

# Extending Statistical Machine Translation with Discriminative and Trigger-Based Lexicon Models

Arne Mauser and Saša Hasan and Hermann Ney  
Human Language Technology and Pattern Recognition Group  
Chair of Computer Science 6, RWTH Aachen University, Germany  
<surname>@cs.rwth-aachen.de

## Abstract

In this work, we propose two extensions of standard word lexicons in statistical machine translation: A discriminative word lexicon that uses sentence-level source information to predict the target words and a trigger-based lexicon model that extends IBM model 1 with a second trigger, allowing for a more fine-grained lexical choice of target words. The models capture dependencies that go beyond the scope of conventional SMT models such as phrase- and language models. We show that the models improve translation quality by 1% in BLEU over a competitive baseline on a large-scale task.

## 1 Introduction

Lexical dependencies modeled in standard phrase-based SMT are rather local. Even though the decision about the best translation is made on sentence level, phrase models and word lexicons usually do not take context beyond the phrase boundaries into account. This is especially problematic since the average source phrase length used during decoding is small. When translating Chinese to English, e.g., it is typically close to only two words.

The target language model is the only model that uses lexical context across phrase boundaries. It is a very important feature in the log-linear setup of today's phrase-based decoders. However, its context is typically limited to three to six words and it is not informed about the source sentence. In the presented models, we explicitly take advantage of sentence-level dependencies including the

source side and make non-local predictions for the target words. This is an important aspect when translating from languages like German and Chinese where long-distance dependencies are common. In Chinese, for example, tenses are often encoded by indicator words and particles whose position is relatively free in the sentence. In German, prefixes of verbs can be moved over long distances towards the end of the sentence.

In this work, we propose two models that can be categorized as extensions of standard word lexicons: A discriminative word lexicon that uses global, i.e. sentence-level source information to predict the target words using a statistical classifier and a trigger-based lexicon model that extends the well-known IBM model 1 (Brown et al., 1993) with a second trigger, allowing for a more fine-grained lexical choice of target words. The log-linear framework of the discriminative word lexicon offers a high degree of flexibility in the selection of features. Other sources of information such as syntax or morphology can be easily integrated.

The trigger-based lexicon model, or simply triplet model since it is based on word triplets, is not trained discriminatively but uses the classical maximum likelihood approach (MLE) instead. We train the triplets iteratively on a training corpus using the Expectation-Maximization (EM) algorithm. We will present how both models allow for a representation of topic-related sentence-level information which puts them close to word sense disambiguation (WSD) approaches. As will be shown later, the experiments indicate that these models help to ensure translation of content words that are often omitted by the baseline system. This is a common problem in Chinese-English translation. Furthermore, the models are often capable to produce a better lexical choice of content words.

The structure of the paper is as follows: In Section 2, we will address related work and briefly pin down how our models differentiate from previous work. Section 3 will describe the discriminative lexical selection model and the triplet model in more detail, explain the training procedures and show how the models are integrated into the decoder. The experimental setup and results will be given in Section 4. A more detailed discussion will be presented in Section 5. In the end, we conclude our findings and give an outlook for further research in Section 6.

## 2 Related Work

Several word lexicon models have emerged in the context of multilingual natural language processing. Some of them were used as a machine translation system or as a part of one such system. There are three major types of models: Heuristic models as in (Melamed, 2000), generative models as the IBM models (Brown et al., 1993) and discriminative models (Varea et al., 2001; Bangalore et al., 2006).

Similar to this work, the authors of (Varea et al., 2001) try to incorporate a maximum entropy lexicon model into an SMT system. They use the words and word classes from the local context as features and show improvements with  $n$ -best rescoring.

The models in this paper are also related to word sense disambiguation (WSD). For example, (Chan et al., 2007) trained a discriminative model for WSD using local but also across-sentence unigram collocations of words in order to refine phrase pair selection dynamically by incorporating scores from the WSD classifier. They showed improvements in translation quality in a hierarchical phrase-based translation system. Another WSD approach incorporating context-dependent phrasal translation lexicons is given in (Carpuat and Wu, 2007) and has been evaluated on several translation tasks. Our model differs from the latter in three ways. First, our approach models word selection of the target sentence based on global sentence-level features of the source sentence. Second, instead of disambiguating phrase senses as in (Carpuat and Wu, 2007), we model word selection independently of the phrases used in the MT models. Finally, the training is done in a different way as will be presented in Sections 3.1.1 and 3.2.1.

Recently, full translation models using discriminative training criteria emerged as well. They are designed to generate a translation for a given source sentence and not only score or disambiguate hypotheses given by a translation system. In (Ittycheriah and Roukos, 2007), the model can predict 1-to-many translations with gaps and uses words, morphologic and syntactic features from the local context.

The authors of (Venkatapathy and Bangalore, 2007) propose three different models. The first one is a global lexical selection model which includes all words of the source sentence as features, regardless of their position. Using these features, the system predicts the words that should be included in the target sentence. Sentence structure is then reconstructed using permutations of the generated bag of target words. We will also use this type of features in our model.

One of the simplest models in the context of lexical triggers is the IBM model 1 (Brown et al., 1993) which captures lexical dependencies between source and target words. It can be seen as a lexicon containing correspondents of translations of source and target words in a very broad sense since the pairs are trained on the full sentence level. The trigger-based lexicon model used in this work follows the training procedure introduced in (Hasan et al., 2008) and is integrated directly in the decoder instead of being applied in  $n$ -best list reranking. The model is very close to the IBM model 1 and can be seen as an extension of it by taking another word into the conditioning part, i.e. the triggering items. Thus, instead of  $p(f|e)$ , it models  $p(f|e, e')$ . Furthermore, since the second trigger can come from any part of the sentence, there is a link to long-range monolingual triggers as presented in (Tillmann and Ney, 1997) where a trigger language model was trained using the EM algorithm and helped to reduce perplexities and word error rates in a speech recognition experiment. In (Rosenfeld, 1996), another approach was chosen to model monolingual triggers using a maximum-entropy based framework. Again, this adapted LM could improve speech recognition performance significantly.

A comparison of a variant of the trigger-based lexicon model applied in decoding and  $n$ -best list reranking can be found in (Hasan and Ney, 2009). In order to reduce the number of overall triplets, the authors use the word alignments for fixing the

first trigger to the aligned target word. In general, this constraint performs slightly worse than the unconstrained variant used in this work, but allows for faster training and decoding.

### 3 Extended Lexicon Models

In this section, we present the extended lexicon models, how they are trained and integrated into the phrase-based decoder.

#### 3.1 Discriminative Lexicon Model

Discriminative models have been shown to outperform generative models on many natural language processing tasks. For machine translation, however, the adaptation of these methods is difficult due to the large space of possible translations and the size of the training data that has to be used to achieve significant improvements.

In this section, we propose a discriminative word lexicon model that follows (Bangalore et al., 2007) and integrate it into the standard phrase-based machine translation approach.

The core of our model is a classifier that predicts target words, given the words of the source sentence. The structure of source as well as target sentence is neglected in this model. We do not make any assumptions about the location of the words in the sentence. This is useful in many cases, as words and morphology can depend on information given at other positions in the sentence. An example would be the character 了 in Chinese that indicates a completed or past action and does not need to appear close to the verb.

We model the probability of the set of target words in a sentence  $e$  given the set of source words  $\mathbf{f}$ . For each word in the target vocabulary, we can calculate a probability for being or not being included in the set. The probability of the whole set then is the product over the entire target vocabulary  $\mathbf{V}_E$ :

$$P(e|\mathbf{f}) = \prod_{e \in e} P(e^+|\mathbf{f}) \cdot \prod_{e \in \mathbf{V}_E \setminus e} P(e^-|\mathbf{f}) \quad (1)$$

For notational simplicity, we use the event  $e^+$  when the target word  $e$  is included in the target sentence and  $e^-$  if not. We model the individual factors  $p(e|\mathbf{f})$  of the probability in Eq. 1 as a log-linear model using the source words from  $\mathbf{f}$  as binary features

$$\phi(f, \mathbf{f}) = \begin{cases} 1 & \text{if } f \in \mathbf{f} \\ 0 & \text{else} \end{cases} \quad (2)$$

and feature weights  $\lambda_{f,:}$ :

$$P(e^+|\mathbf{f}) = \frac{\exp\left(\sum_{f \in \mathbf{f}} \lambda_{f,e^+} \phi(f, \mathbf{f})\right)}{\sum_{e \in \{e^+, e^-\}} \exp\left(\sum_{f \in \mathbf{f}} \lambda_{f,e} \phi(f, \mathbf{f})\right)} \quad (3)$$

Subsequently, we will call this model discriminative word lexicon (DWL).

Modeling the lexicon on sets and not on sequences has two reasons. Phrase-based MT along with  $n$ -gram language models is strong at predicting sequences but only uses information from a local context. By using global features and predicting words in a non-local fashion, we can augment the strong local decisions from the phrase-based systems with sentence-level information.

For practical reasons, translating from a set to a set simplifies the parallelization of the training procedure. The classifiers for the target words can be trained separately as explained in the following section.

##### 3.1.1 Training

Common classification tasks have a relatively small number of classes. In our case, the number of classes is the size of the target vocabulary. For large translation tasks, this is in the range of a hundred thousand classes. It is far from what conventional out-of-the-box classifiers can handle.

The discriminative word lexicon model has the convenient property that we can train a separate model for each target word making parallelization straightforward. Discussions about possible classifiers and the choice of regularization can be found in (Bangalore et al., 2007). We used the freely available MegaM Toolkit<sup>1</sup> for training, which implements the L-BFGS method (Byrd et al., 1995). Regularization is done using Gaussian priors. We performed 100 iterations of the training algorithm for each word in the target vocabulary. This results in a large number of classifiers to be trained. For the Arabic-English data (cf. Section 4), the training took an average of 38 seconds per word. No feature cutoff was used.

##### 3.1.2 Decoding

In search, we compute the model probabilities as an additional model in the log-linear model combination of the phrase-based translation approach. To reduce the memory footprint and startup time of the decoding process, we reduced the number of

<sup>1</sup><http://www.cs.utah.edu/~hal/megam/>

parameters by keeping only large values  $\lambda_{f,e}$  since smaller values tend to have less effect on the overall probability. In experiments we determined that we could safely reduce the size of the final model by a factor of ten without losing predictive power. In search, we compute the model probabilities as an additional model in the log-linear combination. When scoring hypotheses from the phrase-based system, we see the translation hypothesis as the set of target words that are predicted. Words from the target vocabulary which are not included in the hypothesis are not part of the set. During the search process, however, we also have to score incomplete hypotheses where we do not know which words will not be included. This problem is circumvented by rewriting Eq. 1 as

$$P(\mathbf{e}|\mathbf{f}) = \prod_{e \in \mathbf{V}_E} P(e^-|\mathbf{f}) \cdot \prod_{e \in \mathbf{e}} \frac{P(e^+|\mathbf{f})}{P(e^-|\mathbf{f})}.$$

The first product is constant given a source sentence and therefore does not affect the search. Using the model assumption from Eq. 3, we can further simplify the computation and compute the model score entirely in log-space which is numerically stable even for large vocabularies. Experiments showed that using only the first factor of Eq. 1 is sufficient to obtain good results.

In comparison with the translation model from (Bangalore et al., 2007) where a threshold on the probability is used to determine which words are included in the target sentence, our approach relies on the phrase model to generate translation candidates. This has several advantages: The length of the translation is determined by the phrase model. Words occurring multiple times in the translation do not have to be explicitly modeled. In (Bangalore et al., 2007), repeated target words are treated as distinct classes.

The main advantage of the integration being done in a way as presented here is that the phrase model and the discriminative word lexicon model are complementary in the way they model the translation. While the phrase model is good in predicting translations in a local context, the discriminative word lexicon model is able to predict global aspects of the sentence like tense or vocabulary changes in questions. While the phrase model is closely tied to the structure of word and phrase alignments, the discriminative word lexicon model completely disregards the structure in source and target sentences.

## 3.2 Trigger-based Lexicon Model

The triplets of the trigger-based lexicon model, i.e.  $p(e|f, f')$ , are composed of two words in the source language triggering one target language word. We chose this inverse direction since it can be integrated directly into the decoder and, thus, does not rely on a two-pass approach using reranking, as it is the case for (Hasan et al., 2008). The triggers can originate from words of the whole source sentence, also crossing phrase boundaries of the conventional bilingual phrase pairs. The model is symmetric though, meaning that the order of the triggers is not relevant, i.e.  $(f, f' \rightarrow e) = (f', f \rightarrow e)$ . Nevertheless, the model is able to capture long-distance effects such as verb splits or adjustments to lexical choice of the target word given the topic-triggers of the source sentence. In training, we determine the probability of a target sentence  $e_1^I$  given the source sentence  $f_1^J$  within the model by

$$\begin{aligned} p(e_1^I | f_1^J) &= \prod_{i=1}^I p(e_i | f_1^J) \\ &= \prod_{i=1}^I \frac{2}{J(J+1)} \sum_{j=0}^J \sum_{j'=j+1}^J p(e_i | f_j, f_{j'}), \end{aligned} \quad (4)$$

where  $f_0$  denotes the empty word and, thus, for  $f_j = \varepsilon$ , allows for modeling the conventional (inverse) IBM model 1 lexical probabilities as well. Since the second trigger  $f_{j'}$  always starts right of the current first trigger, the model is symmetric and does not need to look at all trigger pairs. Eq. 4 is used in the iterative EM training on all sentence pairs of the training data which is described in more detail in the following.

### 3.2.1 Training

For training the trigger-based lexicon model, we apply the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). The goal is to maximize the log-likelihood  $F_{trip}$  of this model for a given bilingual training corpus  $\{(f_1^{J_n}, e_1^{I_n})\}_1^N$  consisting of  $N$  sentence pairs:

$$F_{trip} := \sum_{n=1}^N \log p(e_1^{I_n} | f_1^{J_n}),$$

where  $I_n$  and  $J_n$  are the lengths of the  $n$ -th target and source sentence, respectively. An auxiliary function  $Q(\mu; \bar{\mu})$  is defined based on  $F_{trip}$

where  $\bar{\mu}$  is the updated estimate within an iteration which is to be derived from the current estimate  $\mu$ . Here,  $\mu$  stands for the entire set of model parameters, i.e. the set of all  $\{\alpha(e|f, f')\}$  with the constraint  $\sum_e \alpha(e|f, f') = 1$ . The accumulators  $\alpha(\cdot)$  are therefore iteratively trained on the training data by using the current estimate, i.e. deriving the expected value (E-step), and maximizing their likelihood afterwards to reestimate the distribution. Thus, the perplexity of the training data is reduced in each iteration.

### 3.2.2 Decoding

In search, we can apply this model directly when scoring bilingual phrase pairs. Given a trained model for  $p(e|f, f')$ , we compute the feature score  $h_{trip}(\cdot)$  of a phrase pair  $(\tilde{e}, \tilde{f})$  as

$$h_{trip}(\tilde{e}, \tilde{f}, f_0^J) = - \sum_i \log \left( \frac{2}{J \cdot (J+1)} \sum_j \sum_{j' > j} p(\tilde{e}_i | f_j, f_{j'}) \right), \quad (5)$$

where  $i$  moves over all target words in the phrase  $\tilde{e}$ , the second sum selects all source sentence words  $f_0^J$  including the empty word, and  $j' > j$  incorporates the rest of the source sentence right of the first trigger. We take negative log-probabilities and normalize to obtain the final score (representing costs) for the given phrase pair. Note that in search, we can only use this direction,  $p(e|f, f')$ , since the whole source sentence is available for triggering effects whereas not all target words have been generated so far, as it would be necessary for the standard direction,  $p(f|e, e')$ .

Due to the enormous number of triplets, we trained the model on a subset of the overall training data. The subcorpus, mainly consisting of newswire articles, contained 1.4M sentence pairs with 32.3M running words on the English side. We trained two versions of the triplet lexicon, one using 4 EM iterations and another one that was trained for 10 EM iterations. Due to trimming of triplets with small probabilities after each iteration, the version based on 10 iterations was slightly smaller, having 164 million triplets but also performed slightly worse. Thus, for the experiments, we used the version based on 4 iterations which contained 291 million triplets.

Note that decoding with this model can be quite efficient if caching is applied. Since the given source sentence does not change, we have to calculate  $p(e|f, f')$  for each  $e$  only once and can re-

	train (C/E)	test08 (NW/WT)	
Sent. pairs	9.1M	480	490
Run. words	259M/300M	14.8K	12.3K
Vocabulary	357K/627K	3.6K	3.2K

Table 1: GALE Chinese-English corpus statistics including two test sets: newswire and web text.

	train C/E — A/E		nist08 C/A
Sent. pairs	7.3M	4.6M	1357
Words (M)	185/196	142/139	36K/46K
Vocab. (K)	163/265	351/361	6.4K/9.6K

Table 2: NIST Chinese-English and Arabic-English corpus statistics including the official 2008 test sets.

trieve the probabilities from the cache for consecutive scorings of the same target word  $e$ . This significantly speeds up the decoding process.

## 4 Experimental Evaluation

In this section we evaluate our lexicon models on the GALE Chinese-English task for newswire and web text translation and additionally on the official NIST 2008 task for both Chinese-English and Arabic-English. The baseline system was built using a state-of-the-art phrase-based MT system (Zens and Ney, 2008). We use the standard set of models with phrase translation probabilities for source-to-target and target-to-source direction, smoothing with lexical weights, a word and phrase penalty, distance-based and lexicalized reordering and a 5-gram (GALE) or 6-gram (NIST) target language model.

We used training data provided by the Linguistic Data Consortium (LDC) consisting of 9.1M parallel Chinese-English sentence pairs of various domains for GALE (cf. Table 1) and smaller amounts of data for the NIST systems (cf. Table 2). The DWL and Triplet models were integrated into the decoder as presented in Section 3.

For the GALE development and test set, we separated the newswire and web text parts and did separate parameter tuning for each genre using the corresponding development set which consists of 485 sentences for newswire texts and 533 sentences of web text. The test set has 480 sentences for newswire and 490 sentences for web text. For NIST, we tuned on the official 2006 eval set and used the 2008 evaluation set as a blind test set.

GALE test08	NW		WT	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Baseline	32.3	59.38	25.3	64.40
DWL	33.1	58.90	26.2	63.75
Triplet	32.9	58.59	26.2	64.20
DWL+Trip.	33.3	58.23	26.3	63.87

Table 3: Results on the GALE Chinese-English test set for the newswire and web text setting (case-insensitive evaluation).

#### 4.1 Translation Results

The translation results on the two GALE test sets are shown in Table 3 for newswire and web text. Both the discriminative word lexicon and the triplet lexicon can individually improve the baseline by approximately +0.6–0.9% BLEU and -0.5–0.8% TER. For the combination of both lexicons on the newswire setting, we observe only a slight improvement on BLEU but also an additional boost in TER reduction, arriving at +1% BLEU and -1.2% TER. For web text, the findings are similar: The combination of the discriminative and trigger-based lexicons yields +1% BLEU and decreases TER by -0.5%.

We compared these results against an inverse IBM model 1 but the results were inconclusive which is consistent with the results presented in (Och et al., 2004) where no improvements were achieved using  $p(e|f)$ . In our case, inverse IBM1 improves results by 0.2–0.4% BLEU on the development set but does not show the same trend on the test sets. Furthermore, combining IBM1 with DWL or Triplets often even degraded the translation results, e.g. only 32.8% BLEU was achieved on newswire for a combination of the IBM1, DWL and Triplet model. In contrast, combinations of the DWL and Triplet model did not degrade performance and could benefit from each other.

In addition to the automatic scoring, we also did a randomized subjective evaluation where the hypotheses of the baseline was compared against the hypotheses generated using the discriminative word lexicon and triplet models. We evaluated 200 sentences from newswire and web text. In 80% of the evaluated sentences, the improved models were judged equal or better than the baseline.

We tested the presented lexicon models also on another large-scale system, i.e. NIST, for two lan-

NIST nist08	Chinese-Eng.		Arabic-Eng.	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Baseline	26.8	65.11	42.0	50.55
DWL	27.6	63.56	42.4	50.01
Triplet	27.7	63.60	42.9	49.76
DWL+Trip.	27.9	63.56	43.0	49.15

Table 4: Results on the test sets for the NIST 2008 Chinese-English and Arabic-English task (case-insensitive evaluation).

guage pairs, namely Chinese-English and Arabic-English. Interestingly, the results obtained for Arabic-English are similar to the findings for Chinese-English, as can be seen in Table 4. The overall improvements for this language pair are +1% BLEU and -1.4% TER. In contrast to the GALE Chinese-English task, the triplet lexicon model for the Arabic-English language pair performs slightly better than the discriminative word lexicon.

These results strengthen the claim that the presented models are capable of improving lexical choice of the MT system. In the next section, we discuss the observed effects and analyze our results in more detail.

## 5 Discussion

In terms of automatic evaluation measures, the results indicate that it is helpful to incorporate the extended lexicon models into the search process. In this section, we will analyze some more details of the models and take a look at the lexical choice they make and what differentiates them from the baseline models. In Table 5, we picked an example sentence from the GALE newswire test set and show the different hypotheses produced by our system. As can be seen, the baseline does not produce the present participle of the verb *restore* which makes the sentence somewhat hard to understand. Both the discriminative and the trigger-based lexicon approach are capable of generating this missing information, i.e. the correct use of *restoring*. Figure 1 gives an example how discontinuous triggers affect the word choice on the target side. Two cases are depicted where high probabilities of triplets including *emergency* and *restoring* on the target side influence the overall hypothesis selection. The non-local modeling advantages of the triplet model can be observed as well: The

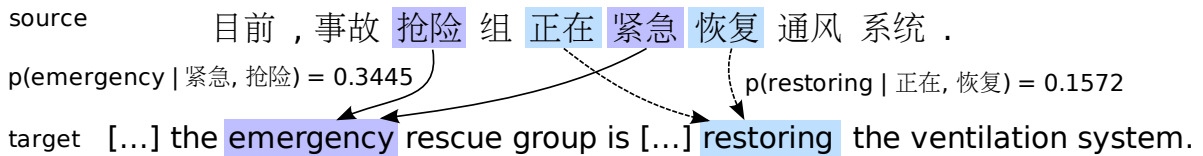


Figure 1: Triggering effect for the example sentence using the triplet lexicon model. The Chinese source sentence is shown in its segmented form. Two triplets are highlighted that have high probability and favor the target words *emergency* and *restoring*.

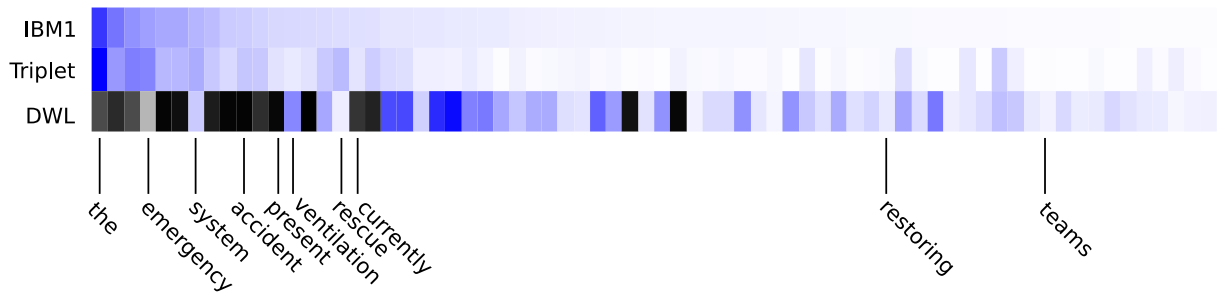


Figure 2: Ranking of words for the example sentence for IBM1, Triplet and DWL model. Ranks are sorted at IBM1, darker colors indicate higher probabilities within the model.

triggering events do not need to be located next to each other or within a given phrase pair. They move across the whole source sentence, thus allowing for capturing of long-range dependencies.

Table 6 shows the top ten content words that are predicted by the two models, discriminative word lexicon and triplet lexicon model. IBM model 1 ranks are indicated by subscripts in the column of the triplet model. Although the triplet model is similar to IBM1, we observe differences in the word lists. Comparing this to the visualization of the probability distribution for the example sentence, cf. Figure 2, we argue that, although the IBM1 and Triplet distributions look similar, the triplet model is sharper and favors words such as the ones in Table 6, resulting in different word choice in the translation process. In contrast, the DWL approach gives more distinct probabilities, selecting content words that are not chosen by the other models.

Table 7 shows an example from the web text test set. Here, the baseline hypothesis contains an incorrect word, *anna*, which might have been mistaken for the name *ying*. Interestingly, the hypotheses of the DWL lexicon and the combination of DWL and Triplet contain the correct content word *remarks*. The triplet model makes an error by selecting *music*, an artifact that might come from words that co-occur frequently with the cor-

responding Chinese verb *to listen*, i.e. 听, in the data. Although the TER score of the baseline is better than the one for the alternative models for this particular example, we still think that the observed effects show how our models help producing different hypotheses that might lead to subjectively better translations.

An Arabic-English translation example is shown in Table 8. Here, the term *incidents of murder in apartments* was chosen over the baseline’s *killings inside the flats*. Both translations are understandable and the difference in the wording is only based on synonyms. The translation using the discriminative and trigger-based lexicons better matches the reference translation and, thus, reflects a better lexical choice of the content words.

## 6 Conclusion

We have presented two lexicon models that use global source sentence context and are capable of predicting context-specific target words. The models have been directly integrated into the decoder and have shown to improve the translation quality of a state-of-the-art phrase-based machine translation system. The first model was a discriminative word lexicon that uses sentence-level features to predict if a word from the target vocabulary should be included in the translation or not. The second model was a trigger-based lexi-

Source	目前,事故抢险组正在紧急恢复通风系统.
Baseline	at present, the accident and rescue teams are currently emergency recovery ventilation systems.
DWL	at present, the emergency rescue teams are currently restoring the ventilation system.
Triplet	at present, the emergency rescue group is in the process of restoring the ventilation system.
DWL +Triplet	at present, the accident emergency rescue teams are currently restoring the ventilation system.
Reference	right now, the accident emergency rescue team is making emergency repair on the ventilation system.

Table 5: Translation example from the GALE newswire test set, comparing the baseline and the extended lexicon models given a reference translation. The Chinese source sentence is presented in its segmented form.

con that uses triplets to model long-range dependencies in the data. The source word triggers can move across the whole sentence and capture the topic of the sentence and incorporate more fine-grained lexical choice of the target words within the decoder.

Overall improvements are up to +1% in BLEU and -1.5% in TER on large-scale systems for Chinese-English and Arabic-English. Compared to the inverse IBM model 1 which did not yield consistent improvements, the presented models are valuable additional features in a phrase-based statistical machine translation system. We will test this setup for other language pairs and expect that languages like German where long-distance effects are common can benefit from these extended lexicon models.

In future work, we plan to extend the discriminative word lexicon model in two directions: extending context to the document level and feature engineering. For the trigger-based model, we plan to investigate more model variants. It might be interesting to look at cross-lingual trigger models such as  $p(f|e, f')$  or constrained variants like  $p(f|e, e')$  with  $pos(e') < pos(e)$ , i.e. the second trigger coming from the left context within a sentence which has already been generated. These

DWL		Triplet	
emergency	0.894	emergency <sub>1</sub>	0.048
currently	0.330	system <sub>2</sub>	0.032
current	0.175	rescue <sub>8</sub>	0.027
emergencies	0.133	accident <sub>3</sub>	0.022
present	0.133	ventilation <sub>7</sub>	0.021
accident	0.119	work <sub>33</sub>	0.021
recovery	0.053	present <sub>5</sub>	0.011
group	0.046	currently <sub>9</sub>	0.010
dealing	0.042	rush <sub>60</sub>	0.010
ventilation	0.034	restoration <sub>31</sub>	0.009

Table 6: The top 10 content words predicted by each model for the GALE newswire example sentence. Original ranks for the related IBM model 1 are given as subscripts for the triplet model.

Source	我听了莹的话,乐得哈哈大笑.
Baseline	i have listened to anna, happy and laugh.
DWL	i have listened to the remarks, happy and laugh.
Triplet	i have listened to the music, a roar of laughter.
DWL +Triplet	i have listened to the remarks, happy and laugh.
Reference	hearing ying's remark, i laughed aloud happily.

Table 7: Translation example from the GALE web text test set. In this case, the baseline has a better TER but we can observe a corrected content word (*remark*) for the extended lexicon models. The Chinese source sentence is shown in its segmented form.

extensions could be integrated directly in search as well and would enable the system to combine both directions (standard and inverse) to some extent which was previously shown to help when applying the standard direction  $p(f|e, e')$  as an additional reranking step, cf. (Hasan and Ney, 2009).

## Acknowledgments

This material is partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023, and was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

The authors would like to thank Christian Buck



Source	و كانت بعض الصحف السعودية قد نشرت عددا من الحالات التي تعرضت ل السجن دون مبرر وكذلك بعض حوادث القتل داخل الشقق و غير ها .
Baseline	some saudi newspapers have published a number of cases that had been subjected to imprisonment without justification, as well as some killings inside the flats and others.
DWL +Triplet	some of the saudi newspapers have published a number of cases which were subjected to imprisonment without justification, as well as some incidents of murder in apartments and others.
Reference	some saudi newspapers have published a number of cases in which people were unjustifiably imprisoned, as well as some incidents of murder in apartments and elsewhere.

Table 8: Translation example from the NIST Arabic-English test set. The DWL and Triplet models improve lexical word choice by favoring *incidents of murder in apartments* instead of *killings inside the flats*. The Arabic source is shown in its segmented form.

and Juri Ganitkevitch for their help training the extended lexicon models.

## References

- S. Bangalore, P. Haffner, and S. Kanthak. 2006. Sequence classification for machine translation. In *Ninth International Conf. on Spoken Language Processing, Interspeech 2006 — ICSLP*, pages 1722–1725, Pittsburgh, PA, September.
- S. Bangalore, P. Haffner, and S. Kanthak. 2007. Statistical machine translation through global lexical selection and sentence reconstruction. In *45th Annual Meeting of the Association of Computational Linguistics*, pages 152–159, Prague, Czech Republic, June.
- P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–312, June.
- R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.
- M. Carpuat and D. Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Prague, Czech Republic, June.
- Y. S. Chan, H. T. Ng, and D. Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic, June.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–22.
- S. Hasan and H. Ney. 2009. Comparison of extended lexicon models in search and rescoring for SMT. In *NAACL HLT 2009, Companion Volume: Short Papers*, pages 17–20, Boulder, Colorado, June.
- S. Hasan, J. Ganitkevitch, H. Ney, and J. Andrés-Ferrer. 2008. Triplet lexicon models for statistical machine translation. In *EMNLP*, pages 372–381, Honolulu, Hawaii, October.
- A. Ittycheriah and S. Roukos. 2007. Direct translation model 2. In *HLT-NAACL 2007: Main Conference*, pages 57–64, Rochester, New York, April.
- I. D. Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.
- 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. 2004. A smorgasbord of features for statistical machine translation. pages 161–168, Boston, MA, May.
- R. Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modeling. *Computer Speech and Language*, 10(3):187–228.
- C. Tillmann and H. Ney. 1997. Word triggers and the EM algorithm. In *Proc. Special Interest Group Workshop on Computational Natural Language Learning (ACL)*, pages 117–124, Madrid, Spain, July.
- I. García Varea, F. J. Och, H. Ney, and F. Casacuberta. 2001. Refined lexicon models for statistical machine translation using a maximum entropy approach. In *ACL ’01: 39th Annual Meeting on Association for Computational Linguistics*, pages 204–211, Morristown, NJ, USA.
- S. Venkatapathy and S. Bangalore. 2007. Three models for discriminative machine translation using global lexical selection and sentence reconstruction. In *SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 96–102, Rochester, New York, April.
- R. Zens and H. Ney. 2008. Improvements in dynamic programming beam search for phrase-based statistical machine translation. In *International Workshop on Spoken Language Translation*, Honolulu, Hawaii, October.