Local lexical adaptation in Machine Translation through triangulation: SMT helping SMT

Josep Maria Crego
LIMSI-CNRS
jmcrego@limsi.fr

Aurélien Max
LIMSI-CNRS
Univ. Paris Sud
amax@limsi.fr

François Yvon
LIMSI-CNRS
Univ. Paris Sud
yvon@limsi.fr

Abstract

We present a framework where auxiliary MT systems are used to provide lexical predictions to a main SMT system. In this work, predictions are obtained by means of pivoting via auxiliary languages, and introduced into the main SMT system in the form of a low order language model, which is estimated on a sentence-by-sentence basis. The linear combination of models implemented by the decoder is thus extended with this additional language model. Experiments are carried out over three different translation tasks using the European Parliament corpus. For each task, nine additional languages are used as auxiliary languages to obtain the triangulated predictions. Translation accuracy results show that improvements in translation quality are obtained, even for large data conditions.

1 Introduction

Important improvements are yet to come regarding the performance of Statistical Machine Translation systems. Dependence on training data and limited modelling expressiveness are the focus of many research efforts, such as using monolingual corpora for the former and syntactic models for the latter.

Another promising approach consists in exploiting complementary sources of information in order to build better translations, as done by consensus-based system combination (e.g. (Matusov et al., 2008)). This, however, requires to have several systems available for the same language pair. Considering that the same training data would be available to all systems, differences in translation modelling are expected to produce redundant and complementary hypotheses. Multisource translation (e.g. (Och and Ney, 2001; Schwartz, 2008)) is a variant, involving source texts available in several languages which can be translated by systems for different language pairs and whose outputs can be successfully combined into better translations (Schroeder et al., 2009). One theoretical expectation of multisource translation is that it can successfully reduce ambiguity of the original source text, but does so under the rare conditions of availability of existing (accurate) translations. In contrast, pivot-based system combination (e.g. (Utiyama and Isahara, 2007; Wu and Wang, 2007)) aims at compensating the lack of training data for a given language pair by producing translation hypotheses obtained by pivoting via an intermediary language for which better systems are available.

These techniques generally produce a search space that differs from that of the direct translation systems. As such, they create a new translation system out of various systems for which diagnosis becomes more difficult.

This paper instead focusses on improving a single system, which should be state-of-the-art as regards data and models. We propose a framework in which information coming from external sources is used to boost lexical choices and guide the decoder into making more informed choices.[1]

[1] We performed initial experiments where the complementary information was exploited during n-best list reranking (Max et al., 2010), but except for the multisource condition the list of hypotheses contained too little useful variation

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 232–240,
Beijing, August 2010

Complementary sources can be of different nature: they can involve other automatic systems (for the same or different language pairs) and/or human knowledge. Furthermore, complementary information is injected at the lexical level, thus making targeted fine-grained lexical predictions useful. Importantly, those predictions are exploited at the sentence level[2], so as to allow for efficient use of source contextual information.

The second contribution of this paper is an instantiation of the proposed framework. Automatically pivoting via auxiliary languages is used to make complementary predictions that are exploited through language model adaptation by the decoder for a given language pair. For this apparently difficult condition, where predictions result from automatic translations involving two systems, we manage to report significant improvements, measured with respect to the target and the source text, under various configurations.

This paper is organized as follows. We first review related work in section 2.1, and describe the distinctive characteristics of our approach in Section 2.2. Section 2.3 presents our instantiation of the framework based on lexical boosting via auxiliary language triangulation. Experiments involving three language pairs of various complexity and different amounts of training data are described in Section 3. We finally conclude by discussing the prospects offered by our proposed framework in Section 4.

2 A framework for sentence-level lexical boosting

2.1 Related work

The idea of using more than one translation system to improve translation performance is not new and has been implemented in many different ways which we briefly review here.

System combination An often used strategy consists in *combining the output of several systems* for a fixed language pair, and to rescore the resulting set of hypotheses taking into account all the available translations and scores. Various proposals have been made to efficiently perform such a combination, using auxiliary data structures such as n-best lists, word lattices or consensus networks (see for instance (Kumar and Byrne, 2004; Rosti et al., 2007; Matusov et al., 2008; Hildebrand and Vogel, 2008; Tromble et al., 2008)). Theses techniques have proven extremely effective and have allowed to deliver very significant gains in several recent evaluation campaigns (Callison-Burch et al., 2008).

Multisource translation A related, yet more resourceful approach, consists in trying to combine several systems providing translations *from different sources into the same target*, provided such *multilingual sources* are available. (Och and Ney, 2001) propose to select the most promising translation amongst the hypotheses produced by several Foreign→English systems, where output selection is based on the translation scores. The intuition that if a system assigns a high figure of merits to the translation of a particular sentence, then this translation should be preferred, is implemented in the MAX combination heuristics, whose relative (lack of) success is discussed in (Schwartz, 2008). A similar idea is explored in (Nomoto, 2004), where the sole target language model score is used to rank competing outputs. (Schroeder et al., 2009) propose to combine the available sources prior to translation, under the form of a multilingual lattice, which is decoded with a multisource phrase table. (Chen et al., 2008) integrate the available auxiliary information in a different manner, and discuss how to improve the translation model of the primary system: the idea is to use the entries in the phrase table of the auxiliary system to filter out those accidental correspondences that pollute the main translation model. The most effective implementation of multisource translation to date however consists in using mono-source system combination techniques (Schroeder et al., 2009).

Translation through pivoting The use of auxiliary systems has also been proposed in another common situation, as a possible remedy to the lack of parallel data for a particular language pair, or for a particular domain. Assume, for instance, that one wishes to build a translation system for

to lead to measurable improvements.

[2]We plan to experiment next on using predictions at the document level.

the pair $A \rightarrow B$, for which the parallel data is sparse; assuming further that such parallel resources exist for pairs $A \rightarrow C$ and for $C \rightarrow B$, it is then tempting to perform the translation indirectly through *pivoting*, by first translating from A to C, then from C to B. Direct implementations of this idea are discussed e.g. in (Utiyama and Isahara, 2007). Pivoting can also intervene earlier in the process, for instance as a means to *automatically generate* the missing parallel resource, an idea that has also been considered to adapt an existing translation systems to new domains (Bertoldi and Federico, 2009). Pivoting can finally be used to fix or improve the translation model: (Cohn and Lapata, 2007) augments the phrase table for a baseline bilingual system with supplementary phrases obtained by pivoting into a third language.

Triangulation in translation Triangulation techniques are somewhat more general and only require the availabily of *one* auxiliary system (or one auxiliary parallel corpus). For instance, the authors of (Chen et al., 2008) propose to use the translation model of an auxiliary $C \rightarrow B$ system to filter-out the phrase-table of a primary $A \rightarrow B$ system.

2.2 Our framework

As in other works, we propose to make use of several MT systems (of any type) to improve translation performance, but contrarily to these works we concentrate on *improving one particular system*. Our framework is illustrated on Figure 1. The main system (henceforth, *direct* system), corresponding to configuration **1**, is a SMT system, translating from German to English in the example. Auxiliary information may originate from various sources (**2**-**6**) and enter into the decoder. A new model is dynamically built and is used to guide the exploration of the search space to the best hypothesis. Several auxiliary models can be used at once and can be weighted by standard optimization techniques using development data, so that bad sources are not used in practice, or by exploiting *a priori* information. In the implementation described in section 2.3, this information is updated by the auxiliary source at each sentence.

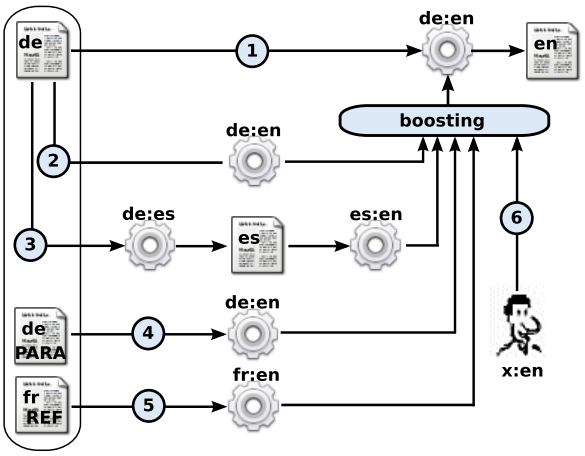

Figure 1: Lexical boosting framework with various configurations for auxiliary predictions

We now briefly describe various possible configurations to make some links to previous works explicit. Configuration **2** translates the same source text by means of another system for the same language pair, as would be done in system combination, except that here a new complete decoding is performed by the direct system. Configuration **3**, which will be detailed in section 2.3, uses translations obtained by triangulating via an auxiliary language (Spanish in the example). Using this two-step translation is common to pivot approaches, but our approach is different in that the result of the triangulation is only used as auxiliary information for the decoding of the direct system. Configurations **4** and **5** are instances of multisource translation, where a paraphrase or a translation of the source text is available. Lastly, configuration **6** illustrates the case where a human translator, with knowledge of the target language and at least of one of the available source languages, could influence the decoding by providing *desired*[3] words (e.g. only for source words or phrases that would be judged difficult to translate). This human supervision through a feedback text in real time is similar to the proposal of (Dymetman et al., 2003).

Given this framework, several questions arise,

[3]The proposal as it is limits the hypotheses produced by the system to those that are attainable given its training data. It is conceivable, however, to find ways of introducing new knowledge in this framework.

the most important underlying this work being whether the performance of SMT systems can be improved by using other SMT systems. Another point of interest is whether improvements made to *auxiliary* systems can yield improvement to the *direct* system, without the latter undergoing any modification.

2.3 Lexical boosting via triangulation

Auxiliary translations obtained by pivoting can be viewed as a source of adaptation data for the target language model of the direct system. Assuming we have computed n-best translation hypotheses of a sentence in the target language, we can then boost the likeliness of the words and phrases occurring in these hypotheses by deriving an auxiliary language model for each test sentence. This allows us to integrate this auxiliary information during the search and thus provides a tighter integration with the direct system. This idea has successfully been used in speech recognition, using for instance close captions (Placeway and Lafferty, 1996) or an imperfect translation (Paulik et al., 2005) to provide auxiliary in-domain adaptation data for the recognizer's language model. (Simard and Isabelle, 2009) proposed a similar approach in Machine Translation in which they use the target-side of an exact match in a translation memory to build language models on a per sentence basis used in their decoder.

This strategy can be implemented in a straightforward manner, by simply training a language model using the n-best list as an adaptation corpus. Being automatically generated, hypotheses in the n-best list are not entirely reliable: in particular, they may contain very unlikely target sequences at the junction of two segments. It is however straightforward to filter these out using the available phrase alignment information.

This configuration is illustrated on Figure 2: the direct system (configuration **1**) makes use of predictions from pivoting through an auxiliary language (configuration **2**), where n-best lists can be used to produce several hypotheses. In order to get a upper bound on the potential gains of this approach, we can run the artificial experiment (configuration **3**) where a reference in the target language is used as a "perfect" source of information. Furthermore, we are interested in the performance of the simple pivot system alone (configuration **4**), as it gives an indication of the quality of the data used for LM adaptation.

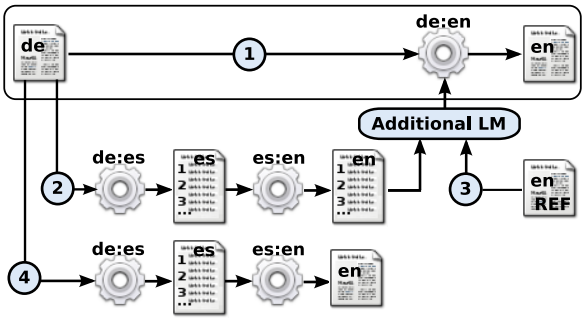

Figure 2: Architecture of a German→English system for lexical boosting via triangulation through Spanish

3 Experiments and results

3.1 Translation engine

In this study, we used our own machine translation engine, which implements the n-gram-based approach to statistical machine translation (Mariño et al., 2006). The translation model is implemented as a stochastic finite-state transducer trained using a n-gram language model of (source,target) pairs.

In addition to a bilingual n-gram model, our SMT system uses six additional models which are linearly combined following a discriminative modeling framework: two *lexicalized reordering* (Tillmann, 2004) models,a *target-language model*, two *lexicon models*, a 'weak' distance-based *distortion model*, a *word bonus model* and a *translation unit bonus model*. Coefficients in this linear combination are tuned over development data with the MERT optimization toolkit[4], slightly modified to use our decoder's n-best lists.

For this study, we used 3-gram bilingual and 3-gram target language models built using modified Kneser-Ney smoothing (Chen and Goodman, 1996); model estimation was performed with the SRI language modeling toolkit.[5] Target language

[4] http://www.statmt.org/moses

[5] http://wwww.speech.sri.com/projects/srilm

models were trained on the target side of the bitext corpora.

After preprocessing the corpora with standard tokenization tools, word-to-word alignments are performed in both directions, source-to-target and target-to-source. In our system implementation, the *GIZA++* toolkit[6] is used to compute the word alignments. Then, the *grow-diag-final-and* heuristic is used to obtain the final alignments from which translation units are extracted. Convergent studies have showed that systems built according to these principles typically achieve a performance comparable to that of the widely used MOSES phrase-based system for the language pairs under study.

3.2 Corpora

We have used the Europarl corpus[7] for our main and auxiliary languages. The eleven languages are: Danish (da), German (de), English (en), Spanish (es), Finnish (fi), French (fr), Greek (el), Italian (it), Dutch (nl), Portuguese (pt) and Swedish (sv).

We focussed on three translation tasks: one for which translation accuracy, as measured by automatic metrics, is rather high ($fr \rightarrow en$), and two for which translation accuracy is lower ($de \rightarrow en$) and ($fr \rightarrow de$). This will allow us to check whether the improvements provided by our method carry over even in situations where the baseline is strong; conversely, it will allow us to assess whether the proposed techniques are applicable when the baseline is average or poor.

In order to measure the contribution of each of the auxiliary languages we used a subset of the training corpus that is common to all language pairs, hereinafter referred to as the *intersection* data condition. We used the English side of all training language pairs to collect the same sentences in all languages, summing up to $320,304$ sentence pairs. Some statistics on the data used in this study are reported in Table 1. Finally, in order to assess the impact of the training data size over the results obtained, we also considered a much more challenging condition for the $fr \rightarrow de$ pair, where we used the entire Europarl data (V5) made available for the fifth Workshop on Statistical Machine Translation[8] for training, and test our system on out-of-domain news data. The training corpus in this condition contains 43.6M French words and 37.2M German words.

Development and test data for the first condition (*intersection*) were obtained by leaving out respectively 500 and 1000 sentences from the common subset (same sentences for all languages), while the first 500 sentences of *newstest2008* and the entire *newstest2009* official test sets were used for the *full* data condition.

	Train		*Dev*			*Test*		
	Words	*Voc.*	*Words*	*Voc.*	*OOV*	*Words*	*Voc.*	*OOV*
da	$8.5M$	$133.5k$	$13.4k$	$3.2k$	104	$25.9k$	$5.1k$	226
de	$8.5M$	$145.3k$	$13.5k$	$3.5k$	120	$26.0k$	$5.5k$	245
en	$8.9M$	$53.7k$	$14.0k$	$2.8k$	39	$27.2k$	$4.0k$	63
es	$9.3M$	$85.3k$	$14.6k$	$3.3k$	56	$28.6k$	$5.0k$	88
fi	$6.4M$	$274.9k$	$10.1k$	$4.3k$	244	$19.6k$	$7.1k$	407
fr	$10.3M$	$67.8k$	$16.1k$	$3.2k$	47	$31.5k$	$4.8k$	87
el	$8.9M$	$128.3k$	$14.1k$	$3.9k$	72	$27.2k$	$6.2k$	159
it	$9.0M$	$78.9k$	$14.3k$	$3.4k$	61	$28.1k$	$5.1k$	99
nl	$8.9M$	$105.0k$	$14.2k$	$3.1k$	76	$27.5k$	$4.8k$	162
pt	$9.2M$	$87.3k$	$14.5k$	$3.4k$	49	$28.3k$	$5.2k$	118
sv	$8.0M$	$140.8k$	$12.7k$	$3.3k$	116	$24.5k$	$5.2k$	226

Table 1: *Statistics for the training, development and test sets of the intersection data condition*

3.3 Results

In this section, we report on the experiments carried out to assess the benefits of introducing an auxiliary language model to the linear combination of models implemented in our SMT system.

Table 2 reports translation accuracy (BLEU) results for the main translation tasks considered in this work ($fr \rightarrow de$), ($fr \rightarrow en$) and ($de \rightarrow en$), as well as for multiple intermediate tasks needed for pivoting via auxiliary systems.

For each triplet of languages (*src*, *aux*, *trg*), columns 4^{th} to 6^{th} show BLEU scores for systems performing ($src \rightarrow aux$), ($aux \rightarrow trg$) and *pivot* translations using *aux* as the bridge language.

The last two columns display BLEU scores for the main translation tasks ($fr \rightarrow de$), ($fr \rightarrow en$) and ($de \rightarrow en$). Column *src-trg* refers to the baseline (direct) systems, for which no additional lan-

[6] http://www.fjoch.com/GIZA++.html

[7] http://www.statmt.org/europarl

[8] http://www.statmt.org/wmt10

src aux trg	*src-aux*	*aux-trg*	*pivot*	*src-trg*	*+auxLM*
Intersection data condition					
fr - de	-	-	-	18.02	
da	22.78	20.02	16.27		+0.44
el	24.54	18.51	15.86		+0.76
en	29.53	17.31	15.69		+0.50
es	**34.94**	18.31	**16.76**		**+0.96**
fi	10.71	14.15	11.39		+0.65
it	31.60	16.86	16.54		-0.05
nl	22.71	**21.44**	**16.76**		+0.55
pt	33.61	17.47	16.34		-0.12
sv	20.73	19.59	13.73		-0.14
average					*+0.39*
- - ref	-	-	-	-	+6.46
fr - en	-	-	-	29.53	
da	22.78	29.54	25.48		+0.02
de	18.02	24.66	23.50		+0.05
el	24.54	29.37	25.31		+0.07
es	**34.94**	**31.05**	**27.76**		**+0.61**
fi	10.71	20.56	19.15		+0.44
it	31.60	25.75	25.79		+0.32
nl	22.71	24.49	25.15		+0.01
pt	33.61	29.44	27.27		+0.01
sv	20.73	30.98	23.74		+0.50
average					*+0.22*
- - ref	-	-	-	-	+11.30
de - en	-	-	-	24.66	
da	24.59	29.54	**22.73**		**+0.96**
el	19.72	29.37	20.88		**+1.02**
es	**25.48**	**31.05**	21.23		+0.77
fi	12.42	20.56	18.02		**+0.94**
fr	25.93	29.53	21.55		+0.19
it	18.82	25.75	18.05		+0.19
nl	24.97	24.49	22.62		+0.64
pt	23.15	29.44	21.93		+0.87
sv	19.80	**30.98**	21.35		+0.69
average					*+0.69*
- - ref	-	-	-	-	+9.53
Full data condition					
fr - de	-	-	-	19.94	
es	38.76	20.18	19.36		+0.61

Table 2: *Translation accuracy (BLEU) results.*

guage model is used; column *+auxLM* refers to the same system augmented with the additional language model. Additional language models are built from hypotheses obtained by means of *pivot* translations, using *aux* as auxiliary language. The last score is shown in the form of the difference (improvement) with respect to the score of the baseline system.

This table additionally displays the BLEU results obtained when building the additional language models directly from the English reference translations (see last row of each translation task). These numbers provide an upper-bound of the expected improvements. Note finally that numbers in boldface correspond to the best numbers in their column for a given language pair.

As detailed above, the additional language models are built using *trg* hypotheses obtained by pivoting via an auxiliary language: $(src \rightarrow aux)$ $+$ $(aux \rightarrow trg)$. Hence, column *pivot* shows the quality (measured in terms of BLEU) of the hypotheses used to estimate the additional model. Note that we did not limit the language model to be estimated from the 1-best *pivot* hypotheses. Instead, we uses n-best translation hypotheses of the $(src \rightarrow aux)$ system and m-best hypotheses of the $(aux \rightarrow trg)$ system. Hence, $n \times m$ target hypotheses were used as training data to estimate the additional models. Column *+auxLM* shows BLEU scores over the test set after performing four system optimizations on the development set to select the best combination of values used for n and m among: $(1, 1)$, $(10, 1)$, $(10, 1)$ and $(10, 10)$. All hypotheses used to estimate a language model are considered equally likely. Language models are learnt using Witten-Bell discounting. Approximately ± 1.0 point must be added to BLEU scores shown in the last 2 columns for 95% confidence levels.

As expected, *pivot* translations yield lower quality scores than the corresponding direct translations hypotheses. However, *pivot* hypotheses may contain better lexical predictions, that the additional model helps transfer into the baseline system, yielding translations with a higher quality, as shown in many cases the *+auxLM* systems results. The case of using Finnish as an auxiliary language is particularly remarkable. Even though *pivot* hypotheses obtained through Finnish have the lowest scores[9], they help improve the baseline performance as additional language models.

As expected, the translation results of the pair

[9]Given the agglutinative nature of morphological processes in Finnish, reflected in a much lower number of words per sentence, and a higher number of types (see Table 1), BLEU scores for this language do not compare directly with the ones obtained for other languages.

with a highest baseline ($fr \rightarrow en$) were on average less improved than those of the pairs with lower baselines.

As can also be seen, the contribution of each auxiliary language varies for each of the three translation tasks. For instance, Danish (da) provides a clear improvement to ($de \rightarrow en$) translations, while no gain is observed for ($fr \rightarrow en$). No clear patterns seems to emerge, though, and the correlation between the quality of the pivot translation and the boost provided by using these pivot hypotheses remains to be better analyzed.

In order to assess whether the improvements obtained carry over larger data conditions, we trained our ($fr \rightarrow de$), ($fr \rightarrow es$) and ($es \rightarrow de$) systems over the entire EPPS data. Results are reported in the bottom part of Table 2. As can be seen, the ($fr \rightarrow de$) system is still improved by using the additional language model. However, the absolute value of the gain under the *full* condition ($+0.61$) is lower than that of the *intersection* data condition ($+0.96$).

3.4 Contrastive evaluation of lexical translation

In some cases, automatic metrics such as BLEU cannot show significant differences that can be revealed by fine-grained focussed human evaluation (e.g. (Vilar et al., 2006)). Furthermore, computing some similarity between a system's hypotheses and *gold standard* references puts a strong focus on the target side of translation, and does not allow evaluating translation performance from the source words that were actually translated. We therefore use the evaluation methodology described in (Max et al., 2010) for a complementary measure of translation performance that focuses on the contrastive ability of two systems to adequately translate source words.

Source words from the test corpus were first aligned with target words in the reference, by automatically aligning the union of the training and test corpus using GIZA++.[10] The test corpus was analyzed by the TREETAGGER[11] so as to identify content words, which have a more direct impact on translation adequacy. When source words are aligned to several target words, each target word should be individually searched for in the candidate translation, and words from the reference can only be matched once.

		Source words' part-of-speech						
aux		**ADJ**	**ADV**	**NOM**	**PRO**	**VER**	all	**+Bleu**
el	-	27	21	114	25	99	286	+0.07
	+	62	29	136	27	114	368	
es	-	33	25	106	26	110	300	+0.61
	+	64	38	136	22	117	377	
fi	-	44	40	106	20	92	302	+0.44
	+	49	31	120	23	106	329	
it	-	55	39	128	35	119	376	+0.32
	+	55	39	145	36	121	396	
sv	-	40	30	138	29	109	346	+0.50
	+	69	46	144	23	134	416	

Table 3: Contrastive lexical evaluation results per part-of-speech between the baseline French→English system and our systems using various auxiliary languages. '-' (resp. '+') values indicate numbers of words that only the baseline system (resp. our system) correctly translated with respect to the reference translation.

Table 3 shows contrastive results per part-of-speech between the baseline fr→en system and systems using various auxiliary languages. Values in the '-' row indicate the number of words that only the baseline system translated as in the reference translation, and values in the '+' row the number of words that only our corresponding system translated as in the reference. The most striking result is the contribution of Greek, which, while giving no gain in terms of BLEU, improved the translation of 82 content words. This could be explained, in addition to the lower Bleu3 and Bleu4 precision, by the fact that the quality of the translation of grammatical words may have decreased. On the contrary, Italian brings little improvement for content words save for nouns. The mostly negative results on the translation of pronouns were expected, because this depends on their antecedent in English and is not the object of specific modelling from the systems. The translation of nouns and adjectives benefits the most from auxiliary translations.

[10] The obtained alignments are thus strongly influenced by alignments from the training corpus. It could be noted that alignments could be manually corrected.

[11] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

Figure 3 illustrates this evaluation by means of two examples. It should be noted that a recurrent type of improvement was that of avoiding missing words, which is here a direct result of their being boosted in the auxiliary hypotheses.

4 Conclusions and future work

We have presented a framework where auxiliary MT systems are used to provide useful information to a main SMT system. Our experiments on auxiliary language triangulation have demonstrated its validity on a difficult configuration and have shown that improvements in translation quality could be obtained even under large training data conditions.

The fact that low quality sources such as pivot translation can provide useful complementary information calls for a better understanding of the phenomena at play. It is very likely that, looking at our results on the contribution of auxiliary languages, improving the quality of an auxiliary source can also be achieved by identifying what a source is good for. For example, in the studied language configurations predictions of translations for pronouns in the source text by auxiliary triangulation does not give access to useful information. On the contrary, triangulation with Greek when translating from French to English seems to give useful information regarding the translation of adjectives, a result which was quite unexpected.

Also, it would be interesting to use richer predictions than short n-grams, such as syntactic dependencies, but this would require significant changes on the decoders used. Using dynamic models at the discourse level rather than only at the sentence level would also be a useful improvement. Besides the improvements just mentioned, our future work includes working on several configurations of the framework described in section 2.2, in particular investigating the new type of system combination.

Acknowledgements

This work has been partially funded by OSEO under the Quaero program.

References

Bertoldi, Nicola and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of WMT*, Athens, Greece.

Callison-Burch, Chris, Cameron Shaw Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of WMT*, Columbus, USA.

Chen, Stanley F. and Joshua T. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of ACL*, Santa Cruz, USA.

Chen, Yu, Andreas Eisele, and Martin Kay. 2008. Improving statistical machine translation efficiency by triangulation. In *Proceedings of LREC*, Marrakech, Morocco.

Cohn, Trevor and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of ACL*, Prague, Czech Republic.

Dymetman, Marc, Aurélien Max, and Kenji Yamada. 2003. Towards interactive text understanding. In *Proceedings of ACL, short paper session*, Sapporo, Japan.

Hildebrand, Almut Silja and Stephan Vogel. 2008. Combination of machine translation systems via hypothesis selection from combined n-best lists. In *Proceedings of AMTA*, Honolulu, USA.

Kumar, Shankar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of NAACL-HLT*, Boston, USA.

Mariño, José, Rafael E. Banchs, Josep Maria Crego, Adria de Gispert, Patrick Lambert, J.A.R. Fonollosa, and Martha Costa-jussà. 2006. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549.

Matusov, Evgeny, Gregor Leusch, Rafael E. Banchs, Nicola Bertoldi, Daniel Dechelotte, Marcello Federico, Muntsin Kolss, Young-Suk Lee, Jose Mariño, Matthias Paulik, Salim Roukos, Holger Schwenk, and Hermann Ney. 2008. System combination for machine translation of spoken and written language. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1222–1237, September.

Max, Aurélien, Josep M. Crego, and François Yvon. 2010. Contrastive Lexical Evaluation of Machine Translation. In *Proceedings of LREC*, Valletta, Malta.

ref #357	this concession to the unions ignores the reality that all airlines have different safety procedures which even differ between aircrafts within each airline .
bas	this **concession unions ignores** the *fact* that all **airlines have different safety procedures** which are **even within each** of the *companies* in accordance with the types of equipment .
w.r.t. src	cette **concession** aux **syndicats ignore** la *réalité* selon laquelle toutes les **compagnies aériennes ont** des **procédures de sécurité différentes** qui diffèrent **même** au **sein** de chacune des *compagnies* en fonction des types d ' *appareils* .
+aux	this **concession** to the trade **unions ignores** the **reality** according to which all the **airlines have different safety procedures** which **differ even within each** of the *companies* in accordance with the types of equipment .
w.r.t. src	cette **concession** aux **syndicats ignore** la **réalité** selon laquelle toutes les **compagnies aériennes ont** des **procédures de sécurité différentes** qui **diffèrent même** au **sein** de chacune des *compagnies* en fonction des types d ' *appareils* .

Figure 3: Example of automatic translations from French to English for the baseline system and when using Spanish as the auxiliary language. Bold marking indicates source/target words which were correctly translated according to the reference translation.

Nomoto, Tadashi. 2004. Multi-engine machine translation with voted language model. In *Proceedings of ACL*, Barcelona, Catalunya, Spain.

Och, Franz Josef and Hermann Ney. 2001. Statistical multi-source translation. In *Proceedings of MT Summit*, Santiago de Compostela, Spain.

Paulik, Matthias, Christian Fügen, Thomas Schaaf, Tanja Schultz, Sebastian Stüker, and Alex Waibel. 2005. Document driven machine translation enhanced automatic speech recognition. In *Proceedings of InterSpeech*, Lisbon, Portugal.

Placeway, Paul and John Lafferty. 1996. Cheating with imperfect transcripts. In *Proceedings of ICSLP*, Philadelphia, USA.

Rosti, Antti-Veikko, Necip Fazil Ayan, Bin Xiang, Spyros Matsoukas, Richard Schwatz, and Bonnie J. Dorr. 2007. Combining outputs from multiple machine translation systems. In *Proceedings of NAACL-HTL*, Rochester, USA.

Schroeder, Josh, Trevor Cohn, and Philipp Koehn. 2009. Word lattices for multi-source translation. In *Proceedings of EACL*, Athens, Greece.

Schwartz, Lane. 2008. Multi-source translation methods. In *Proceedings of AMTA*, Honolulu, USA.

Simard, Michel and Pierre Isabelle. 2009. Phrase-based machine translation in a computer-assisted translation environment. In *Proceedings of Machine Translation Summit XII*, Ottawa, Canada.

Tillmann, Christoph. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of NAACL-HLT*, Boston, USA.

Tromble, Roy, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice Minimum Bayes-Risk decoding for statistical machine translation. In *Proceedings of EMNLP*, Honolulu, USA.

Utiyama, Masao and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of NAACL-HLT*, Rochester, USA.

Vilar, David, Jia Xu, Luis Fernando d'Haro, and Hermann Ney. 2006. Error Analysis of Statistical Machine Translation Output. In *Proceedings of LREC*, Genoa, Italy.

Wu, Hua and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of ACL*, Prague, Czech Republic.