Translation Model Adaptation for an Arabic/French News Translation System by Lightly-Supervised Training

Holger Schwenk
LIUM, University of Le Mans
72085 Le Mans cedex 9, FRANCE
schwenk@lium.univ-lemans.fr

Jean Senellart
SYSTRAN SA
92044 Paris La Défense cedex, FRANCE
senellart@systran.fr

Abstract

Most of the existing, easily available parallel texts to train a statistical machine translation system are from international organizations that use a particular jargon. In this paper, we consider the automatic adaptation of such a translation model to the news domain. The initial system was trained on more than 200M words of UN bitexts. We then explore large amounts of in-domain *monolingual texts* to modify the probability distribution of the phrase-table and to learn new task-specific phrase-pairs. This procedure achieved an improvement of 3.5 points BLEU on the test set in an Arabic/French statistical machine translation system. This result compares favorably with other large state-of-the-art systems for this language pair.

1 Introduction

Adaptation of a statistical machine translation system (SMT) is a topic of increasing interest during the last years. Statistical (n-gram) language models are used in many domains and several approaches to adapt such models were proposed in the literature, for instance in the framework of automatic speech recognition. Many of these approaches were successfully used to adapt the language model of an SMT system. On the other hand, it seems more challenging to adapt the other components of an SMT system, namely the translation and reordering model. In this work we consider the adaptation of the translation model of a phrase-based SMT system.

While rule-based machine translation rely on rules and linguistic resources built for that purpose, SMT systems can be developed without the need of any language-specific expertise and are only based on bilingual sentence-aligned data ("*bitexts*") and large monolingual texts. However, while monolingual data is usually available in large amounts and for a variety of tasks, bilingual texts are a sparse resource for most language pairs.

Current parallel corpora mostly come from one domain (proceedings of the Canadian or European Parliament, or of the United Nations). This is problematic when SMT systems trained on such corpora are used for general translations, as the language jargon heavily used in these corpora is not appropriate for everyday life translations or translations in some other domain. This problem could be attacked by either searching for more in-domain training data, e.g. by exploring comparable corpora or the WEB, or by adapting the translation model to the task. In this work we consider translation model adaptation without using additional *bilingual data*. One can distinguish two types of translation model adaptation: first, adding new source words or/and new translations to the model; and second, modifying the probabilities of the existing model to better fit the topic of the task. These two directions are complementary and could be simultaneously applied. In this work we focus on the second type of adaptation.

A common way to modify a statistical model is to use a mixture model and to optimize the coefficients to the adaptation domain. This was investigated in the framework of SMT by several authors, for instance for word alignment (Civera and Juan, 2007), for language modeling (Zhao et al., 2004; Koehn and Schroeder, 2007) and to a lesser extent for the translation model (Foster and Kuhn, 2007; Chen et al., 2008). This mixture approach has the advan-

tage that only few parameters need to be modified, the mixture coefficients. On the other hand, many translation probabilities are modified at once and it is not possible to selectively modify the probabilities of particular phrases.

Comparable corpora are commonly used to find additional parallel texts, candidate sentences being often identified with help of information retrieval techniques, for instance (Hildebrand et al., 2005). Recently, a similar idea was applied to adapt the translation and language model using monolingual texts in the *target language* (Snover et al., 2008). Cross-lingual information retrieval was applied to find texts in the target language that are related to the domain of the source texts. However, it was difficult to get the alignments between the source and target phrases and an over-generalizing IBM1-style approach was used.

Another direction of research is self-enhancing of the translation model. This was first proposed by (Ueffing, 2006). The idea is to translate the test data, to filter the translations with help of a confidence score and to use the most reliable ones to train an additional small phrase table that is jointly used with the generic phrase table. This could be also seen as a mixture model with the in-domain component being build on-the-fly for each test set. In practice, such an approach is probably only feasible when large amounts of test data are collected and processed at once, e.g. a typical evaluation set up with a test set of about 50k words. This method of self-enhancing the translation model seems to be more difficult to apply for on-line SMT, e.g. a WEB service, since often the translation of some sentences only is requested. In follow up work, this approach was refined (Ueffing, 2007). Domain adaptation was also performed simultaneously for the translation, language and reordering model (Chen et al., 2008).

A somehow related approach was named lightly-supervised training (Schwenk, 2008). In that work an SMT system is used to translate large amounts of monolingual texts, to filter them and to add them to the translation model training data. We could obtain small improvements in the BLEU score in a French/English translation system. Although this technique seems to be close to *self enhancing* as proposed by (Ueffing, 2006), there is a conceptual difference. We do not use the test data to adapt the translation model, but large amounts of *monolingual training data* in the source language and we create a *complete new model* that can be applied to *any test data* without additional modification of the system. This kind of adapted system can be used in WEB service.

In this paper, we use the same type of approach to adapt an *generic* Arabic/French translation system to the news domain. This task is interesting for several reasons: there is only a limited amount of in-domain bitexts available (about 1.2M words), but large amounts of out-of-domain bitexts ($\approx$150M words of UN data) and both languages have a rich morphology. Usually, the Arabic source words are decomposed to detach pre- and suffixes. This helps to significantly reduce the size of the translation vocabulary and is reported to improve the translation quality. This morphological decomposition also results in many different and infrequent phrases which may lead to bad relative frequency estimates of the phrase translation probabilities. We are aiming in improving those estimates by using large amount of monolingual in-domain data. Finally, there seems to be a real need to translate between Arabic and French for the population in the Mediterranean area.

This paper is organized as follows. In the next section we first describe the considered task and the available bilingual and monolingual resources. Section 3 describes the baseline SMT systems. The following section describe our adaptation technique. Results are summarized in section 5 and the paper concludes with a discussion and perspectives of this work.

2 Task Description and resources

In this paper, we consider the translation of news texts from Arabic into French. We are not aware of easily available aligned parallel corpora for this language pair. Fortunately, Arabic and French are both official languages of the United Nations. We crawled data from various sources of the United Nations over the period 1988-2008. This totals in almost 150M Arabic words. The Arabic and French texts were automatically sentence aligned. This amount of parallel texts is usually considered as more than sufficient to train an SMT system. Note however, that a particular jargon is used in the UN

texts that is not appropriate for news-text translation.

The French TRAMES[1] project considered the translation of Arabic Speeches to French. In the framework of this project, about 90h of Arabic TV and radio broadcast news were recorded, transcribed and translated into French. The sources are Orient, Qatar, BBC, Alarabiya, Aljazeera and Alalam. These high-quality domain specific bitexts of about 262k Arabic words were made available to us by the DGA.[2]

Additional bilingual training data was obtained from the Project Syndicate WEB-site.[3] This data source is already used to build SMT systems to translate between European languages, in particular in the framework of the evaluations organized in junction with the workshops on statistical machine translation (Callison-Burch et al., 2007; Callison-Burch et al., 2008). Some of the texts are also translated into Arabic. The scripts to access this WEB-site were kindly made available by P. Koehn. We crawled and aligned a total of 1.6M words. Note that these texts are not exactly broadcast news texts.

The characteristics of the translation model training data is summarized in table 1. The number of words is given after tokenization.

	Arabic		French	
	Words	Vocab	Words	Vocab
DGA TRAMES	262k	30k	400k	18k
News commentary	1.1M	67k	1.3M	41k
UN	149M	712k	212M	420k

Table 1: Characteristics of the available bitexts

The DGA also provided a test set that was created in the same way than the in-domain bitexts. Four high-quality reference translations are available. We randomly split this data into a development set for system tuning and an internal test set. The details of the development and test set are given in Table 2.

We are only aware of one other large Arabic/French news translation system, the one that was developed during the TRAMES project (Hasan and Ney, 2008). In that work, results are reported on the same test set, but different bitexts for training were used, namely the TRAMES bitexts, UN data from the period 2001 to April 2007, archives of Amnesty International and articles from Le Monde Diplomatique. The authors report a BLEU score of 41.1 on the whole test data of 466 lines. Tuning was done on some held-out data which was not exactly specified. Due to the differences in the resources used by both systems, it is of course not possible to directly compare the results. Nevertheless, we can probably infer that MT systems achieving BLEU scores close to 42 on this test set are state-of-the-art. This is also the range of scores obtained by the best systems in the NIST 2008 evaluations when translating from Arabic to English. Translation from Arabic to French could be also approached by pivoting through English, but we don't have comparable BLEU scores for this kind of approach.

	Arabic	French
Dev data:		
Sentences	235	940 (4x)
Words	9931	45038
Test data:		
Sentences	231	924 (4x)
Words	10246	47066

Table 2: Characteristics of the development and test set.

3 Baseline system

The baseline system is a standard phrase-based SMT system based on the the Moses SMT toolkit (Koehn et al., 2007). It uses fourteen features functions, namely phrase and lexical translation probabilities in both directions, seven features for the lexicalized distortion model, a word and a phrase penalty, and a target language model. It constructed as follows. First, word alignments in both directions are calculated. We used a multi-threaded version of the GIZA++ tool (Gao and Vogel, 2008).[4] This speeds up the process and corrects an error of GIZA++ that can appear with rare words. Phrases and lexical reorderings are extracted using the default settings of the Moses toolkit. All the bitexts were concatenated. The parameters of Moses are tuned on the development data using the CMERT tool.

[1] Traduction Automatique par Méthodes Statistiques

[2] Direction générale de l'armement

[3] http://www.project-syndicate.org

[4] The source is available at http://www.cs.cmu.edu/~qing/

3.1 Tokenization

There is a large body of work in the literature showing that a morphological decomposition of the Arabic words can improve the word coverage and by these means the translation quality, see for instance (Habash and Sadat, 2006). It is clear that such a decomposition is most helpful when the translation model training data is limited, but this is less obvious for tasks where several hundreds of millions of words of bitexts are available. Most of the published work is based on the freely available tools, like the Buckwalter transliterator and the MADA and TOKAN tools for morphological analysis from Columbia University.

In this work, we compare two different tokenization of the Arabic source text: a full word mode and the morphological decomposition provided by the sentence analysis module of SYSTRAN's rule-based Arabic/English translation software. Sentence analysis represents a large share of the computations in a rule-based system. This process first applies decomposition rules coupled with a word dictionary. For words that are not known in the dictionary, the most likely decomposition is guessed. In general, all possible decompositions of each word are generated and then filtered in the context of the sentence. This step uses lexical knowledge and a global analysis of the sentences. This morphological decomposition drastically reduced the vocabulary of the Arabic bitexts: it was almost divided by two.

The French texts were tokenized using the tools of the Moses suite. Punctuation and case were preserved.

3.2 LM training

In contrast to the translation model, many resources are available to train a LM optimized on French broadcast news texts. We used the French side of the bitexts, data form the European and the Canadian parliament, news-data crawled in the framework for the 2009 WMT evaluation,[5] other WEB data collected by ourselves and finally LDC's Gigaword corpus. Separate 4-gram back-off LMs were build on each data source with the SRI LM toolkit (Stolcke, 2002) and then linearly interpolated, optimizing the coefficients with an EM procedure. The perplexities of these LMs are given in Table 3.

[5] newstrain08 as available on `http://www.statmt.org/wmt2009`

	#words	Perplex	Coeff
Trames	400k	223.4	0.174
NC	1.3M	267.6	0.004
UN	237M	234.7	0.035
Europarl	45M	187.9	0.048
Hansard	72M	320.6	0.031
WEB	33M	139.4	0.044
Newstrain	159M	139.3	0.037
AFP9x	236M	112.6	0.097
AFP2x	334M	87.4	0.407
APW	200M	104.0	0.125
interpolated	all	59.1	-

Table 3: LM training data, perplexities and interpolation coefficients (NC=news-commentary, AFP and APW are from LDC's Gigaword corpus)

4 Translation model adaptation by lightly-supervised training

The goal of this work is to adapt the translation model without using additional bilingual data. Instead we will use in-domain monolingual data in the source language. Usually it is relatively easy to obtain large collections of such data, in particular in the news domain as considered in this work. We use parts of the LDC Arabic Gigaword corpus, but more recent texts could be easily found on the Internet.

These texts are translated by an initial, unadapted SMT system. We then need to filter the automatic translations in order to keep only the "good ones" for addition to the translation model training data. This selection could take advantage of word-based confidence scores (Ueffing, 2007). We use the sentence-length normalized log-likelihoods of the decoder. These selected translations are used as additional in-domain bitexts and the standard procedure to build a new SMT system is performed, i.e. word alignment with GIZA++, phrase extraction and tuning of the system parameters.

Alternately, we could reuse the alignments established by the translation process since the Moses decoder is able to output the phrase and *word alignments*. This would speed up the process of creating the adapted SMT system since we skip the time-consuming word alignment performed by GIZA++.

Source	Arabic	French
AFP	**145M**	570M
APW	-	200M
ASB	7M	-
HYT	175M	-
NHR	188M	-
UMH	1M	-
XIN	58M	-

Table 4: Characteristics of the available monolingual Gigaword corpora (number of words).

It could also be that the decoder-induced word alignments are more appropriate than those performed by GIZA++. This was partially investigated in the framework of pivot translation to produce artificially bitexts in another language (Bertoldi et al., 2008). Finally, instead of only using the 1-best translation we could also use the n-best list.

LDC's Arabic and French Gigaword corpora are described in Table 4. There is only one source that does exist in both languages: the AFP collection. It is likely that the Arabic and French texts partially cover the same facts, but they are usually not direct translations.[6] In fact, we were informed that journalists at AFP have the possibility to freely change the sentences when they report on a fact based on text already available in another language. Nevertheless, it can be expected that using these texts in the *target language model* helps the SMT system to produce good translations. This language model training data can be considered as some form of light supervision and we will therefore use the term *lightly-supervised training* (Schwenk, 2008). This can be compared to the research in speech recognition where the same term was used for supervision that either comes from approximate transcriptions of the audio signal (closed captions) or related language model training data.

In this work, we have processed the AFP Arabic text only, but the other texts (ASB, HYT, ...) could be processed in the same manner. In fact it is an interesting question whether the availability of related or even "comparable" texts for the target language model is a necessary condition for our approach to work. All these issues will be explored in future research.

[6]Note that there are almost four times as much texts in French than in Arabic.

5 Experimental Evaluation

We first performed experiments using different amounts of bitexts to train the translation model and analyzed the benefits of the morphological decomposition of the Arabic words. The results are summarized in Table 5. The TRAMES training corpus contains several lines with more than 100 words. These can't be processed by the GIZA++ tool and the were discarded. This was the case for about 6% of the data. In future research, we will try to split those lines into shorter sentences.

As expected, the morphological decomposition of the Arabic words is very helpful when only a small amount of training data is available: using only the TRAMES and news-commentary in-domain data we observed an improvement of 4.6 BLEU points (first and second line in Table 5). Note that in this case we have actually less training data since the morphological decomposition leads to longer phrases out of which many are discarded by the 100 words limit of GIZA++. Somewhat surprisingly, the morphological decomposition still achieves a significant improvement of 1 BLEU point when more than 200M words of bitexts are available (last two lines in Table 5).

5.1 Translation model adaptation

The best system we were able to build with all human provided translations was used to translate all the AFP news texts from Arabic to French. The phrase table of this system has 329M entries and occupies 7.9GB on disk (gzipped). Translating the

Bitexts	#words Arabic	Morphol. decomp.	Dev	Test
TRAMES	1021k	-	32.10	30.85
+NC	858k	+	36.68	35.45
UN	142M	-	38.97	37.53
	203M	+	40.02	37.91
TRAMES	143M	-	39.64	38.99
+NC+UN	204M	+	41.88	40.04

Table 5: BLEU scores for different unadapted systems (NC=news-commentary).

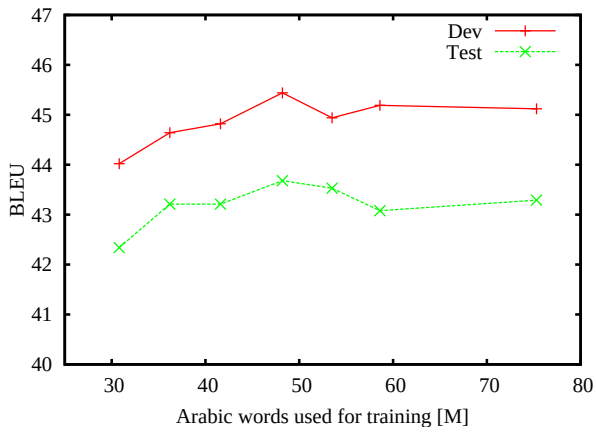

Figure 1: BLEU scores when adding automatic translations to the TRAMES and NC bitexts.

	#words Arabic	Dev	Test
Unadapted	217M	41.88	40.04
Adapted	48M	45.44	43.68

Table 6: Improvements obtained by the automatically adapted systems (BLEU scores).

whole AFP Gigaword corpus with such a large system is a computational challenge since it is impossible to charge the whole phrase table into memory. The Moses system supports two procedures to deal with this problem: filtering of the phrase table or binary on-disk phrase tables. Neither technique can be applied here. The phrase table is still too big in the first case and the binary representation of the whole phrase table occupies too much space on disk. This problem could be eventually approached with a distributed representation of the data structures. We finally adopted a combination of both techniques: the AFP corpus is split into parts with 50k lines (approximately 1.5M words), the phrase table is filtered for this data and then binarized. This made it possible to load the LM into memory and to have a process size of less then 20GB. The total translation time was more than 2700 hours.[7]

These automatic translations were filtered according to the sentence-length normalized log-likelihood and the most likely ones were used as bitexts. Different amounts of data can be obtained by varying the threshold on the likelihood. The BLEU scores on the development and test data in function of the total size of the bitexts are shown in figure 1. In these experiments we only use the in-domain human-provided bitexts (TRAMES and news-commentary) – the UN data being *replaced* by the automatic translations.

The best value on the development data was obtained for a total of 48M words of bitexts. The BLEU score on the development data is 45.44 and 43.68 on the test data respectively (see also table 6). This is an improvement of 3.5 BLEU points on both data sets.

We analyzed the phrase table of the original system trained on all human provided data, including UN, and the one of the automatically adapted system. This is summarized in table 7. The original phrase table had 329M entries out of which 22.9M could be potentially applied on the test data. The phrase table of the adapted system on the other hand used only 700k out of a total of 8.6M phrases. It seems clear that the phrase table obtained by training on the UN data contains many entries that are not useful, or eventually even correspond to wrong translations. Surprisingly, the phrase table of the adapted system is not only substantially smaller, but even contains about 11% more entries (18029 with respect to 16263). All these entries correspond to new sequences of known words since lightly-supervised training cannot extend the source or target side vocabulary. We conjecture that this is particularly important with the morphological decomposition of the Arabic words. This decomposition reduces the vocabulary size of the source language, but produces on the other hand many possible se-

	Unadapted	Adapted
Number of entries	22.9M	700k
Number of different source phrases	16263	18029
Average number of translations	1406.4	38.8
Average length of source phrases	2.65	2.81

Table 7: Characteristics of the phrase tables of the unadapted and adapted system. In both cases the table was filtered to include only entries that could be applied on the test data.

[7] translation was of course performed in parallel on a cluster.

Source:	اَلمَحكَمَةُ العِرَاقِيَّة بَدَأَت مُنذُ قَلِيل بِتَوجِيهِ لَائِحَةِ تُهَم ضِدَّ اَلرَئِيس اَلعِرَاقِيِّ السَّابِق.
Base:	le tribunal irakien a commencé depuis peu par la direction du règlement des accusations contre l'ancien président irakien.
Adapt:	le tribunal irakien a commencé depuis peu une liste d'accusations contre l'ancien président irakien.
Ref:	La Cour irakienne a commencé à dresser la liste des inculpations de l'ancien président irakien.
Source:	أَفَادَت مَصَادِرُ عَسكَرِيَّةٌ إِسرَائِيلِيَّة أَنَّ الجَيشَ الإِسرَائِيلِيّ اِعتَقَلَ لَيلاً نَاشِطاً فِي رَام اَللّه فِي الضِّفَّةِ الغَربِـيَّة كَمَا تَمَّ اعتِقَالُ نَاشِطَينِ آخَرَين كَانُوا يُحَصِّرُونَ
Base:	De source militaire israélienne a indiqué que l'armée israélienne a arrêté dans la nuit militants à Ramallah en Cisjordanie ont été arrêtés autres militants qui ...
Adapt:	Selon des sources militaires israéliennes, l'armée israélienne a arrêté dans la nuit de militants à Ramallah, en Cisjordanie, a également été arrêté deux autres activistes qui ...
Ref:	Des sources militaires israéliennes ont indiqué que l'armée israélienne a arrêté de nuit un activiste à Ramallah en Cisjordanie, ainsi que deux autres activistes qui ...
Source:	مُحَمَّدُ الغُبَارِي، جَولَةُ الصَّحَافَة، اَليَمَن.
Base:	Mohammed du brouillard, le cycle de la presse, au Yémen.
Adapt:	Mohammed, une tournée de la presse le Yémen.
Ref:	Mohamed Al-Ghobari, tour de la presse, Yémen.
Source:	مِن جِهَةٍ أُخرَى شَرَعَت تَايلَاند أَيضاً فِي سَحبِ جُنُودِهَا مِنَ العِرَاق.
Base:	d'autre part commencé aussi embarrassée à retirer ses troupes d'Irak.
Adapt:	D'autre part, la Thaïlande a commencé à retirer ses troupes d'Irak.
Ref:	D'autre part, la Thaïlande a également commencé à retirer ses troupes d'Irak.

Figure 2: Example translations from the test set of the baseline and the automatically adapted system. We also give the closest reference translation.

.

quences of tokens. It seems important to a include sequences of these tokens in the phrase table that appear in in-domain data. As a side effect, the smaller phrase-table of the adapted system also leads to a 40% faster translation.

We compared the translations of the unadapted and adapted systems: the TER is about 30 in both directions, meaning that the outputs differ substantially. Some example translations are shown in figure 2. The adapted system clearly produces better output in these examples. There remain of course some errors in these sentences, but we argue that the quality is high enough for an human being to capture most of the meaning of the sentences.

6 Conclusion

Statistical machine translation is today used to rapidly create automatic translations systems for a variety of tasks. In principle, we only need aligned example translations and monolingual data in the target language. However, for many application domains and language pairs there are no appropriate in-domain parallel texts to train the translation model. On the other hand, large generic bitexts may be available.

In this work we consider such a configuration: the translation of broadcast news texts from Arabic to French. We have a little more than 1M words of in-domain bitexts and about 150M words of generic bitexts from the United Nations. This system is automatically adapted to the news domain by using large amounts of *monolingual texts*, namely LDC's collection of Arabic AFP texts from 1994 to 2006. These texts were processed by the initial SMT system and the most reliable automatic translations were added to the bitexts and a new system was trained. By these means we achieved an improvement in the BLEU score of 3.5 points on the test set. This system actually uses less bitexts than the generic one since the generic UN bitexts are not

used any more.

An analysis of the created phrase table seems to indicate that the adaptation of the translation model leads to much smaller and more concise phrases. Our best system achieves a BLEU score of 43.68 on the test set which compares favorably with other large state-of-the-art systems for this language pair.

The proposed algorithm is generic and could be applied to other language pairs and applications domains. We only need a good initial generic SMT system and in-domain monolingual texts in the source language.r

It is interesting to compare our approach to self-learning as proposed in (Ueffing, 2006) Self-learning was applied to small amounts of test data only while we use several hundreds of million words of training data in the source language. We build a complete new phrase table instead of interpolating a small "adapted phrase table" with a generic one. Finally, self-learning was applied during the translation process and must be repeated for each new test data. This is computationally expensive and is difficult to use in on-line translation. The approach proposed in this paper applies translation model adaptation once and builds a new SMT system that can be then applied to any test data (ideally from the same domain).

Several extensions of the proposed approach can be envisioned, namely improved confidence scores for filtering the most reliable translations, processing of n-best lists instead of using the most likely translation only, and reuse of the decoder-induced word alignments instead of rerunning GIZA++. We are currently working on these issues.

Acknowledgments

The Arabic/French corpus of broadcasts news as well as the corresponding test set were made available to us by the DGA. This work has been partially funded by the French Government under the project INSTAR (ANR JCJC06_143038) and the European Commission under the project EuromatrixPlus.

References

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-based statistical machine translation with pivot languages. In *IWSLT*, pages 143–149.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. Meta-evaluation of machine translation. In *Second Workshop on SMT*, pages 136–158.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Third Workshop on SMT*, pages 70–106.

Boxing Chen, Min Zhang, Aiti Aw, and Haizhou Li. 2008. Exploiting n-best hypotheses for SMT self-enhancement. In *ACL*, pages 157–160.

Jorge Civera and Alfons Juan. 2007. Domain adaptation in statistical machine translation with mixture modelling. In *Second Workshop on SMT*, pages 177–180, June.

G. Foster and R. Kuhn. 2007. Mixture-model adaptation for smt. In *EMNLP*, pages 128–135.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June. Association for Computational Linguistics.

N. Habash and F. Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *NAACL*, pages 49–52.

Sasa Hasan and Hermann Ney. 2008. A multi-genre SMT system for Arabic to French. In *LREC*, pages 2167–2170.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. In *EAMT*, pages 133–142.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Second Workshop on SMT*, pages 224–227, June.

Philipp Koehn et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL, demonstration session*.

Holger Schwenk. 2008. Investigations on large-scale lightly-supervised training for statistical machine translation. In *IWSLT*, pages 182–189.

Matthew Snover, Bonnie Dorr, and Richard Schwartz. 2008. Language and translation model adaptation using comparable corpora. In *EMNLP*.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *ICSLP*, pages II: 901–904.

Nicola Ueffing. 2006. Using monolingual source-language data to improve MT performance. In *IWSLT*, pages 174–181.

Nicola Ueffing. 2007. Transductive learning for statistical machine translation. In *ACL*, pages 25–32.

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language model adaptation for statistical machine translation with structured query models. In *Cooling*.