A Combination of Hierarchical Systems with Forced Alignments from Phrase-Based Systems

Carmen Heger, Joern Wuebker, David Vilar, Hermann Ney

Human Language Technology and Pattern Recognition Group
RWTH Aachen University
Aachen, Germany
surname@cs.rwth-aachen.de

Abstract

Currently most state-of-the-art statistical machine translation systems present a mismatch between training and generation conditions. Word alignments are computed using the well known IBM models for single-word based translation. Afterwards phrases are extracted using extraction heuristics, unrelated to the stochastic models applied for finding the word alignment. In the last years, several research groups have tried to overcome this mismatch, but only with limited success. Recently, the technique of forced alignments has shown to improve translation quality for a phrase-based system, applying a more statistically sound approach to phrase extraction. In this work we investigate the first steps to combine forced alignment with a hierarchical model. Experimental results on IWSLT and WMT data show improvements in translation quality of up to 0.7% BLEU and 1.0% TER.

1. Introduction

During the past years, several works investigated training of phrase translation probabilities for phrase-based statistical machine translation (cf. [1, 2, 3, 4]). By *training* we mean a genuine statistical training that goes beyond pure counting of phrases in word-aligned training data. While most of these approaches suffer from overfitting problems, the first work that successfully counteracts overfitting is [4]. The authors train phrase models by applying a forced alignment procedure where a slightly modified phrase-based decoder is used to find a phrase alignment between source and target sentences. Phrase translation probabilities are then updated based on this alignment. By applying leaving-one-out in the training procedure, overfitting effects can be diminished. The phrase table which is learnt from forced alignments can be used as phrase table itself in the translation system or can be combined with the original phrase-based system. Experiments in [4] show that the latter gives better results.

In recent years, conventional phrase-based systems have been outerperformed by hierarchical phrase-based or syntax-based systems. The papers [5, 6] describe first training approaches with forced alignment techniques with hierarchical translation systems. However, they report difficulties when aligning training sentences because of the restrictions in the phrase extraction process. Our work is intended to neither solve this problem nor propose any forced alignment training method on hierarchical systems. Instead, we want to see the effects of combining a hierarchical system with forced alignments from a phrase-based decoder.

The next section will recall phrase-based and hierarchical phrase-based translation models. In Section 3 we describe the phrase training method with forced alignments and in Section 4 we explain how we combine these forced alignments with hierarchical translation models. An empirical evaluation on two different tasks is done in Section 5. Finally, Section 6 concludes the paper.

2. Translation Models

In this work we will study the combination of two widely used approaches to statistical machine translation. The main difference between the two models lies in the basic units that are used for the translation.

2.1. Phrase-based Translation Model

The phrase based translation model is based on the concept of *phrase*, a bilingual pair of sequences of words that are translations of each other [7].

Given a word-aligned training corpus, we extract those phrases for which the source words are aligned only to target words within the phrase and vice-versa. This set can be formalized for a sentence pair (f_1^J, e_1^I) as

$$\begin{aligned}\mathcal{P}(f_1^J, e_1^I, A) = \{ \langle f_{j_1}^{j_2}, e_{i_1}^{i_2} \rangle \mid j_1, j_2, i_1, i_2 \text{ s.t.} \\ \forall (j,i) \in A : j_1 \leq j \leq j_2 \Leftrightarrow i_1 \leq i \leq i_2 \\ \wedge \exists (j,i) \in A : (j_1 \leq j \leq j_2 \wedge i_1 \leq i \leq i_2) \} ,\end{aligned} \tag{1}$$

where A is the alignment between the source and target sentences expressed as a set of position pairs.

2.2. Hierarchical Phrase-based Translation Model

The hierarchical approach [8] to machine translation is a generalization of the above model where the phrases are allowed to have "gaps". Those gaps are linked in the source and target language such that a translation rule specifies the location of the translation of the text filling a gap in the source side.

The model is formalized as a synchronous context-free grammar. The set of rules extracted from an aligned bilingual sentence pair is best described in a recursive way. Given a source sentence f_1^J, a target sentence e_1^I, an alignment A between them and N the maximum number of gaps allowed (usually $N = 2$), we can define the set of hierarchical phrases $\mathcal{H}(f_1^J, e_1^I, A)$ as

$$\mathcal{H}(f_1^J, e_1^I, A) = \bigcup_{n=0}^{N} \mathcal{H}_n(f_1^J, e_1^I, A)\,, \tag{2}$$

where the $\mathcal{H}_n$ are the subsets of hierarchical phrases with n gaps. For $n = 0$ the set $\mathcal{H}_0$ corresponds to the set of standard phrases given in Equation 1, but expressed in the form of rules of the grammar:

$$\begin{aligned}\mathcal{H}_0(f_1^J,&e_1^I, A) = \\ &\left\{X \rightarrow \left\langle f_{j_1}^{j_2}, e_{i_1}^{i_2})\right\rangle \mid j_1, j_2, i_1, i_2 \text{ s.t.}\right. \\ &\forall (j,i) \in A : j_1 \leq j \leq j_2 \Leftrightarrow i_1 \leq i \leq i_2 \\ &\left.\wedge\, \exists (j,i) \in A : (j_1 \leq j \leq j_2 \wedge i_1 \leq i \leq i_2)\right\}.\end{aligned} \tag{3}$$

We then proceed to define the following sets in a recursive manner. Denoting with $\mathcal{F}$ and $\mathcal{E}$ the vocabulary of the source and target languages respectively, and with $\mathcal{N}$ the set of non-terminals, we can define

$$\begin{aligned}\mathcal{H}_n(f_1^J,&e_1^I, A) = \\ &\left\{X \rightarrow \left\langle \alpha X^{\sim n} \beta, \delta X^{\sim n} \gamma\right\rangle \mid\right. \\ &\alpha, \beta \in (\mathcal{F} \cup \mathcal{N})^\star, \delta, \gamma \in (\mathcal{E} \cup \mathcal{N})^\star \\ &\wedge\, \exists j_1, j_2, i_1, i_2 : j_1 < j_2, i_1 < i_2 : \\ &\left(X \rightarrow \left\langle \alpha f_{j_1}^{j_2} \beta, \delta e_{i_1}^{i^2} \gamma\right\rangle \in \mathcal{H}_{n-1}(f_1^J, e_1^I, A)\right. \\ &\left.\left.\wedge X \rightarrow \left\langle f_{j_1}^{j_2}, e_{i_1}^{i_2}\right\rangle \in \mathcal{H}_0(f_1^J, e_1^I, A)\right)\right\},\end{aligned} \tag{4}$$

where the $^\sim$ denotes the relationship between non-terminals in the source and the target side.

The total set of hierarchical phrases extracted from a parallel corpus is the union of the hierarchical phrases extracted from each of its sentences. As can be seen from Equation 4, in the standard approach only one generic non-terminal is used. There are works which propose to extend the set of non-terminals, see e.g. [9].

It is common practice to include two additional rules to the set of hierarchical rules

$$S \rightarrow \left\langle S^{\sim 0} X^{\sim 1}, S^{\sim 0} X^{\sim 1}\right\rangle \tag{5}$$

$$S \rightarrow \left\langle X^{\sim 0}, X^{\sim 0}\right\rangle \tag{6}$$

where S is the initial symbol in the grammar. Rule (5), usually denoted as "glue rule", allows the concatenation of hierarchical phrases in a manner similar to monotonic phrase-based translation. Rule (6) allows the substitution of the initial symbol in the grammar with the generic non-terminal that allows the translation process to be carried out.

2.3. Common Features

Up to this point we have focused on the description of the structural part of the translation models, but we did not specify how to compute the probabilities associated with the translations. For both the phrase-based and the hierarchical-based models we define the probability using a log-linear model combination, as is standard practice in current state-of-the-art systems.

Given a source sentence f_1^J that is to be translated into a target sentence e_1^I, the translation probability is defined directly as

$$p(e_1^I | f_1^J) = \frac{\exp\left(\sum_{m=1}^{M} \lambda_m h_m(f_1^J, e_1^I)\right)}{\sum_{\tilde{e}_1^{\tilde{I}}} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(f_1^J, \tilde{e}_1^{\tilde{I}})\right)}\,. \tag{7}$$

In most state-of-the-art systems only an approximation to this equation is used. Instead of taking into account all the possible ways to generate a translation by using the basic unit of the models (phrases or hierarchical rules), we only consider the combination with the maximum probability. To formally define the probability we would need to add an additional variable to represent this set of units and how they are combined with each other. At this point, this would however only clutter the exposition.

The set of *feature functions* $h_m(f_1^J, e_1^I)$ is very similar in both models and is composed of

- Phrase translation probabilities in source-to-target and target-to-source directions. This is an estimation of the probabilities of the basic units in the models (simple phrases or hierarchical phrases). Standard practice is to estimate these probabilities as relative frequencies. In this work we will study alternative ways to compute these probabilities.
- IBM1-like word-based probabilities computed at the phrase level, also in source-to-target and target-to-source directions. These probabilities can be seen as a smoothing of the afore mentioned phrase-based probabilities. In the case of the hierarchical model, the non-terminals in the rules are simply ignored in the computation.

- Language model probabilities of the produced translation.
- Different penalties. These heuristic features help in controlling different aspects of the translation. A word penalty can guide the translation process into choosing longer or shorter translations, a penalty for rules not in the set $\mathcal{H}_0$ can favour hierarchical rules over lexical phrases, etc. The set of chosen features is dependent of the model used, but they are similar in spirit.
- Phrase count features which penalize phrases with low counts.

The values for the scaling factors λ_m are estimated by *minimum error rate* training, a numerical method that optimizes a measure of translation quality (usually BLEU) on a held-out development set [10].

3. Forced Alignments

The standard phrase extraction procedure defined in Equation 1 requires a word alignment to be provided. Normally this alignment is computed using probabilistic models, usually the word-based IBM translation models [11] as implemented in the GIZA++ toolkit [12], which are different to the ones later used in the translation process. This produces a mismatch between the phrase extraction procedure and the translation procedure, as they are based on different stochastic models, which are applied independently of each other. Other approaches have tried to bridge this mismatch between training and decoding, e.g. [1, 2, 3]. Recently, consistent improvements in translation quality could be achieved [4].

In this work we investigate a first approach to incorporate the techniques proposed in [4] in the hierarchical phrase-based approach. In this section we describe the training procedure and the model by which the phrase counts are computed, which we will later incorporate into the hierarchical translation system.

3.1. Forced Alignment for Phrase-Based Models

An illustration of the basic idea of the forced alignment training can be seen in Figure 1. The forced alignment procedure performs a phrase segmentation and alignment of each sentence pair of the training data using a modification of the translation decoder. To do this, the translation decoder is constrained to produce the reference translation for each bilingual sentence pair. No language model is used in search, as the target side is already given, but otherwise the set of models used is identical to unconstrained translation listed in Section 2.3.

Given a source and a target sentence, we search for the best segmentation and alignment that covers both sentences. I.e., we want to find a set of phrases and their disposition in order to maximize the probability defined by extending

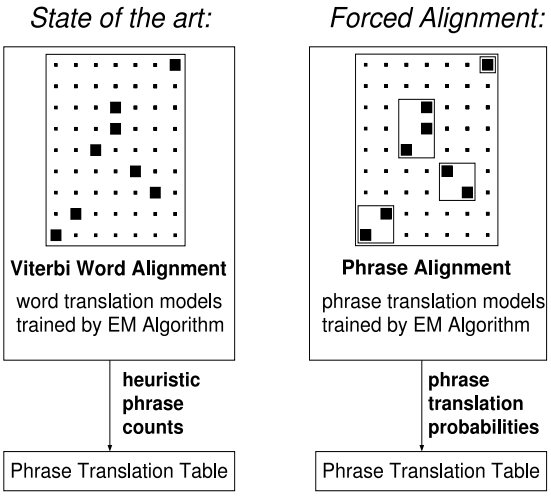

Figure 1: Illustration of phrase training with forced alignment.

Equation (7) to take the phrase segmentation into account, as noted in Section 2.3.

For efficiency, a phrase matching is performed on both source and target side before the search. Sentences for which the decoder cannot find an alignment are discarded for the phrase model training. In order to avoid overfitting, leaving-one-out is applied in search which modifies the phrase translation probabilities for each sentence pair. For a training example (f_n, e_n), we discount the occurrences of a given phrase in this sentence pair from the phrase counts obtained from the full training data. Let $C_n(\tilde{f}, \tilde{e})$ be the count for phrase $\langle \tilde{f}, \tilde{e} \rangle$, that were extracted from this sentence pair, and similarly the marginal counts $C_n(\tilde{e})$ and $C_n(\tilde{f})$. The resulting leaving-one-out phrase probability for training sentence pair n is

$$p_{l1o,n}(\tilde{f}|\tilde{e}) = \frac{C(\tilde{f}, \tilde{e}) - C_n(\tilde{f}, \tilde{e})}{C(\tilde{e}) - C_n(\tilde{e})} \,. \tag{8}$$

In order to avoid zero-probabilities (and thus possibly untranslated sentence pairs), phrase pairs for which the count would be reduced to zero are assigned a small probability close to zero. Here, we follow the length-based leaving-one-out strategy described in [4], setting the probability for singleton phrases to $\alpha = \beta^{(|\tilde{f}|+|\tilde{e}|)}$, with $\beta = e^{-5}$ where e denotes Euler's constant. This corresponds to a penalty of 5 in the logarithmic space. The exact value for β has proven inconsequential. For our experiments, one iteration of forced alignment was performed to produce the final phrase table.

3.2. Phrase Model

The phrase model we used corresponds to the *count model* described in [4]. From the n-best list we compute the phrase

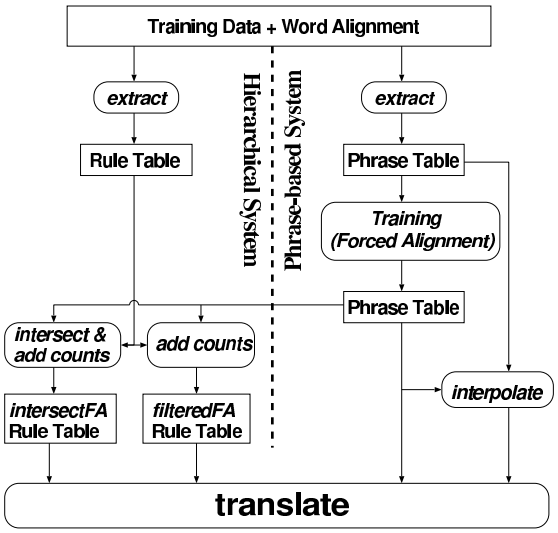

Figure 2: Experimental setup for forced alignment, including the original setup from [4] on the right-hand side and the setup proposed in this paper on the left-hand side.

Table 1: *Corpus Statistics on IWSLT 2010, BTEC*

		Arabic	English
Train:	Sentences	23,940	
	Running Words	206,008	240,125
	Vocabulary	15,861	8,258
	Singletons	7,152	3,516
test04:	Sentences	500	
	Running Words	3,537	–
	Vocabulary	1,183	–
	OOV rate	3%	–
test08:	Sentences	507	
	Running Words	3,478	–
	Vocabulary	1,141	–
	OOV rate	3.9%	–
test05:	Sentences	506	
	Running Words	3,421	–
	Vocabulary	1,185	–
	OOV rate	3.5%	–

counts $C_{FA}(\tilde{f},\tilde{e})$. Each item of the n-best list is weighted equally, and the size is set to $n = 100$. The translation probability of a phrase pair $(\tilde{f},\tilde{e})$ is estimated as

$$p_{FA}(\tilde{f}|\tilde{e}) = \frac{C_{FA}(\tilde{f},\tilde{e})}{\sum_{\tilde{f}'} C_{FA}(\tilde{f}',\tilde{e})}. \quad (9)$$

4. Combining Forced Alignments with Hierarchical Phrase-based Translation

There are several ways to use the estimated model from Section 3 in translation systems. The simplest method is to take the phrase table from forced alignments and use this directly as phrase table in the decoder. Another possiblity is to combine this table with the original phrase table.

In both [1] and [4] the authors show that a fixed log-linear interpolation of the phrase translation probabilities of the estimated model with the original model gives the best results. Inspired by this, we want to combine the forced alignment table with a hierarchical phrase table.

In order to perform an interpolation of the phrase tables containing different sets of phrase pairs, the intersection of both is retained. This approach cannot be adapted to the hierarchical model because an intersection would result in a phrase table without containing any hierarchical rules. This is not our intention, rather we would like to obtain a hierarchical model with improved translation probabilites. Therefore we filter the phrase table trained with forced alignment with the hierarchical phrases such that only rules are retained that are already known to the hierarchical system. The new probabilities are then estimated in the following way. First, for every phrase pair $(\tilde{f},\tilde{e})$, we compute the new count $C_M(\tilde{f},\tilde{e})$ by

$$C_M = C_{FA}(\tilde{f},\tilde{e}) + C_H(\tilde{f},\tilde{e}), \quad (10)$$

where $C_{FA}(\tilde{f},\tilde{e})$ is the count of the phrase pair in the filtered forced alignment table and $C_H(\tilde{f},\tilde{e})$ denotes the count from the original hierarchical phrase table. The same is done with the source and target marginals of all phrases, i.e., the counts from both tables are added up. In the end, the phrase probabilities can be estimated by renormalizing the counts as given in Equation (9). We will denote this method by *filteredFA*.

We also performed experiments with a different combination method which comes closer to what is reported in [4]. Instead of filtering the forced alignment table we intersect it with the non-hierarchical phrases from our hierarchical rules set. This set is then joined with the pure hierarchical phrases as follows:

$$\begin{aligned}\mathcal{H}_{FA} = \{X \rightarrow \langle\alpha,\delta\rangle \mid \alpha \in (\mathcal{F}\cup\mathcal{N})^\star, \delta \in (\mathcal{E}\cup\mathcal{N})^\star: \\ (X \rightarrow \langle\alpha,\delta\rangle \in \mathcal{H}_n \wedge n > 0), \\ \vee (X \rightarrow \langle\alpha,\delta\rangle \in \mathcal{H}_0 \wedge \langle\alpha,\beta\rangle \in \mathcal{P}_{FA})\},\end{aligned} \quad (11)$$

where $\mathcal{H}_n$ is the set of hierarchical rules as defined in Equations 3 and 4 and $\mathcal{P}_{FA}$ is the set of rules which we obtain from forced alignments.

When intersecting and joining phrase pairs, counts are added and probabilities are renormalized as described above. This method will be denoted by *intersectFA*. Note that the hierarchical and the phrase-based system are trained independently in all cases.

Table 2: *BLEU[%] results on IWSLT.*

	test04	test08	test05
hierarchical baseline	58.8	57.1	62.5
hierarchical+filteredFA	59.3	57.5	62.2
hierarchical+intersectFA	58.6	56.3	61.5
hierarchical+syntax	59.0	57.5	61.4
hierarchical+syntax+filteredFA	60.6	57.5	62.1

Table 3: *TER[%] results on IWSLT.*

	test04	test08	test05
hierarchical baseline	27.6	29.6	25.4
hierarchical+filteredFA	26.6	28.6	25.2
hierarchical+intersectFA	27.8	29.9	26.0
hierarchical+syntax	26.7	28.6	25.9
hierarchical+syntax+filteredFA	26.2	28.6	25.5

5. Experimental Results

We conducted our experiments on two different tasks. One is on the Arabic-English data published for the International Workshop on Spoken Language Translation (IWSLT) 2010 BTEC task. The other task is on German-English Europarl and news data.

We trained a phrase-based and a hierarchical statistical MT system (cf. [13] and [14]). The phrase-based system was used to train forced alignments as described in Section 3. The hierarchical system is used as baseline and combined with the retrained phrase table from the forced alignment training using the phrase-based system.

5.1. Results

First, we present our results on the IWSLT 2010 BTEC task for Arabic-to-English for which corpus statistics are given in Table 1. We give BLEU scores in Table 2 as well as TER scores in Table 3. We chose test04 as development set and test08 and test05 as blind test sets.

We experimented with a hierarchical baseline system and a hierarchical system enriched with soft syntactic labels [15]. The second will be simply denoted by the label *syntax* in our tables of results. The filtering method from Section 4 improves both the baseline and the system with syntax information, though improvements can be found only on one of the test sets each. The intersection method was only tested on the simple baseline and performed worse on all test sets.

In order to investigate our method on a larger corpus, we experimented on the English-to-German Quaero project corpus from 2010. This corpus mainly contains the data from the Workshop of Machine Translation (WMT) 2010, namely Europarl and news-commentary data. In addition to that, our corpus contains project internal data which is also from the news-commentary domain. Table 4 shows an overview of the corpus. The development and test sets are official WMT sets. Our development set is the combined nc-dev07 and nc-test06 and our blind test set is nc-test07.

Table 4: *Corpus Statistics on Quaero 2010*

		English	German
Train:	Sentences	1,799,293	
	Running Words	46,708,144	44,626,470
	Vocabulary	121,857	369,859
	Singletons	45,648	176,657
Dev:	Sentences	2,121	
	Running Words	51,343	52,946
	Vocabulary	7,313	9,863
	OOV rate	0.6%	1.1%
Test:	Sentences	2,007	
	Running Words	49,763	51,119
	Vocabulary	7,119	9,680
	OOV rate	0.5%	1.1%

Table 5: *BLEU[%] results for English-German news.*

	Dev	Test
hierarchical base	17.6	18.6
hierarchical filteredFA	17.8	19.0
phrase-based base	17.1	17.8
phrase-based filteredFA	17.3	18.4

Table 5 and Table 6 show phrase-based and hierarchical baselines as well as both systems using forced alignment. The improvement that forced alignments yield on the phrase-based system are not fully carried over to the hierarchical system. Still we find an improvement of 0.4 BLEU on the test set.

5.2. Translation Examples

In this section, we present some translation examples by comparing the hierarchical phrase-based baseline system to the one using forced alignments as presented in Section 4. Examples are taken from the test set of the English-to-German Quaero 2010 system.

Table 6: *TER[%] results for English-German news.*

	Dev	Test
hierarchical base	66.9	65.0
hierarchical filteredFA	66.9	64.6
phrase-based base	67.8	65.7
phrase-based filteredFA	66.4	64.5

Table 7: *Translation examples from the WMT nc-test07 set for English-German.*

source	This makes an increase in immigration unavoidable.
hierarchical	Dies ist ein Anstieg der Einwanderung unvermeidlich.
hierarchical+FA	Das macht eine zunehmende Einwanderung unvermeidlich.
reference	Schon das macht eine vermehrte Einwanderung unvermeidlich.
source	Saving Alstom by nationalizing the company is obviously wrong.
hierarchical	Die Rettung von Alstom durch das Unternehmen verstaatlicht ist offenkundig falsch.
hierarchical+FA	Die Rettung von Alstom durch Verstaatlichung des Unternehmens ist offenkundig falsch.
reference	Alstom durch die Nationalisierung des Unternehmens zu retten, ist offensichtlich falsch.
source	So it would be highly imprudent to extrapolate Bolivia's current crisis to the rest of Latin America.
hierarchical	Daher wäre es höchst unvorsichtig, rechnet die gegenwärtige Krise in Bolivien auf die restliche Lateinamerika.
hierarchical+FA	Daher wäre es höchst unvorsichtig zu extrapolieren die aktuelle Krise in Bolivien auf die restliche Lateinamerika.
reference	Es wäre also äußerst unvorsichtig, Boliviens aktuelle Krise auf das übrige Lateinamerika zu übertragen.

Table 7 shows three examples with improved translation quality. In the first sentence, the baseline system is missing a translation of the verb *makes/macht* which is present in the variant using forced alignments. The second example shows an improvement when translating the English phrase *by nationalizing the company*. In the baseline system, this is translated to *durch das Unternehmen verstaatlicht* which can be (back-)translated to *nationalized by the company*, i.e. a wrong wording is produced. The system using forced alignments generates the term *durch Verstaatlichung des Unternehmens* which is a correct translation. Note that for this sentence the reference provides a synonym as correct translation of *nationalizing*, namely *Nationalisierung* instead of *Verstaatlichung*. In the third example, the English infinitive *extrapolate* is mistranslated as *rechnet (calculates,expects)* in the baseline system. The system with forced alignments generates the correct *zu extrapolieren*.

6. Conclusion

We have shown that forced alignments trained for phrase-based systems can improve not only phrase-based systems themselves but also show a positive effect on hierarchical systems. Our proposed techniques yield improvements on the IWSLT 2010 task of up to 0.7% BLEU and up to 1.0% TER depending on the test set. Also on a large scale task as Quaero 2010 we gain up to 0.6% BLEU. Translation examples show that the forced alignments from the phrase-based decoder lead to better lexical choice during the decoding process of the hierarchical system.

Since our hierarchical translation systems with forced alignments contain the same phrase pairs as the hierarchical baseline system, only with different probabilites, we cannot report smaller and faster translation systems as done in [4]. However, this work is intended to report first experiments with forced alignments in hierarchical systems and shall stimulate research on training forced alignments directly with hierarchical phrases. Further, our method can be seen as a step towards a genuine combination of phrase-based and hierarchical translation systems that goes beyond standard system combination.

7. Acknowledgements

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation, and also partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-08-C-0110.

8. References

[1] J. DeNero, D. Gillick, J. Zhang, and D. Klein, "Why Generative Phrase Models Underperform Surface Heuristics," in *Proceedings of the Workshop on Statistical Machine Translation*, New York City, June 2006, pp. 31–38.

[2] P. Liang, A. Buchard-Côté, D. Klein, and B. Taskar, "An End-to-End Discriminative Approach to Machine Translation," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006, pp. 761–768.

[3] J. Ferrer and A. Juan, "A phrase-based hidden semi-markov approach to machine translation." in *Proceedings of European Association for Machine Translation (EAMT)*. Barcelona, Spain: European Association for Machine Translation, May 2009.

[4] J. Wuebker, A. Mauser, and H. Ney, "Training phrase translation models with leaving-one-out," in *Proceedings of the 48th Annual Meeting of the Assoc. for Computational Linguistics*, Uppsala, Sweden, July 2010, pp. 475–484.

[5] P. Blunsom, T. Cohn, and M. Osborne, "A discriminative latent variable model for statistical machine translation," in *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, June 2008, pp. 200–208.

[6] M. Cmejrek, B. Zhou, and B. Xiang, "Enriching SCFG Rules Directly From Efficient Bilingual Chart Parsing," in *Proc. of the International Workshop on Spoken Language Translation*, Tokyo, Japan, 2009, pp. 136–143.

[7] F. Och, C. Tillmann, and H. Ney, "Improved alignment models for statistical machine translation," in *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP99)*, University of Maryland, College Park, MD, USA, June 1999, pp. 20–28.

[8] D. Chiang, "A Hierarchical Phrase-Based Model for Statistical Machine Translation," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, Michigan, USA, June 2005, pp. 263–270.

[9] A. Zollmann and A. Venugopal, "Syntax Augmented Machine Translation Via Chart Parsing," in *Proceedings of the Workshop on Statistical Machine Translation*, New York City, New York, USA, June 2006, pp. 138–141.

[10] F. J. Och, "Minimum Error Rate Training for Statistical Machine Translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, July 2003, pp. 160–167.

[11] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–312, June 1993.

[12] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, Mar. 2003.

[13] R. Zens and H. Ney, "Improvements in dynamic programming beam search for phrase-based statistical machine translation," in *International Workshop on Spoken Language Translation*, Honolulu, Hawaii, Oct. 2008.

[14] D. Vilar, D. Stein, M. Huck, and H. Ney, "Jane: Open source hierarchical translation, extended with reordering and lexicon models," in *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, Uppsala, Sweden, July 2010, pp. 262–270.

[15] D. Vilar, D. Stein, and H. Ney, "Analysing soft syntax features and heuristics for hierarchical phrase based machine translation," in *International Workshop on Spoken Language Translation*, Honolulu, Hawaii, Oct. 2008, pp. 190–197.