of slots to be filled by the CRF model, i.e. in a phrase with M source and N target words, the CRF model was forced to produce exactly N target words. During search similarly an unconstrained PBT system was used to generate a phrase sequence. In n-best rescoring the source was reordered as in training with this phrase sequence, applying the same slot constrain. Where in SSSD search only the source was reordered and no restriction on the slots was applied. We have to note that with considering the reordering, some information from a PBT system is passed to the SSSD search.

Even though the intermediate classes speed up the training, and we also implemented the concept of (Lavergne et al., 2010), a full training of CRFs using target bigram features is not possible. To permit the training of CRFs with bigram features we included a posterior pruning step in training before applying the bigram features and after the source-n-gram features have been applied. The posterior pruning restricts the number of possible translations in $\sum_{\tilde{y}_1^I, I(A)}$ of equation 3 to a reduced number of summands. As only paths with low scores where removed, most of the probability mass is preserved.

5 Experimental Results

The experiments were conducted on training and test sets extracted from the English to French data of the International Workshop on Spoken Language Translation (IWSLT) 2011. The (H)CRF models were trained on the TED part of the training data, and the IWSLT 2011 development and test sets were applied in optimization and evaluation. As baseline system we were provided with the best single PBT system from (Peitz et al., 2012) for English to French and a PBT system with forced alignments only trained on the TED training data. The first phrase-based system used the SCSS software variant of the Jane software package (Wuebker et al., 2012) and made use of all available in-domain and out-domain data, part-of-speech-based adjective reordering as preprocessing step, a LM with all available monolingual training data, and a 7-gram word class language model. The second system also uses the SCSS software variant of the Jane software package. However, it was trained solely on the TED portion of the training data with the generative training scheme presented in (Wuebker et al., 2010) applying forced alignment. Models include translation probabilities and lexical smoothing in both directions, word and phrase penalty, distance-based reordering model, an n-gram target language model and three binary count features, and a 4-gram language model trained on the TED, Europarl, News-Commentary and Shuffled News corpora from the workshop. From the Shuffled News data 1/8 of the sentences were selected with the technique presented in (Moore and Lewis, 2010).

	word class.	re-ord.	dev 2010 Bleu	dev 2010 Ter	test 2010 Bleu	test 2010 Ter
1	PBT 1 (all data)		28.3	55.9	31.8	50.2
2	PBT 2 (TED)		25.8	58.3	29.4	52.3
	unigram, with LM					
3	250	no	22.7	60.8	25.7	55.1
4	500	no	22.9	60.7	25.7	55.2
5	250	yes	22.8	60.9	26.6	54.1
6	500	yes	22.8	60.5	26.6	54.1
	bigram, without LM					
7	250	no	21.6	60.5	24.6	55.1
8	250	yes	21.5	62.3	25.4	55.4
	bigram, with LM					
9	250	no	21.8	61.3	25.2	55.6
10	250	yes	21.7	61.5	25.1	55.3

Table 3: Results with SSSD search within the CRF framework, and without using a PBT system. Language model scales and word penalty were selected to optimize BLEU on the dev set.

5.1 SSSD search within CRF framework

Table 3 contrasts the results of the two PBT systems (line 1 and line 2) with an independent CRF translation tandem $p(c_1^I|x_1^J)$ / $p(y_1^I|c_1^I, x_1^J)$. Line 3 to 6 replaced the bigram features in model $p(c_1^I|x_1^J)$ with a LM estimated on all available monolingual training data in IWSLT projected to classes, while line 7 and line 8 used bigram features instead of a LM, and line 9 and line 10 used both. Results were very less affected by the choice of features in $p(y_1^I|c_1^I, x_1^J)$. A LM or bigram features in $p(y_1^I|c_1^I, x_1^J)$ did not result in a change in translation performance. Using reordering improves the generalization of the systems testified by approximately 1 bleu improvement on test except in the case of line 10. With $p(c_1^I|x_1^J)$ there seems to be a unfavorable interaction between the bigram feature and the LM avoiding an addition of the modeling power of both. The translation performance was not competitive to the PBT systems even when the PBT was restricted to the same training data (line 2). PBT systems are well designed systems with a decade of detailed development, and it could be expected that some aspects of a translation are captured in a PBT but not in this CRF system, e.g. reordering, why we decided to design an combination of both via n-best rescoring.

5.2 N-best rescoring

Both PBT systems were used to create 1000-best lists. Of all hypotheses with the same resulting word sequence but different phrase segmentation only the one with the best score was added to the n-best list, resulting to a search space with an average of 475 n-best hypotheses per sentence for the development set and 505 n-best hypotheses per sentence for the test set were created with the first PBT system and 312 n-best hypotheses per sentence for the development set and 329 n-best hypotheses per sentence for the test set with the second PBT system. To augment these lists CRF-models modeling $p(c_1^I|x_1^J)$ and $p(y_1^I|c_1^I, x_1^J)$ for 250 and 500 classes with and without reordering were trained without bigram features. Additionally, two CRF-models were trained to capture $p(c_1^I|x_1^J)$ for 250 classes with bigram features. Training CRF-models with bigrams on 500 classes was still to computational demanding. These models were used to augment the n-best lists with the scores N_{sum}, and N_{max}. Using fully normalized probabilities did not change the translation quality. On the final augmented n-best lists the weights for n-best list scores were retrained via Minimum Error Rate Training (MERT) (Och and Ney, 2004), initialized with the best weights of the n-best list generating SCSS system.

	word class.	re-ord.	dev 2010 Bleu	dev 2010 Ter	test 2010 Bleu	test 2010 Ter
1	PBT 2 (TED)		25.8	58.3	29.4	52.3
	(oracle best)		37.3	47.6	43.7	39.7
	(oracle worst)		14.3	75.2	16.3	70.2
	PBT 2 + N_{sum}					
2	250	no	25.8	58.3	29.4	52.3
3	500	no	26.6	57.5	30.2	51.5
4	250	yes	**26.7**	**57.6**	**29.7**	**51.9**
5	500	yes	25.8	58.3	29.4	52.3
	PBT 2 + N_{max}					
6	250	no	25.8	58.3	29.4	52.3
7	500	no	**26.5**	**57.6**	**30.0**	**51.4**
8	250	yes	26.5	57.7	29.9	51.6
9	500	yes	25.8	58.3	29.4	52.3
	PBT 2 + N_{sum} (including target bigram)					
10	250	no	26.4	57.7	29.4	51.9
11	250	yes	**27.1**	**56.9**	**30.1**	**51.2**

Table 4: Results of N-best rescoring adding the (H)CRF scores on top of the scores in the n-best lists of the second PBT system trained on TED data. Line 1 indicates the result of the baseline system. Bold face numbers mark the best result with respect to the dev set.

Experiments have shown that the second model $p(y_1^I|c_1^I, x_1^J)$ did not change the translation quality, and got a zero weight by the MERT training. The results with only the first model $p(c_1^I|x_1^J)$ are shown in table 4 and table 5. To have a fair comparison the parameters of the baseline system (line 1) were reoptimized, too. We have marked the systems giving the best results with respect to the development set. The best systems on the development set produce the best results on the test set, but in some cases the MERT optimization was not able to include the CRF score in a useful way and do not gain an improvement in performance. Best improvements in table 4 were +0.3, +0.6, +0.7 in bleu and -0.4, -0.9, -1.1 in ter, and +0.4, +0.3,

	word class.	re-ord.	dev 2010 Bleu	dev 2010 Ter	test 2010 Bleu	test 2010 Ter
1	PBT 1		28.3	55.9	31.8	50.2
	(oracle best)		40.8	43.9	46.8	36.7
	(oracle worst)		16.4	72.8	18.4	68.6
	PBT 1 + N_{sum}					
2	250	yes	28.3	55.9	31.8	50.2
3	500	no	**28.7**	**55.6**	**32.2**	**49.5**
4	250	yes	28.4	55.6	32.0	49.7
5	500	yes	28.5	55.6	32.2	49.5
	PBT 1 + N_{max}					
6	250	no	28.3	55.9	31.8	50.2
7	500	no	**28.6**	**55.8**	**32.1**	**50.0**
8	250	yes	28.3	55.9	31.8	50.2
9	500	yes	28.3	55.8	31.8	50.2
	PBT 1 + N_{sum} (including target bigram)					
10	250	no	**28.8**	**55.3**	**32.2**	**49.5**
11	250	yes	28.3	55.9	31.8	50.2

Table 5: Results of N-best rescoring adding the (H)CRF scores on top of the scores in the n-best lists of the first and stronger PBT system. Line 1 indicates the result of the baseline system. Bold face numbers mark the best result with respect to the dev set.

+0.4 in bleu and -0.7, -0.2, -0.7 in ter in table 5. The size of improvements were not influenced by the used n-best lists, thus the improvement is not due to adaptation effects. The results with N_{sum} seem to be a bit more stable than the results with N_{max}, which can be explained by a reduced sensitivity to single bad alignments between the source and target sequence. A gain in using the reordering could not been verified in the n-best experiments.

6 Conclusion

In this paper, we have presented the combination of intermediate classes, which became popular over the last years in the context of neural network language models, and conditional random fields (CRFs) for statistical machine translation. We have shown that intermediate classes give a useful alternative to other speed up techniques already tested by different authors. The technique could e.g. be further extended by a better class selection, and an integrated training of the models $p(y_1^I|c_1^I, x_1^J)$ and $p(c_1^I|x_1^J)$. Additionally we have provided a strategy to include CRF scores with PBT systems via n-best rescoring.

Acknowledgments

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. H. Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.

References

Blunsom, Phil and Trevor Cohn. 2006. Discriminative Word Alignment with Conditional Random Fields. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL)*, pages 65–72, Morristown, NJ, USA, July. Association for Computational Linguistics.

Blunsom, Phil, Trevor Cohn, and Miles Osborne. 2008. A discriminative latent variable model for statistical machine translation. In *Proc. of ACL-HLT*.

Goodman, J. 2001. Classes for fast maximum entropy training. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 1, pages 561 –564 vol.1.

Hahn, S., M. Dinarelli, C. Raymond, F. Lefevre, P. Lehnen, R. De Mori, A. Moschitti, H. Ney, and G. Riccardi. 2010. Comparing stochastic approaches to spoken language understanding in multiple languages. *Audio, Speech, and Language Processing, IEEE Transactions on*, PP(99):1.

He, Xiaodong and Li Deng. 2012. Maximum expected bleu training of phrase and lexicon translation models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 292–301, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kneser, Reinhard and Hermann Ney. 1993. Forming Word Classes by Statistical Clustering for Statistical Language Modelling. In Köhler, Reinhard and BurghardB. Rieger, editors, *Contributions to Quantitative Linguistics*, pages 221–226. Springer Netherlands.

Koo, Terry and Michael Collins. 2005. Hidden-variable models for discriminative reranking. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 507–514, Morristown, NJ, USA. Association for Computational Linguistics.

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

Lavergne, Thomas, Olivier Cappé, and François Yvon. 2010. Practical Very Large Scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics, July.

Lavergne, Thomas, Alexandre Allauzen, Josep Maria Crego, and François Yvon. 2011. From n-gram-based to crf-based translation models. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 542–553, Edinburgh, Scotland, July. Association for Computational Linguistics.

Lehnen, Patrick, Stefan Hahn, and Hermann Ney. 2011. N-grams for conditional random fields or a failure-transition posterior for acyclic fsts. In *Interspeech*, Florence, Italy, August.

Lehnen, Patrick, Stefan Hahn, Vlad-Andrei Guta, and Hermann Ney. 2012. Hidden conditional random fields with m-to-n alignments for grapheme-to-phoneme conversion. In *Interspeech*, Portland, OR, USA, September.

Mohri, Mehryar. 2009. Weighted automata algorithms. *Handbook of weighted automata*, pages 213–254.

Moore, Robert C and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224. Association for Computational Linguistics.

Morin, Frederic and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In *Proceedings of the international workshop on artificial intelligence and statistics*, pages 246–252.

Niehues, Jan and Stephan Vogel. 2008. Discriminative word alignment via alignment matrix modeling. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 18–25, Columbus, Ohio, June. Association for Computational Linguistics.

Och, Franz Josef and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational linguistics*, 30(4):417–449.

Och, Franz Josef. 1995. Maximum-Likelihood-Schaetzung von Wortkategorien mit Verfahren der kombinatorischen Optimierung. Studienarbeit, Universität Erlangen-Nuernberg, Germany.

Peitz, Stephan, Saab Mansour, Markus Freitag, Minwei Feng, Matthias Huck, Joern Wuebker, Malte Nuhn, Markus Nußbaum-Thom, and Hermann Ney. 2012. The rwth aachen speech recognition and machine translation system for iwslt 2012. In *International Workshop on Spoken Language Translation*, pages 69–76, Hong Kong, December.

Quattoni, Ariadna, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell. 2007. Hidden Conditional Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:1848–1852.

Ramshaw, Lance and Mitchell Marcus. 1995. Text Chunking using Transformation-Based Learning. In *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 84–94, Cambridge, MA, USA, June.

Roark, Brian, Murat Saraclar, Michael Collins, and Mark Johnson. 2004. Discriminative language modeling with conditional random fields and the perceptron algorithm. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sha, Fei and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 134–141, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wuebker, Joern, Arne Mauser, and Hermann Ney. 2010. Training phrase translation models with leaving-one-out. In *Annual Meeting of the Assoc. for Computational Linguistics*, pages 475–484, Uppsala, Sweden, July.

Wuebker, Joern, Matthias Huck, Stephan Peitz, Malte Nuhn, Markus Freitag, Jan-Thorsten Peter, Saab Mansour, and Hermann Ney. 2012. Jane 2: Open source phrase-based and hierarchical statistical machine translation. In *International Conference on Computational Linguistics*, pages 483–491, Mumbai, India, December.

Yu, Xiaofeng and Wai Lam. 2008. Hidden Dynamic Probabilistic Models for Labeling Sequence Data. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 739–745, Chicago, IL, USA, July.

Zweig, Geoffrey and Patrick Nguyen. 2009. A segmental CRF approach to large vocabulary continuous speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding*.