Lexicon Models for Hierarchical Phrase-Based Machine Translation

Matthias Huck, Saab Mansour, Simon Wiesler, Hermann Ney

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

Abstract

In this paper, we investigate lexicon models for hierarchical phrase-based statistical machine translation. We study five types of lexicon models: a model which is extracted from word-aligned training data and—given the word alignment matrix—relies on pure relative frequencies [1]; the IBM model 1 lexicon [2]; a regularized version of IBM model 1; a triplet lexicon model variant [3]; and a discriminatively trained word lexicon model [4]. We explore source-to-target models with phrase-level as well as sentence-level scoring and target-to-source models with scoring on phrase level only. For the first two types of lexicon models, we compare several scoring variants. All models are used during search, i.e. they are incorporated directly into the log-linear model combination of the decoder.

Phrase table smoothing with triplet lexicon models and with discriminative word lexicons are novel contributions. We also propose a new regularization technique for IBM model 1 by means of the Kullback-Leibler divergence with the empirical unigram distribution as regularization term.

Experiments are carried out on the large-scale NIST Chinese→English translation task and on the English→French and Arabic→English IWSLT TED tasks. For Chinese→English and English→French, we obtain the best results by using the discriminative word lexicon to smooth our phrase tables.

1. Introduction

Lexical scoring on phrase level is the standard technique for phrase table smoothing in statistical machine translation [1, 5]. As most of the longer phrases appear only sparsely in the training data, their translation probabilities are overestimated when using relative frequencies to obtain conditional probabilities. One way to counteract overestimation of phrase pairs for which little evidence in the training data exists is to score phrases with word-based models and to interpolate these lexical probabilities with the phrase translation probabilities. Interpolation of the models is usually done log-linearly as part of the combination of feature functions of the translation system [6]. In this way the interpolation parameter can be tuned directly against the metric of translation quality, e.g. BLEU or TER, on a held-out development set.

Lexicon models in both source-to-target and target-to-source direction are thus a crucial component of state-of-the-art phrase-based systems, including hierarchical ones. Hierarchical SMT systems use a generalization of the standard phrase model where in addition to contiguous *lexical* phrases, *hierarchical* phrases with usually up to two gaps are extracted from the parallel training data [7]. The hierarchical phrase-based paradigm thus enables modeling of reorderings and long-distance dependencies in a consistent way.

In addition to phrase table smoothing, lexicon models are often applied on sentence level to rerank the n-best candidate translations of the decoder [8, 9, 10]. In reranking, the complete target sentence is available and the model can account for global sentence-level context to judge the selection of target words which was determined by the decoder. Both source-to-target and target-to-source models may be used.

Lexicon models in source-to-target direction are sometimes applied to score the target side of phrases given the whole source sentence during decoding already [4]. This can be accomplished quite efficiently since the given source sentence does not change. Phrase-level models, on the other hand, have the advantage that their scores do not have to be calculated on demand for each hypothesis expansion, but can be precomputed in advance and written to the phrase table.

Two of the models that we study in this paper, the triplet lexicon model and the discriminative word lexicon (DWL), have only been applied using sentence-level context before. For the DWL model, results in target-to-source direction have never been reported. We demonstrate that especially the DWL model performs very well on phrase level in both directions compared to the other types of lexicon models, and that limiting the context to phrase level does not harm translation quality in the hierarchical system.

While phrase table smoothing with the DWL model performs better as with IBM model 1 with respect to both metrics we use (BLEU and TER) on two of our three tasks, the conceptually appealing approach of extending IBM model 1 with a regularization term reduces the errors made by the system with regard to our secondary metric (TER) only. We show that the DWL model and both standard and regularized IBM model 1 clearly outperform the lexicon model which is extracted from word-aligned training data, though the latter one is probably most commonly used in setups reported in the literature.

2. Related Work

The well-known IBM model 1 lexicon was introduced by Brown et al. [2]. IBM model 1 is still employed within the widely used GIZA++ toolkit [11] as part of the word alignment training, which is the basis of modern phrase-based machine translation. Besides, it can be helpfull as an additional model in the log-linear combination or in n-best reranking [8, 9, 10]. Moore [12] suggested improvements to IBM model 1 parameter estimation, including an add-n smoothing technique which could be modeled within our IBM model 1 regularization framework. Recently, Toutanova and Galley [13] pointed out that the optimization problem for IBM model 1 is not strictly convex.

Word lexicon models extracted from the alignment have been proposed by Koehn, Och and Marcu [1] and Zens and Ney [5] and applied in their respective translation systems for phrase table smoothing. Foster et al. [14] compare several strategies for phrase table smoothing, including the former two. Chiang et al. [15] suggested morphology-based and provenance-based improvements to the Koehn-Och-Marcu method recently.

Hasan et al. [10] proposed triplet lexicon models for statistical machine translation for the first time and applied them in an n-best reranking framework. Hasan and Ney [3] investigated triplet lexicon scoring in a phrase-based decoder and compared the translation performance of triplet models applied in reranking to a direct application in search, Vilar et al. [16] integrated triplet as well as DWL models into a hierarchical decoder. Variants of discriminatively trained lexicon models have been utilized effectively within a phrase-based system [4], within a hierarchical system [17] and within a treelet translation system [18] before. The model we use is most similar to the one proposed by Mauser et al. [4]. It follows the approach described by Bangalore et al. [19].

3. Lexicon Models

We describe the source-to-target directions of the models in the following sections. The reverse models and scoring functions are computed similarly.

3.1. Word Lexicon from Word-Aligned Data

Given a word-aligned parallel training corpus, we are able to estimate single-word based translation probabilities $p_{\text{RF}}(e|f)$ by relative frequency.

Let $[f_1^{J_s}; e_1^{I_s}; \{a_{ij}\}_s]_s$, $1 \leq s \leq S$, be training samples of S word-aligned sentence pairs, where $\{a_{ij}\}_s$ denotes the alignment matrix of the s-th sentence pair. Let $j \in \{a_i\}$ express that f_j is aligned to the target word e_i.

We can now define (possibly fractional) counts

$$N_s(e,f) = \sum_{e_{i_s}:e_{i_s}=e} \sum_{f_{j_s}:f_{j_s}=f, j\in\{a_i\}_s} \frac{1}{|\{a_i\}_s|} \qquad (1)$$

for $1 \leq s \leq S$. If an occurence e_i of e has multiple aligned source words, each of the $|\{a_i\}| > 1$ alignment links contributes with a fractional count of $\frac{1}{|\{a_i\}|}$.

By summing over the whole corpus we obtain a count of aligned cooccurrences of target word e and source word f

$$N(e,f) = \sum_s N_s(e,f). \qquad (2)$$

The probabilities $p_{\text{RF}}(e|f)$ can then be computed as

$$p_{\text{RF}}(e|f) = \frac{N(e,f)}{\sum_{e'} N(e',f)}. \qquad (3)$$

This model is most similar to the one presented by Koehn et al. [1]. One difference we make is that we do not assume unaligned words to be aligned to the empty word (NULL). Probabilities with the empty word are thus not included in our lexicon. If scoring with the empty word is desired, we use a constant value of 0.05. The model does not comprise the discounting technique of Zens and Ney [5]. We are going to denote it as relative frequency (RF) word lexicon throughout this paper.

3.2. IBM Model 1

The IBM model 1 lexicon (IBM-1) is the first and most simple one in a sequence of probabilistic generative models [2]. The following assumptions are made for IBM-1: The target length I depends on the length J of the source sentence only, each target word is aligned to exactly one source word, the alignment of the target word depends on its absolute position and the sentence lengths only, and the target word depends on the aligned source word only. The alignment probability is in addition assumed to be uniform for IBM-1.

The probability of a target sentence e_1^I given a source sentence f_0^J (with $f_0 = $ NULL) can thus be written as

$$Pr(e_1^I|f_1^J) = \frac{1}{(J+1)^I} \prod_{i=1}^{I} \sum_{j=0}^{J} p_{\text{ibm1}}(e_i|f_j). \qquad (4)$$

The parameters of IBM-1 are estimated iteratively by means of the Expectation-Maximization (EM) algorithm [20] with maximum likelihood as training criterion.

3.3. Scoring Variants

Several methods to score phrase pairs with RF word lexicons or IBM-1 models have been suggested in the literature and are in common use. We apply and compare four of them.

In hierarchical phrase-based translation, we deal with rules $X \rightarrow \langle \alpha, \beta, ^\sim \rangle$ where $\langle \alpha, \beta \rangle$ is a bilingual phrase pair that may contain symbols from a non-terminal set, i.e. $\alpha \in (\mathcal{N} \cup V_F)^+$ and $\beta \in (\mathcal{N} \cup V_E)^+$, where V_F and V_E are the source and target vocabulary, respectively, and $\mathcal{N}$ is a non-terminal set which is shared by source and target. The left-hand side of the rule is a non-terminal symbol $X \in \mathcal{N}$, and the $^\sim$ relation denotes a one-to-one correspondence between the non-terminals in α and in β. For notational convenience, we define J_α to be the number of terminal symbols in α and I_β to be the number of terminal symbols in β. Indexing α with j, i.e. the symbol α_j, $1 \leq j \leq J_\alpha$, denotes the j-th terminal symbol on the source side of the phrase pair $\langle \alpha, \beta \rangle$, and analogous with β_i. $1 \leq i \leq I_\beta$, on the target side.

Our first scoring variant $t_{\text{Norm}}(\cdot)$ uses an IBM-1 or RF lexicon model $p(e|f)$ to rate the quality of a target side β given the source side α of a hierarchical rule with an included length normalization:

$$t_{\text{Norm}}(\alpha, \beta) = \sum_{i=1}^{I_\beta} \log \left(\frac{p(\beta_i|\text{NULL}) + \sum_{j=1}^{J_\alpha} p(\beta_i|\alpha_j))}{1 + J_\alpha} \right) \tag{5}$$

This variant has e.g. been used by Vilar et al. [16].

By dropping the length normalization we arrive at our second variant $t_{\text{NoNorm}}(\cdot)$:

$$t_{\text{NoNorm}}(\alpha, \beta) = \sum_{i=1}^{I_\beta} \log \left(p(\beta_i|\text{NULL}) + \sum_{j=1}^{J_\alpha} p(\beta_i|\alpha_j)) \right) \tag{6}$$

Among others, Mauser et al. [9] apply this variant in their standard phrase-based system.

Our third scoring variant $t_{\text{NoisyOr}}(\cdot)$ is the noisy-or model proposed by Zens and Ney [5]:

$$t_{\text{NoisyOr}}(\alpha, \beta) = \sum_{i=1}^{I_\beta} \log \left(1 - \prod_{j=1}^{J_\alpha} (1 - p(\beta_i|\alpha_j)) \right) \tag{7}$$

The fourth scoring variant $t_{\text{Moses}}(\cdot)$ is due to Koehn, Och and Marcu [1] and is the standard method in the open-source Moses system [21]:

$$\begin{aligned} t_{\text{Moses}}(\alpha, \beta, \{a_{ij}\}) \quad &= \qquad (8) \\ \sum_{i=1}^{I_\beta} \log & \left(\begin{cases} \frac{1}{|\{a_i\}|} \sum_{j \in \{a_i\}} p(\beta_i|\alpha_j)) & \text{if} \quad |\{a_i\}| > 0 \\ p(\beta_i|\text{NULL}) & \text{otherwise} \end{cases} \right) \end{aligned}$$

This last variant requires the availability of word alignments $\{a_{ij}\}$ for phrase pairs $\langle \alpha, \beta \rangle$. We store the most frequent alignment during phrase extraction and use it to compute $t_{\text{Moses}}(\cdot)$.

Note that all of these scoring methods generalize to hierarchical phrase pairs which may be only partially lexicalized. Unseen events are scored with a small floor value.

If not stated otherwise explicitly, we score with $t_{\text{Norm}}(\cdot)$ (Eq. (5)) in our experiments. Source-to-target sentence-level scores are calculated analogous to Eq. (5), but with the difference that the quality of the target side β of a rule currently chosen to expand a partial hypothesis is rated given the whole input sentence f_1^J instead of the source side α of the rule only.

3.4. Regularized IBM Model 1

Despite the wide use of the IBM model 1, basic modeling problems as non-strict convexity, overfitting and the use of heuristics for unseen events were not resolved algorithmically so far. We propose extending IBM-1 with the Kullback-Leibler (KL) divergence of the IBM-1 probabilities with respect to a smooth reference distribution p_{ref} as a regularization term:

$$\begin{aligned} r(p) &= \sum_f D_{\text{KL}}(p_{\text{ref}}(\cdot|f) \| p(\cdot|f)) \\ &= \sum_f \sum_e p_{\text{ref}}(e|f) \log \frac{p_{\text{ref}}(e|f)}{p(e|f)} \end{aligned} \tag{9}$$

For p_{ref} we choose the empirical unigram distribution

$$p_{\text{ref}}(e|f) = p(e) \ . \tag{10}$$

An advantage of the KL regularization term is that it can be easily integrated into the EM algorithm. Taking the derivative of the new auxiliary function which includes the regularization term, we obtain a weighted average of the reference distribution and the unregularized update as the EM update formula of the regularized IBM-1 model:

$$p(e|f) = \frac{1}{Z(f)} \big(\sum_s c_s(e|f) + C \cdot p_{\text{ref}}(e|f) \big) \ , \tag{11}$$

where

$$Z(f) = \sum_{e'} \sum_s c_s(e'|f) + C \ . \tag{12}$$

With s we denote the training samples, $c_s(e'|f)$ is the expected count of e' given f calculated exactly as in the original IBM-1 model, $C > 0$ denotes the regularization constant.

By using regularization, we gain two advantages: (i) over-fitting is avoided and training can be performed until "convergence"; (ii) the use of small probabilities for unseen events is not required anymore, and unseen event probabilities can be computed on the fly when the model is applied during decoding.

3.5. Triplet Lexicon

The triplet lexicon relies on triplets which are composed of two source language words triggering one target language word, i.e. it models probabilities $p_{\text{triplet}}(e|f, f')$. We use the *path-constrained* (or *path-aligned*) triplet model variant in this work. In the path-constrained triplet model, the first

Table 1: Data statistics for the preprocessed Chinese-English parallel training corpus.

	Chinese	English
Sentences	3.0M	
Running words	77.5M	81.0M
Vocabulary	83K	213K

trigger f is restricted to the aligned target word e. The second trigger f' is allowed to range over all remaining source words. Like IBM model 1, triplets are trained iteratively with the EM algorithm. We refer to Hasan et al. [10] for details about the path-constrained triplet model and the triplet training procedure.

With the same notational conventions as in Sections 3.3 and 3.1, we apply $t_{\text{Triplet}}(\cdot)$ to score a phrase pair with the path-constrained triplet lexicon model:

$$t_{\text{Triplet}}(\alpha, \beta, \{a_{ij}\}) = \sum_{i=1}^{I_\beta} \log \left(\frac{1}{Z_i} \sum_{j \in \{a_i\}} \sum_{j'=1}^{J_\alpha} p_{\text{triplet}}(\beta_i | \alpha_j, \alpha_{j'}) \right) \tag{13}$$

The double summation is normalized with $Z_i = J_\alpha \cdot |\{a_i\}|$. In fact, we score with NULL as a trigger as well. In favor of notational convenience, we omitted this in the formula.

3.6. Discriminative Word Lexicon

The discriminative word lexicon model acts as a classifier that predicts the words contained in the translation from the words given on the source side. The sequential order or any other structural interdependencies between the words on the source side as well as on the target side are ignored.

The model we use is very similar to the one of Mauser et al. [4], and we refer to their description for a more in-depth exposition. Our model differs in the training algorithm: we use the improved RProp+ algorithm [22] instead of the L-BFGS method. The scoring procedure has been transfered to phrase pairs. In our English→French and Arabic→English experiments, we employed sparse models comparable to the sparse DWLs presented by Huck et al. [17].

4. Experiments

We present empirical results obtained with the different lexicon models and scoring variants on the Chinese→English 2008 NIST task[1] as well as on the English→French and Arabic→English 2011 IWSLT TED tasks[2].

Table 2: Data statistics for the preprocessed English-French parallel training corpus.

	English	French
Sentences	2.0M	
Running words	54.3M	59.9M
Vocabulary	136K	159K

Table 3: Data statistics for the preprocessed Arabic-English parallel training corpus.

	Arabic	English
Sentences	89.8K	
Running words	1.6M	1.7M
Vocabulary	56.3K	34.0K

4.1. Hierarchical System

We employ the open source Jane toolkit [16] as a basis for our translation setups. The cube pruning algorithm [23] is used to carry out the search. For Arabic→English and English→French, we translate with a shallow grammar [24].

Word alignments are created by aligning the parallel training data in both directions with GIZA++ and applying the refined heuristic that was proposed by Och and Ney [11] on the two trained alignments to obtain a symmetrized alignment. The symmetrized alignment is used to compute the counts for the RF lexicon model, to train path-constrained triplets and to extract the phrase table. For language model training the SRILM toolkit [25] is utilized. We optimize the model weights against BLEU with standard Minimum Error Rate Training [26] on 100-best lists.

All our lexicon models are trained on the full parallel data, the DWL models have been pruned after training with a threshold of 0.01 for the Arabic→English task and 0.1 for the other two tasks, respectively. The IBM-1 models are produced with GIZA++. Phrase-level scores are precomputed and added to the phrase tables.

The performance of the systems is evaluated using the two metrics BLEU and TER. As BLEU is the optimized measure, TER mainly serves as an additional metric to verify the consistency of our improvements and avoid over-tuning. The results on the test sets are checked for statistical significance over the baseline. Confidence intervals have been computed using bootstrapping for BLEU and Cochran's approximate ratio variance for TER [27].

4.2. Chinese→English NIST Task

For the Chinese→English task we work with a parallel training corpus of 3.0M Chinese-English sentence pairs. The English target side of the data is lowercased, truecasing is part of the postprocessing pipeline. We employ MT06 as development set to tune the model weights, MT08 is used as unseen test set. Detailed statistics about the parallel training

[1] http://www.itl.nist.gov/iad/mig/tests/mt/2008/

[2] http://iwslt2011.anthropomatik.kit.edu/doku.php?id=06_evaluation

Table 4: Comparison of phrase table smoothing with different lexicon models for the NIST Chinese→English translation task (truecase). *s2t* denotes source-to-target scoring, *t2s* target-to-source scoring. The 95% confidence interval is given for the baseline system. Results in bold are significantly better than the baseline.

	MT06 (Dev)		MT08 (Test)	
NIST Chinese→English	BLEU[%]	TER[%]	BLEU[%]	TER[%]
Baseline 1 (no phrase table smoothing)	32.0	62.2	$24.3_{\pm 0.9}$	$67.8_{\pm 0.8}$
+ phrase-level s2t+t2s RF word lexicons	32.6	61.2	25.2	**66.6**
+ phrase-level s2t+t2s IBM-1	33.9	60.5	**26.7**	**65.6**
+ phrase-level s2t+t2s regularized IBM-1	33.7	60.2	**26.6**	**65.2**
+ phrase-level s2t+t2s path-constrained triplets	32.6	61.8	**25.5**	**66.7**
+ phrase-level s2t+t2s DWL	33.7	60.5	**27.0**	**65.6**

Table 5: Comparison of lexical scoring variants for the NIST Chinese→English translation task (truecase). *s2t* denotes source-to-target scoring, *t2s* target-to-source scoring. The 95% confidence interval is given for the baseline system. Results in bold are significantly better than the baseline.

	MT06 (Dev)		MT08 (Test)	
NIST Chinese→English	BLEU[%]	TER[%]	BLEU[%]	TER[%]
Baseline 1 (no phrase table smoothing)	32.0	62.2	$24.3_{\pm 0.9}$	$67.8_{\pm 0.8}$
+ phrase-level s2t+t2s RF word lexicons, Eq. (5): $t_{\text{Norm}}(\cdot)$	32.6	61.2	25.2	**66.6**
+ phrase-level s2t+t2s RF word lexicons, Eq. (6): $t_{\text{NoNorm}}(\cdot)$	32.7	61.8	**25.6**	**66.7**
+ phrase-level s2t+t2s RF word lexicons, Eq. (7): $t_{\text{NoisyOr}}(\cdot)$	32.4	61.2	**25.5**	**66.4**
+ phrase-level s2t+t2s RF word lexicons, Eq. (8): $t_{\text{Moses}}(\cdot)$	32.7	61.8	**25.4**	**66.9**
+ phrase-level s2t+t2s IBM-1, Eq. (5): $t_{\text{Norm}}(\cdot)$	33.9	60.5	**26.7**	**65.6**
+ phrase-level s2t+t2s IBM-1, Eq. (6): $t_{\text{NoNorm}}(\cdot)$	33.8	60.5	**26.6**	**65.7**
+ phrase-level s2t+t2s IBM-1, Eq. (7): $t_{\text{NoisyOr}}(\cdot)$	33.7	60.5	**26.7**	**66.0**
+ phrase-level s2t+t2s IBM-1, Eq. (8): $t_{\text{Moses}}(\cdot)$	33.2	61.3	**26.0**	**66.0**

Table 6: Results by adding sentence-level or phrase-level lexicon models in source-to-target or target-to-source direction to a standard baseline for the NIST Chinese→English translation task (truecase). *s2t* denotes source-to-target scoring, *t2s* target-to-source scoring. The 95% confidence interval is given for the baseline system. Results in bold are significantly better than the baseline.

	MT06 (Dev)		MT08 (Test)	
NIST Chinese→English	BLEU[%]	TER[%]	BLEU[%]	TER[%]
Baseline 2 (with s2t+t2s RF word lexicons)	32.6	61.2	$25.2_{\pm 0.8}$	$66.6_{\pm 0.7}$
+ sentence-level s2t IBM-1	32.9	61.6	25.7	66.6
+ sentence-level s2t path-constrained triplets	33.1	61.1	26.0	66.3
+ sentence-level s2t DWL	33.0	61.0	**26.2**	**65.5**
+ phrase-level s2t IBM-1	33.0	61.4	**26.4**	66.1
+ phrase-level s2t path-constrained triplets	33.1	61.3	26.0	66.3
+ phrase-level s2t DWL	33.4	61.3	**26.4**	66.3
+ phrase-level t2s IBM-1	33.4	60.7	**26.5**	**65.7**
+ phrase-level t2s path-constrained triplets	33.0	61.5	**26.3**	66.3
+ phrase-level t2s DWL	33.8	60.5	**26.5**	**65.7**
+ phrase-level s2t+t2s IBM-1	33.8	60.5	**26.9**	**65.4**
+ phrase-level s2t+t2s path-constrained triplets	33.3	61.3	**26.3**	66.1
+ phrase-level s2t+t2s DWL	34.0	60.2	**27.2**	**65.2**

Table 7: Comparison of phrase table smoothing with different lexicon models for the IWSLT English→French TED translation task (truecase). *s2t* denotes source-to-target scoring, *t2s* target-to-source scoring. The 95% confidence interval is given for the baseline system. Results in bold are significantly better than the baseline.

	Dev		Test	
IWSLT English→French	BLEU[%]	TER[%]	BLEU[%]	TER[%]
Baseline (no phrase table smoothing)	25.7	58.9	28.8$_{\pm 0.9}$	53.3$_{\pm 0.9}$
+ phrase-level s2t+t2s RF word lexicons	26.0	58.1	29.6	**51.8**
+ phrase-level s2t+t2s IBM-1	26.3	58.1	**30.0**	**52.0**
+ phrase-level s2t+t2s regularized IBM-1	26.3	57.9	**30.0**	**51.5**
+ phrase-level s2t+t2s path-constrained triplets	25.9	58.6	29.2	52.9
+ phrase-level s2t+t2s DWL	26.3	58.0	**30.2**	**51.8**

Table 8: Comparison of phrase table smoothing with different lexicon models for the IWSLT Arabic→English TED translation task (truecase). *s2t* denotes source-to-target scoring, *t2s* target-to-source scoring. The 95% confidence interval is given for the baseline system. Results in bold are significantly better than the baseline.

	Dev		Test	
IWSLT Arabic→English	BLEU[%]	TER[%]	BLEU[%]	TER[%]
Baseline (no phrase table smoothing)	25.0	56.6	23.6$_{\pm 0.9}$	59.3$_{\pm 1.0}$
+ phrase-level s2t+t2s RF word lexicons	26.3	55.0	**24.9**	**57.7**
+ phrase-level s2t+t2s IBM-1	26.9	54.0	**25.5**	**56.8**
+ phrase-level s2t+t2s regularized IBM-1	26.9	53.8	**25.3**	**56.8**
+ phrase-level s2t+t2s path-constrained triplets	26.0	55.4	**24.6**	**57.8**
+ phrase-level s2t+t2s DWL	27.1	53.7	**25.4**	**56.9**

data are given in Table 1. The language model is a 4-gram with modified Kneser-Ney smoothing which was trained on a large collection of monolingual data including the target side of the parallel corpus and the LDC Gigaword v3 corpus.

The empirical evaluation of all our Chinese→English setups is presented in Tables 4, 5 and 6. In the experiments shown in Table 4, we applied each one of the five types of lexicon models separately for phrase table smoothing—i.e. on phrase level in both translation directions—over a baseline that does not comprise any lexical features (*Baseline 1*). The impact of the scoring variant on the performance of RF word lexicons and IBM-1 models is examined in the series of experiments presented in Table 5. In Table 6, we took a standard setup including lexical smoothing with the RF word lexicon as a baseline (*Baseline 2*) to which we added IBM-1, path-constrained triplet and DWL models separately in either source-to-target direction or target-to-source direction or both. For the source-to-target direction, we also set up systems with sentence-level scoring for each of these three models.

Applying IBM-1 for phrase table smoothing brings about a considerably better result than resorting to lexical smoothing with the RF lexicon model (+1.5% BLEU / -1.0% TER). The regularized IBM-1 yields improvements over standard IBM-1 in TER only (-0.4% TER). Path-constrained triplets perform slightly better than the RF lexicon. The best phrase table smoothing result is obtained with the DWL model (+1.8% BLEU / -1.0% TER over the RF lexicon model and +0.3% BLEU over IBM-1).

For the RF word lexicon, scoring with $t_{\text{Norm}}(\cdot)$ is a bit worse than the other scoring variants. For IBM-1, $t_{\text{Moses}}(\cdot)$ does not perform very well, which could be explained by the fact that this scoring variant is little consistent with the training conditions of IBM-1.

Source-to-target sentence-level scoring is not better than phrase-level scoring in any of our experiments. Adding target-to-source triplet or DWL models to a standard baseline (*Baseline 2*), which was not done in any previous work, results in significantly better translations. The best hypotheses are produced with the system that includes phrase-level DWLs in both directions in addition to lexical smoothing with RF lexicon models (+2.0% BLEU / -1.4% TER over Baseline 2). Note that, though they perform worse in the phrase table smoothing experiments, RF lexicon models are still valuable in combination with IBM-1, triplet or DWL models.

4.3. English→French IWSLT Task

The parallel training data of our setups for the English→French IWSLT TED translation task is taken from TED talks, news-commentary and Europarl sources and totals to 2.0M sentence pairs. Training data statistics are given in Table 2. Our systems include a 4-gram language model which was trained using additional monolingual data, in particular a selection of French shuffled news.

Table 7 presents lexical smoothing results for the English→French task. IBM-1 outperforms RF word lexicons by +0.4% BLEU. As for the Chinese→English task, we attain the best BLEU result by smoothing the phrase table with DWL models (+0.2% BLEU / -0.2% TER over IBM-1) and the best TER result by smoothing with regularized IBM-1 (±0.0% BLEU / -0.5% TER over standard IBM-1).

4.4. Arabic→English IWSLT Task

We finally experimented with a small-scale setup for the Arabic→English IWSLT TED translation task. Here, we restrict the amount of parallel training data to the in-domain TED talks only. Table 3 contains statistics about the corpus we used. The 4-gram language model we employ was trained with a large amount of additional monolingual data from news-commentary, Europarl, UN and shuffled news sources.

Lexical smoothing results for the Arabic→English task are given in Table 8. IBM-1, regularized IBM-1 and DWL are again clearly better than the RF word lexicons (up to +0.6% BLEU / -0.9% TER). Unlike our findings for Chinese→English and English→French, DWL models do not yield improvements over IBM-1 here.

5. Discussion

Strictly speaking, our improvements over well-known models—more precisely, over source-to-target and target-to-source IBM-1 on phrase level—are rather small (e.g. up to +0.3% BLEU / -0.2% TER with DWL models instead of IBM-1 on top of Baseline 2 on the Chinese→English task). The potentially large gain by simply resorting to a stronger lexical smoothing method is however easily overlooked. As an example, phrase table smoothing with the method we found to perform weakest for the Chinese→English task—word lexicons obtained with relative frequencies from the word alignment and phrase scoring according to Eq. (5)—is the standard technique in the freely available Jane toolkit and has been applied by several system builders in their baseline setups. We thus do not only give a survey and a comparison of known as well as several novel lexical smoothing techniques in this paper, but also point out the weakness of established lexical feature functions that have been widely used in state-of-the-art systems.

6. Conclusion

We investigated five types of lexicon models in source-to-target and target-to-source direction with sentence-level or phrase-level context in a hierarchical phrase-based decoder. For triplet and discriminative word lexicon models, we presented a novel restriction to the phrase level. Restricting the scoring to phrase level has the advantage that the model scores can be precomputed and written to the phrase table. In our translation experiments on the Chinese→English NIST task, we were able to obtain the same or better results by phrase-level scoring as by considering sentence-level lexical context.

On three different translation tasks, we showed that phrase table smoothing with IBM model 1 or discriminative word lexicons clearly outperforms smoothing with lexicon models which are extracted from word-aligned training data. Furthermore, our novel lexical smoothing with DWL models yields improvements over IBM model 1 on two large-scale translation tasks for the Chinese-English and English-French language pairs. Our best Chinese→English system scores +2.0% BLEU / -1.4% TER better than a standard baseline.

We gave an empirical comparison of several commonly applied scoring variants. We finally suggested a regularization technique for IBM model 1 and evaluated it within our systems, obtaining reduced error rates with respect to TER.

7. Acknowledgments

This work was achieved as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

8. References

[1] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-Based Translation," in *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, Edmonton, Canada, May/June 2003, pp. 127–133.

[2] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, June 1993.

[3] S. Hasan and H. Ney, "Comparison of Extended Lexicon Models in Search and Rescoring for SMT," in *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, Boulder, CO, June 2009, pp. 17–20.

[4] A. Mauser, S. Hasan, and H. Ney, "Extending Statistical Machine Translation with Discriminative and Trigger-Based Lexicon Models," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Singapore, Aug. 2009, pp. 210–218.

[5] R. Zens and H. Ney, "Improvements in Phrase-Based Statistical Machine Translation," in *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, Boston, MA, May 2004, pp. 257–264.

[6] F. J. Och and H. Ney, "Discriminative Training and Maximum Entropy Models for Statistical Machine

Translation," in *Proc. of the 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Philadelphia, PA, July 2002, pp. 295–302.

[7] D. Chiang, "Hierarchical Phrase-Based Translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, June 2007.

[8] F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev, "A Smorgasbord of Features for Statistical Machine Translation," in *Proc. Human Language Technology Conf. / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL)*, Boston, MA, May 2004, pp. 161–168.

[9] A. Mauser, R. Zens, E. Matusov, S. Hasan, and H. Ney, "The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Kyoto, Japan, Nov. 2006, pp. 103–110.

[10] S. Hasan, J. Ganitkevitch, H. Ney, and J. Andrés-Ferrer, "Triplet Lexicon Models for Statistical Machine Translation," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Honolulu, Hawaii, Oct. 2008, pp. 372–381.

[11] 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.

[12] R. C. Moore, "Improving IBM Word-Alignment Model 1," in *Proc. of the 42nd Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Barcelona, Spain, July 2004, pp. 518–525.

[13] K. Toutanova and M. Galley, "Why Initialization Matters for IBM Model 1: Multiple Optima and Non-Strict Convexity," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Portland, OR, June 2011, pp. 461–466.

[14] G. Foster, R. Kuhn, and H. Johnson, "Phrasetable Smoothing for Statistical Machine Translation," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Sydney, Australia, July 2006, pp. 53–61.

[15] D. Chiang, S. DeNeefe, and M. Pust, "Two Easy Improvements to Lexical Weighting," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Portland, OR, June 2011, pp. 455–460.

[16] 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.

[17] M. Huck, M. Ratajczak, P. Lehnen, and H. Ney, "A Comparison of Various Types of Extended Lexicon Models for Statistical Machine Translation," in *Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, Denver, CO, Oct./Nov. 2010.

[18] M. Jeong, K. Toutanova, H. Suzuki, and C. Quirk, "A Discriminative Lexicon Model for Complex Morphology," in *Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, Denver, CO, Oct./Nov. 2010.

[19] S. Bangalore, P. Haffner, and S. Kanthak, "Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction," in *Proc. of the 45th Annual Meeting of the Assoc. of Computational Linguistics*, Prague, Czech Republic, June 2007, pp. 152–159.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statist. Soc. Ser. B*, vol. 39, no. 1, pp. 1–22, 1977.

[21] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, *et al.*, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Prague, Czech Republic, June 2007, pp. 177–180.

[22] C. Igel and M. Hüsken, "Empirical Evaluation of the Improved Rprop Learning Algorithms," *Neurocomputing*, vol. 50, pp. 105–123, 2003.

[23] L. Huang and D. Chiang, "Forest Rescoring: Faster Decoding with Integrated Language Models," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Prague, Czech Republic, June 2007, pp. 144–151.

[24] G. Iglesias, A. de Gispert, E. R. Banga, and W. Byrne, "Rule Filtering by Pattern for Efficient Hierarchical Translation," in *Proc. of the 12th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL)*, Athens, Greece, March 2009, pp. 380–388.

[25] A. Stolcke, "SRILM – an Extensible Language Modeling Toolkit," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, vol. 3, Denver, CO, Sept. 2002.

[26] F. J. Och, "Minimum Error Rate Training for Statistical Machine Translation," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Sapporo, Japan, July 2003, pp. 160–167.

[27] G. Leusch and H. Ney, "Edit Distances with Block Movements and Error Rate Confidence Estimates," *Machine Translation*, Dec. 2009.