Pivot-based Machine Translation between Statistical and Black Box systems

Antonio Toral
School of Computing
Dublin City University
Dublin, Ireland
atoral@computing.dcu.ie

Abstract

This paper presents a novel approach to pivot-based machine translation (MT): while the state-of-the-art uses two statistical systems, this proposal treats the second system as a black box. Our approach effecively provides pivot-based MT to target languages for which no suitable bilingual corpora are available to build statistical systems, as long as any other kind of MT system is available. We experiment with an algorithm that uses two features to find the best translation: the translation score provided by the first system and fluency of the final translation. Despite its simplicity, this approach yields significant improvements over the baseline, which translates the source sentences using the two MT systems sequentially. We have experimented with two scenarios, technical documentation in Romance languages and newswire in Slavic languages, obtaining 11.88% and 13.32% relative improvements in terms of BLEU, respectively.

1 Introduction

Pivot-based machine translation (MT) refers to the use of an intermediate language, called pivot language (PL), to translate from the source (SL) to the target language (TL). Unlike typical MT systems, which translate directly from SL to TL, pivot-based systems translate sequentially from SL to PL and then from PL to TL. The main motivation for building pivot-based MT systems is the lack of language resources for a language pair SL–TL, in contrast with the availability of such resources for both language pairs SL–PL and PL–TL.

Much of the research carried out in pivot-based MT concentrates on a scenario where the translation both from SL to PL and from PL to TL is carried out by statistical machine translation systems (SMT). It is also often assumed that the developer has access not only to the output of the systems but also to their internal data structures. Hence, for these methods to work, bilingual corpora for both SL–PL and PL–TL are required in order to train the corresponding SMT systems.

Our research concentrates on pivot-based MT for cases where there is no access to the internals of the second system (PL to TL), i.e. we treat it as a black box: only the output translations produced by this system are available. Because of this our approach is applicable to a much broader set of scenarios than the current state-of-the-art; i.e. it can be applied when there is no access to the internals of the second system, which is the case for many online MT systems, or when the second system does not provide the required data (such as n-best lists), which is the case for many rule-based machine translation systems (RBMT).

The remainder of this paper is organised as follows. Section 2 presents an overview of the state-of-the-art for pivot-based MT. This is followed by the description of our methodology. Subsequently, we carry out the evaluation and present the results of the proposal. Finally, we conclude and outline lines of future work.

2 Related Work

Pivot-based strategies that use SMT systems can be classified into three categories (Wu and Wang, 2009): phrase table multiplication (also known as triangulation), transfer (also referred to as cascade)

© 2012 European Association for Machine Translation.

and synthetic corpus.

Phrase table multiplication methods (Wu and Wang, 2007; Cohn and Lapata, 2007) induce a new SL–TL translation model by combining the corresponding translation probabilities of the translations models for SL–PL and PL–TL.

The transfer method (Utiyama and Isahara, 2007; Khalilov et al., 2008) translates the text in the SL to the PL using a SL–PL translation model and then to the TL using a PL–TL translation model. A source sentence s can be translated into n PL sentences. Each of these n sentences can then be translated into m TL sentences. Therefore we have $n \times m$ translation candidates which can be rescored using the translation scores from both the SL–PL and PL–TL models. The translation that gets the highest ranking is considered to be the best translation.

The synthetic corpus method (Gispert and Mariño, 2006; Bertoldi et al., 2008; Utiyama et al., 2008) obtains a SL–TL corpus using the SL–PL or the PL–TL corpora. One way to do this is to translate the PL sentences in the SL–PL corpus into TL with the PL–TL system. Another possibility is to translate the PL sentences in the PL–TL corpus into SL with the SL–PL system. Obviously, both methods could be applied and the two resulting synthetic corpora be merged into a single SL–TL corpus.

Wu and Wang (2009) compare the performance of the phrase table multiplication, transfer and synthetic corpus methods. They also present a hybrid method that combines RBMT and SMT to fill up the data gap, assuming the SL–PL and PL–TL corpora are independent. In this approach, RBMT systems are used to translate the PL sentences in the SL–PL or PL–TL corpus into TL or SL sentences, respectively. Then these synthetic corpora can be used to enrich the initial SL–PL and PL–TL corpora so that the SMT systems can take advantage of the availability of additional bilingual data.

System combination has also been exploited to improve pivot-based MT. Wu and Wang (2009) build systems following the three aforementioned approaches (phrase table multiplication, transfer and synthetic corpus) and combine the outputs produced by the different systems. Leusch et al. (2010) generate intermediate translations in several PLs, then translate them separately into the TL, and finally generate a consensus translation out of all of them.

The closest research strand to the work presented in this paper is the transfer method. The main difference is that the transfer method uses n-best lists and features from both systems and language pairs (SL–PL and PL–TL) in order to obtain the best translation while our proposal only has access to the n-best list and to internal features of the MT system for the language pair SL–PL. In our approach we treat the MT system for PL–TL as a black box. Because of this its application is wider: while the state-of-the-art requires access to the internals of this system, ours does not.

3 Methodology

In this section we introduce our methodology to perform pivot-based MT. We use a SMT system to translate from SL to PL (System1 from here onwards) and any kind of MT system to translate from PL to TL (System2).

For each source sentence we obtain the best n translations (n-best list) produced by System1 from SL to PL. Then we translate this n-best list from PL to TL using System2. Finally we select the best of these n translations in TL, using features from three different sources: (i) system internal features from System1, (ii) output from System1 (translations in PL) and (iii) output from System2 (translations in TL). In other words, we rerank the n-best list of translations in PL produced by System1 based on features of this system (and the translations in PL) but also using features from the output of System2 (the translations in TL).

The method uses two features in order to perform re-ranking:

- $-ts$, the translation score assigned by System1 to translations from SL into PL. This is an internal confidence measure common in SMT decoders. It is a log probability in the range $[-\infty, 0]$). We take its negative (range $[0, \infty]$); the lower the value the better the translation is considered to be.

- $\log_2(\mathrm{perp})$, the fluency of the translation produced by System2 in the TL. This is the logarithm of the perplexity given by a language model, in the range $[0, \infty]$. The lower the value, the better the fluency is considered to be.

The translations of the n-best list from PL to TL are scored using these two features according to

equation 1. The best translation (the one with the lowest score) is kept.

$$\text{score} = (-\text{ts}) \cdot \alpha + \log_2(\text{perp}) \cdot (1 - \alpha) \quad (1)$$

The parameter α, which can take any value in the interval $[0, 1]$, assigns complementary weights to the two features. An iterative process is followed in order to find the optimal value of α in the development set. The pseudocode of the algorithm is shown in Algorithm 1.

Algorithm 1 Find optimal α

```
score_best ← 0
α_best ← 0.5
α ← 0.5
depth ← 1
max_depth ← 16
while depth < max_depth do
    α_1 ← α + 0.5/2^depth
    α_2 ← α − 0.5/2^depth
    score_1 ← MT score at α_1
    score_2 ← MT score at α_2
    if score_1 = score_2 then
        break
    end if
    α ← α of max(score_1, score_2)
    if score_best < max(score_1, score_2) then
        score_best ← max(score_1, score_2)
        α_best ← α
    end if
    depth = depth + 1
end while
return α_best
```

The procedure starts with $\alpha = 0.5$ (the average value in the range $[0, 1]$). At each step it calculates the scores of the translations selected when using $\alpha_1 = \alpha - \frac{0.5}{2^{depth}}$ and $\alpha_2 = \alpha + \frac{0.5}{2^{depth}}$, sets as new α the one for which the MT score is higher between the two, increments the value of $depth$ and starts again. The procedure stops when the maximum value of $depth$ is reached or when both MT scores at α_1 and α_2 are equal. The best value of α selected during the procedure is then used to select the translations for the test set.

4 Evaluation

4.1 Experimental Setting

The experiments have been carried out for two scenarios (involving different languages and domains). The first scenario translates from Italian (SL) to Catalan (TL), passing through Spanish (PL). The test set consists of technical documentation data. We refer to this scenario as it–es–ca. The second scenario involves English as the SL, Bulgarian as the PL and Macedonian as the TL. The test set consists of newswire data. This scenario is referred to as en–bg–mk.

For System1 we use the phrase-based SMT Moses (Koehn et al., 2007)[1] in both scenarios. This system is trained and tuned on Europarl (Koehn, 2005)[2] Italian–Spanish for the first scenario and Europarl English–Bulgarian for the second. The corpora are tokenised and lowercased, and sentences where the source or the target is longer than 40 words are discarded. From the sentences extracted, we set aside 1,000 as development set for parameter tuning using MERT (Och, 2003) and we use the rest for training, i.e. 1,278,411 sentences for Italian–Spanish and 196,113 for English–Bulgarian.

For each SL sentence we obtain the n-best (up to 3,000) PL translations. We ensure that all translations in the n-best list are different (using the Moses parameter `distinct`). In order to obtain different translations, Moses considers the best $n \cdot m$ translations ($m = 200$), therefore it is not guaranteed that n different translations will be found (in fact, for some sentences we obtain a number of translations slightly lower than n). Apart from this, we use Moses' default settings. The translations in PL are recased using Moses' built-in recaser trained on the target side of the SL–PL training data.

For System2 in both scenarios we use Apertium, a RBMT system that uses a shallow-transfer engine (Forcada et al., 2011).[3] We use Apertium systems developed for Spanish–Catalan (Corbí-bellot et al., 2005) and Bulgarian–Macedonian (Rangelov, 2011).

The development and test sets for it–es–ca are extracted from the KDE4 multilingual documentation corpus in the OPUS project (Tiedemann, 2009).[4] The Italian–Catalan bilingual corpus contains 146,372 sentence pairs. We discarded sentence pairs where the source or target side is

[1] http://www.statmt.org/moses/
[2] http://www.statmt.org/europarl/
[3] http://www.apertium.org/
[4] http://urd.let.rug.nl/tiedeman/OPUS/KDE4v2.php

shorter than 10 words[5] or longer than 30,[6] where the difference of number of words between the source and target sentences is higher than 10% as well as sentences that contain URLs, Copyright notices and source code. This leads to a candidate set of 6,927 sentences, from it we randomly selected 1,000 sentences for development and 1,000 for test. The development set is used for the tuning procedure shown in Algorithm 1.

The development and test sets for en–bg–mk are taken from the SETimes multilingual corpus (Tyers and Alperen, 2010).[7] The development set contains 1,000 sentences whilst the test set holds 1,003 sentences.

5-gram word-based Language Models (LMs) are built for the TL with the IRSTLM toolkit (Federico et al., 2008)[8] using modified Kneser-Ney smoothing (Chen and Goodman, 1996). We use two monolingual corpora for the TL in the first scenario: one in-domain, from the KDE4 corpus, which consists of 53,776 sentences from the Catalan side which are not present in the aforementioned development nor test sets and one out-of-domain, consisting on up to 800,000 sentences gathered from news monolingual sources. A single monolingual corpus is used for the second scenario, in this case in-domain as it consists of sentences from the SETimes corpus. Up to 150,000 sentences are used.

Two automatic MT metrics are used to evaluate our approach, these are BLEU (Papineni et al., 2002) and NIST (Doddington, 2002). Statistical significance tests are carried out using paired bootstrap resampling (Koehn, 2004) with ARK's code.[9] Sentence-level scores for the oracles are computed with smoothed BLEU.[10] BLEU is also used as the MT score to tune the procedure shown in Algorithm 1.

4.2 Experiments

4.2.1 Baseline and Oracles

First we establish the baseline, which consists of combining the two MT systems sequentially in a cascade fashion, i.e. for each source sentence this is the translation by System2 of the best translation obtained by System1.

In order to determine the margin for improvement that can be attained when taking into consideration all the translations in the n-best list, we have developed an oracle system which yields the maximum reachable score. The oracle is based on the one described in (Och et al., 2004) where only one reference translation is available; for each sentence, it translates all the translations in the n-best list in PL with System2 into TL, scores BLEU at sentence level, and picks the translation with the highest score. Finally it builds a set in the TL with the translation picked for each sentence and scores BLEU at document level.

Table 1 shows the BLEU and NIST scores for the baselines and the oracles for both scenarios and for different sizes of the PL n-best list (100, 1,000, 2,000 and 3,000). Apart from the absolute scores, for each metric and oracle we report its relative improvement over the baseline (in columns labelled $\Delta\%$).

For the first scenario, the oracle is almost 6 absolute points higher than the baseline (0.2878 vs 0.2289) according to BLEU with just 100 sentences in the n-best list. Incrementing the list to 1,000 sentences yields approximately 2.5 additional BLEU points (0.3133 vs 0.2878). This can be incremented by almost 3.5 further points by considering the top 2,000 translations (0.3476 vs 0.3133); this is 11.87 absolute points over the baseline (0.3476 vs 0.2289) or a 51.86% relative improvement. In comparison, having access to the best 3,000 translations brings about only modest further improvements: about half a BLEU point over the oracle that uses 2,000 translations (0.3525 vs 0.3476).

Sustained improvements are reported also by NIST, although they are lower (the maximum relative improvement is 24.29%).

A similar pattern is observed for the second scenario. In this case the relative improvements are even higher; 44.66% for 100-best, 68.02% for 1,000-best and almost 75% both for 2,000-best and 3,000-best in terms of BLEU.

The results clearly indicate that methods that exploit the n-best list to translate from PL to TL have potential to improve performance considerably over the baseline. Given the similar results obtained when using 2,000-best and 3,000-best lists and taking into consideration the computa-

[5] Those sentences are usually not fluent sentences but menu items, isolated terms, etc.

[6] Such long sentences are not ideal for potential tasks such as word alignment.

[7] http://opus.lingfil.uu.se/SETIMES.php

[8] http://hlt.fbk.eu/en/irstlm

[9] http://www.ark.cs.cmu.edu/MT/

[10] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl

Scenario	MT system	n-best size	BLEU	$\Delta\%$	NIST	$\Delta\%$
it–es–ca	Baseline	-	0.2289	0.00	5.6706	0.00
	Oracle	100	0.2878	25.73	6.2778	10.70
	Oracle	1,000	0.3133	36.87	6.5884	16.19
	Oracle	2,000	0.3476	51.86	6.9909	23.28
	Oracle	3,000	0.3525	54.00	7.0482	24.29
en–bg–mk	Baseline	-	0.1104	0.00	4.3274	0.00
	Oracle	100	0.1597	44.66	5.1126	18.14
	Oracle	1,000	0.1855	68.02	5.4734	26.48
	Oracle	2,000	0.1931	74.90	5.5646	28.59
	Oracle	3,000	0.1927	74.55	5.4222	25.30

Table 1: MT scores for the baselines and oracles

tional cost involved, we consider lists of 2,000-best sentences for the rest of the experiments.

4.2.2 Pivot Systems

We now turn to our pivot systems that rank translation output according to SL–PL translation score and TL perplexity (rather than oracle selection). We evaluate the pivot method using different LMs. For the first scenario there are four systems that use out-of-domain LMs (newswire) made up of a different number of sentences: News-100k (100,000), News-200k (200,000), News-500k (500,000) and News-800k (800,000). Finally there is a system that uses an in-domain LM, KDE-50k, derived from 50,000 sentences of the KDE corpus.

Regarding the second scenario, we have built three in-domain LMs, using 50,000 (SET-50k), 100,000 (SET-100k) and 150,000 (SET-150k) sentences from the SETimes corpus. The results obtained according to the BLEU and NIST metrics and the improvements over the baseline using 2,000-best lists are shown in Table 2.

The results obtained by the pivot systems using out-of-domain LMs are slightly higher than the baseline (except for the BLEU score for the system News-100K, which is slightly lower). However, only the NIST score for the system News-800K is significantly better than the baseline ($p = 0.05$).

Although using a much smaller LM, the system that uses an in-domain LM made up of 50,000 sentences from the KDE corpus reaches notably higher scores, achieving almost 3 absolute BLEU points over the baseline (0.2561 vs 0.2289, or 11.88% relative improvement). Both the BLEU and NIST scores are statistically significantly better than the baseline ($p = 0.01$). As the testset comes from a very specific and technical domain, having a LM from that same domain (even if it is rather small) to re-rank the translations proves to be very useful.

All the pivot systems for the second scenario obtain significantly better scores compared to the baseline ($p = 0.01$). The highest improvement is achieved by SET-150k (13.32% relative and 1.47 absolute in terms of BLEU).

For all the runs using out-of-domain LMs, the value of α is very high (the values range from 0.9453 to 0.9824), meaning that almost all the weight to choose the best translation is given to the feature that measures translation score in PL, while the one that measures fluency in the TL remains marginal ($1 - \alpha$, see equation 1). As the original n-best list is sorted by translation score, we can expect that in these runs most sentences selected are very near the top of this list; hence the results do not differ much from the baseline. Conversely, the value of α is considerably lower for the runs using in-domain LMs; 0.8125 for KDE-50k in the first scenario, even lower values for the second scenario, in the range $[0.5390, 0.6250]$. This suggests that the fluency in TL plays a more important role in the selection of translations from the n-best list when using an in-domain LM. More details on this are provided in Section 4.3, where the results are analysed.

4.3 Analysis

We provide an analysis of all the systems evaluated by looking at the distribution of the ranking positions of the sentences selected in the n-best lists. For each of the systems we report on the following statistical measures:

- Minimum (*min*), the rank of the highest sen-

Scenario	MT system	α	BLEU	$\Delta\%$	NIST	$\Delta\%$
it–es–ca	News-100k	0.9453	0.2283	-0.26	5.6739	0.05
	News-200k	0.9551	0.2301	0.52	5.6909	0.35
	News-500k	0.9824	0.2299	0.43	5.6844	0.24
	News-800k	0.9824	0.2300	0.48	**5.6853**	0.25
	KDE-50k	0.8125	**0.2561**	11.88	**6.0130**	6.03
en–bg–mk	SET-50k	0.6250	**0.1238**	12.14	**4.4060**	1.81
	SET-100k	0.5390	**0.1245**	12.77	**4.4085**	1.87
	SET-150k	0.5469	**0.1251**	13.32	**4.4115**	1.94

Table 2: MT scores for the pivot method

tence picked by the method.

- Maximum (*max*), the rank of the lowest sentence picked by the method.
- *Mean*, the average value of the ranking positions of the sentences chosen by the method.
- Standard deviation (*stddev*), the standard deviation from the average of the sentences selected.

Table 3 provides these values for the oracle systems over the different sizes of the n-best list (100, 1,000, 2,000 and 3,000). The high values of the *max* and *stddev* show that the oracles select sentences from the whole range of translations available in the n-best lists. At least one of the lowest ranked translations was taken for 100-best (*max* 99), while one very near the end of the list was selected from 1,000-best (*max* 999 and 998), 2,000-best (*max* 1,990 and 1,995) and 3,000-best (*max* 2,997 and 2,995).

	n-best size	min	max	mean	stddev
it–es–ca	100	0	99	17.53	27.34
	1,000	0	999	192.31	279.54
	2,000	0	1,990	472.34	579.35
	3,000	0	2,997	716.36	880.05
en–bg–mk	100	0	99	39.66	29.95
	1,000	0	998	400.10	297.25
	2,000	0	1,995	818.78	611.48
	3,000	0	2,995	1,002.88	862.01

Table 3: Statistics for oracles

Table 4 shows the statistics for the pivot systems. The previous hypothesis that systems using out-of-domain LMs select most sentences very near the top due to the very high value of α is corroborated here by the statistical measures. Although the systems have access to 2,000 translations, the lowest ranked sentence picked by one of the systems using an out-of-domain LM (News-100k) is at position 133, while two of them (News-500k and News-800k) do not even select any translation beyond a rank as high as 9. The very low values of the mean, which range from 0.18 to 1.51, indicate that most translations are taken from the very highest ranked sentences.

The statistics are very different for the systems that use in-domain LMs. The values in this case resemble much more the pattern observed for the statistics shown for the oracles (Table 3). These systems do extract translations from all the range of ranks as indicated by the values of the lowest translation selected (1,990 for KDE-50k, 1,998 for systems using LMs built on SETimes), which are figures similar to those reported for the 2,000-best oracles (1,990 for es–it–ca and 1,995 for en–bg–mk). The high values of both the mean (214.65 for the first scenario and $[615.25, 801.17]$ for the second) and the standard deviation (420.99 for the first scenario and $[585.17, 593.20]$ for the second) confirm this trend.

	LM	min	max	mean	stddev
it–es–ca	News-100k	0	133	1.51	6.84
	News-200k	0	88	0.97	4.00
	News-500k	0	9	0.19	0.69
	News-800k	0	9	0.18	0.66
	KDE-50k	0	1,990	214.65	420.99
en–bg–mk	SET-50k	0	1,998	615.25	585.17
	SET-100k	0	1,998	801.17	593.20
	SET-150k	0	1,998	784.55	588.19

Table 4: Statistics for pivot systems

5 Conclusions and Future Work

This paper has presented, to the best of our knowledge, the first pivot-based MT methodology in which the second MT system is treated as a black box.

Compared to the state-of-the-art, our methodology is applicable to a broader set of scenarios, as no access to the internals of the second system is required. This opens the applicability of MT pivot-based approaches to target languages for which no suitable bilingual corpora to build PL–TL SMT systems are available, as long as there is any kind of PL–TL MT system available.

We have presented a method which exploits two types of features: internal of the system that translates from SL to PL, and from the output of the final TL translation. An algorithm that uses two features (translation score of the first system and perplexity of the final translation) is presented. Complementary weights are given to the features and the optimal values are tuned on the development set. The source code that implements this procedure is available under the GPL-v3 license.[11] The data used in the experiments is also available.

We have evaluated this approach comparing it to a baseline, which consists of translating the input sentences using the two MT systems sequentially. We have experimented with two scenarios that involve different language families and domains, technical documentation in Romance languages and newswire in Slavic languages, obtaining up to 11.88% and 13.32% relative improvements in terms of BLEU, respectively.

Using just two features yields significant improvements for both scenarios, but the scores obtained by the oracles indicate that there is still considerable room for improvement, e.g. for the 2,000-best configuration, the best pivot-based systems (KDE-50k and SET-150k) obtain 0.2561 and 0.1251 BLEU points, while the oracles yield 0.3476 and 0.1931 (over 9 absolute points better in the first case and nearly 7 in the second).

Therefore we plan to extend the methodology presented here in several ways. First, we will explore other possible features, looking for example at features successfully used in other MT-related tasks, such as (He et al., 2010). Second, we will experiment with other algorithms that allow us to combine an arbitrary number of features in order to rescore the translations. Finally, the methodology will be applied to different MT systems, language pairs and domains in order to further validate the applicability of this approach.

[11] `http://nclt.computing.dcu.ie/~atoral/#Resources`

Acknowledgements

We would like to thank Francis Tyers and Tihomir Rangelov for their help and ideas regarding the experiment that involves Bulgarian and Macedonian. This work has been funded by the European Association for Machine Translation through its 2011 sponsorship of activities program.

References

Bertoldi, Nicola, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-Based Statistical Machine Translation with Pivot Languages. In *Proc. of the International Workshop on Spoken Language Translation*, pages 143–149, Hawaii, USA.

Chen, Stanley F. and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cohn, Trevor and Mirella Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 728–735, Prague, Czech Republic.

Corbí-bellot, Antonio M., Mikel L. Forcada, Sergio Ortiz-rojas, Juan Antonio Pérez, Gema Ramírez-sánchez, Felipe Sánchez-martínez, Iñaki Alegria, and Kepa Sarasola. 2005. An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain. In *In Proceedings of the 10th European Associtation for Machine Translation Conference*, pages 79–86.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Federico, Marcello, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *INTERSPEECH*, pages 1618–1621. ISCA.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez Felipe Sánchez-Martínez, and Francis M. Tyers. 2011.

Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144. Special Issue: Free/Open-Source Machine Translation.

Gispert, Adrià De and José B. Mariño. 2006. Statistical machine translation without parallel corpus: bridging through Spanish. In *Processdings of LREC 5th Workshop on Strategies for developing Machine Translation for Minority Languages*, pages 65–68.

He, Yifan, Yanjun Ma, Josef van Genabith, and Andy Way. 2010. Bridging SMT and TM with translation recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 622–630, Stroudsburg, PA, USA. Association for Computational Linguistics.

Khalilov, M., Marta R. Costa-jussà, José A. R. Fonollosa, Rafael E. Banchs, B. Chen, M. Zhang, A. Aw, H. Li, José B. Mariño, Adolfo Hernández, and Carlos A. Henríquez Q. 2008. The TALP and I2R SMT Systems for IWSLT 2008. In *International Workshop on Spoken Language Translation (IWSLT 2008)*, pages 116–123, Hawaii, USA.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Prague, Czech Republic.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In Lin, Dekang and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Leusch, Gregor, Aurélien Max, Josep M. Crego, and Hermann Ney. 2010. Multi-Pivot Translation by System Combination. In Federico, Marcello, Ian Lane, Michael Paul, and François Yvon, editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 299–306.

Och, Franz Josef, Daniel Gildea, Sanjeev Khudanpur, Kenji Yamada, Alex Fraser, Shankar Kumar, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. Final report of johns hopkins 2003 summer workshop on syntax for statistical machine translation. Technical report.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania, USA.

Rangelov, Tihomir. 2011. Rule-based machine translation between Bulgarian and Macedonian. In *Proceedings of the Second International Workshop on Free/Open-SourceRule-Based Machine Translation*, pages 53–60, Barcelona, Spain.

Tiedemann, Jörg. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In Nicolov, N., K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Tyers, Francis M. and Murat Serdar Alperen. 2010. South-East European Times: A parallel corpus of the Balkan languages. In *Proceedings of the Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages, LREC 2010*.

Utiyama, Masao and Hitoshi Isahara. 2007. A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York, April. Association for Computational Linguistics.

Utiyama, Masao, Andrew Finch, Hideo Okuma, Michael Paul, Hailong Cao, Hirofumi Yamamoto, Keiji Yasuda, and Eiichiro Sumita. 2008. The NICT/ATR Speech Translation System for IWSLT 2008. In *Proc. of the International Workshop on Spoken Language Translation*, pages 77–84, Hawaii, USA.

Wu, Hua and Haifeng Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 856–863, Prague, Czech Republic, June. Association for Computational Linguistics.

Wu, Hua and Haifeng Wang. 2009. Revisiting Pivot Language Approach for Machine Translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 154–162, Suntec, Singapore, August. Association for Computational Linguistics.