Multi-Pivot Translation by System Combination

Gregor Leusch[*], *Aurélien Max*[†‡], *Josep Maria Crego*[†], *Hermann Ney*[*]

[*]RWTH Aachen University, Aachen, Germany
[†]LIMSI-CNRS, Orsay, France
[‡] Univ. Paris Sud, Orsay, France
{leusch,ney}@cs.rwth-aachen.de, {aurelien.max,jmcrego}@limsi.fr

Abstract

This paper describes a technique to exploit multiple pivot languages when using machine translation (MT) on language pairs with scarce bilingual resources, or where no translation system for a language pair is available. The principal idea is to generate intermediate translations in several pivot languages, translate them separately into the target language, and generate a consensus translation out of these using MT system combination techniques. Our technique can also be applied when a translation system for a language pair is available, but is limited in its translation accuracy because of scarce resources.

Using statistical MT systems for the 11 different languages of Europarl, we show experimentally that a direct translation system can be replaced by this pivot approach without a loss in translation quality if about six pivot languages are available. Furthermore, we can already improve an existing MT system by adding two pivot systems to it. The maximum improvement was found to be 1.4% abs. in BLEU in our experiments for 8 or more pivot languages.

1. Introduction

Over the last decade, statistical machine translation (SMT) has shown to produce adequate translation results for language pairs and domains where large amounts of mono- and bilingual training data are available. Unfortunately, if one wants to be able to translate from many possible source languages into many possible target languages, separate MT systems for each possible pair of source and target language have to be trained, on bilingual data in this specific language pair. Quite often this is not possible, especially where rare or unrelated languages are involved. Significant amounts of bilingual in-domain training data may be unavailable; the number of systems to train and to tune may be too high. One approach to overcome this problem has been proposed e.g. by Utiyama and Isahara [1]: A third, more frequent language is utilized as a *pivot* or *bridge language*. Ideally, sufficient bilingual language resources are available for both the pair of source and pivot language, and for the pair of pivot and target language. The final translation is then obtained by going via the bridge language, either by generating full translations of the source sentence in this bridge language, or by using the bilingual data to build translation models for the source–target language pair. The disadvantage of this approach is that both the translation into the pivot language, and the translation into the target language are error-prone – and typically, these errors add up. As a result, on comparable training resources, we can expect the translation quality of a pivot system to be significantly lower that the quality of a "direct" system[1].

Since Kay [2] has first predicted the usefulness of multilingual resources, several approaches have been proposed to utilize resources and data available in more than two languages for MT. *Multi-source machine translation* [3, 4, 5] denotes techniques to translate documents which are available in two or more source languages. One approach that has recently been shown to be very effective [4, 6] is to use individual bilingual MT systems to translate the source documents independently of each other into one document each in the target language, and then to use MT system combination (ibid.) to generate a consensus translation out of these different target translations.

In this paper, we will investigate to what extent

[1]Within this paper, we use the term "direct system" to denote a (statistical) MT system that has been trained on a bilingual corpus between source and target language, and does not utilize any pivot or bridge languages.

this multi-source technique can be applied to extend the pivot language approach. We will use multiple pivot languages simultaneously, instead of only one. Our intention is that we expect certain translation errors to be limited to specific language pairs, and that they will thus be voted down in system combination. We will also look at scenarios where both a direct source-to-target translation and pivot translations are available. Here, our plan is to improve translation output by using these pivot translations to indirectly cancel out noise in the translation models, and also to indirectly avoid ambiguities with the help of the language models in the pivot languages. In these scenarios, our approach turns out to be similar to the approach of Koehn et al. [7].

The rest of this paper is organized as follows: We will review related work in pivot and multi-source MT in Section 2. Our approach to multi-source translation will be briefly presented in Section 3. We will then describe how this approach can be used for multi-pivot translation in Section 4. An experimental framework to study this approach will be introduced in Section 5, and its results will be shown in Section 6. We will conclude this paper with a final discussion in Section 7.

2. Related Work

Over the last years, pivot or bridge languages have been utilized at several steps of statistical MT. Kumar et al. [8] have shown how bridge languages can be used to improve word alignments, and thus to improve an (existing) direct source-target translation model. A different approach with the same objective is to use bridge languages to generate paraphrases for the training data [9, 10]. Bridge languages have alsobeen used to find paraphrases of the source text that yield improvement in translation performance [9, 11]. Other approaches do not require an existing translation model between source and target, but generate such models from one [1] or several [12] pivot translation models. MT systems for the pivot languages can also be used to build bilingual training data for the source-target pair [13]. Probably the most obvious way for using multilingual resources in MT is to generate an individual translation for each source sentence in the pivot language, which is then translated into the target language. We will extend this approach in this paper.

Multi-source translation has probably been first described as a MT task by Och and Ney [3], although the utilization of multiple languages within the translation process goes back at least to Kay [2]. Och's methods have later been revisited and improved by Schwartz [14]. Here, Schwartz describes this also in the context of a consensus translation/system combination approach, although in the style of hypothesis-selection system combination. Confusion-network and lattice-based approaches have been successfully applied for multi-source translations as well [4, 5, 6].

Eisele [15] mentioned the idea to use multiple pivot translations to overcome both translation errors in individual pivot systems, and the lack of bilingual data for rare language pairs, although there is no detailed description of the combination process in his paper. A hypothesis selection approach for this has been described by Wu and Wang [13].

Koehn et al. [7] later described a confusion network approach similar to the one used in this paper, though to improve an existing "direct" system. On a large corpus of multilingual law texts, they investigated experimentally whether the output of an existing MT system can be improved using multiple (three to six) pivot systems by creating a consensus translation. Focussing on English as their target language, Koehn et al. reported improvements of up to 0.9% abs. in BLEU. They did not investigate whether it is possible to omit the "direct" system in this approach.

3. Multi-source translation by system combination

Combining outputs from different systems was first shown to produce superior results in automatic speech recognition (ASR). Voting schemes like the ROVER approach of Fiscus [16] create confusion networks (CNs) from the output of different ASR systems for the same audio input. The consensus recognition hypothesis is generated by weighted majority voting.

This approach has later been adapted to MT as well [17]. In this paper, we follow the approach of Matusov et al [4, 18]: An unsupervised monolingual word alignment is trained between all pairs of hypotheses for each source sentence using the GIZA++ toolkit [19]. These alignments are then used to reorder all individual hypotheses to one selected ("primary") hypothesis, which defines the word order in the consensus translation. A CN is then generated from these reordered hypotheses. As there is no obvious way to determine the best primary hypothesis, separate CNs are generated for all possible primary hypotheses, which are then combined

to a single word lattice. This lattice is then rescored using system weights and a language model (LM). The latter is typically trained on the input hypotheses, not on a "classical" training corpus. As the last step, the best path within this lattice is calculated, and the corresponding sentence (after removal of empty arcs) is then considered to be the consensus translation of the input hypotheses.

As proposed by Matusov et al. [4] and Schroeder et al. [5], this approach can also be used for multi-source translation, where each document to be translated is simultaneously available in several languages: Each source sentence is translated separately by a MT system trained for this source language and the common target language. The individual hypotheses in the target language are then combined by the CN-based system combination approach. Note than since our specific CN-based approach to system combination does not depend on a (single) source language, it can thus be used independently of the individual source sentences, and consequently in multi-source translation.

4. Multi-pivot translations

The multi-source approach from the previous section can also be applied to pivot translation – more precisely to the pivot approach where an explicit intermediate translation of each source sentence is being generated. The difference here is that we can now use multiple pivot languages: First, we use MT engines from the source language to the different pivot languages to generate pivot translations. We then combine their translations into the target language using the technique described in Section 3. The structure of this approach is shown in Figure 1. The advantage over single-pivot translation is that many translation errors will likely be canceled out, and that there is a higher chance to resolve ambiguities in the intermediate translations.

In a scenario with a large number of possible source and target sentences, this leads to a significant reduction in the number of required systems, despite the necessity to build systems to and from all pivot languages: Given n source/target languages, and m pivot languages, the number of required systems would be $n(n-1)$ in the non-pivot scenario, and $m(2n-m-1)$ in the pivot-scenario. With, e.g., $n = 23$ EU languages, and $m = 4$ pivot languages, this would result in 506 language pairs, but only 164 required systems.

Table 1: Corpus statistics for the experimental setup.

	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

5. Experimental setup

5.1. Training and Development data

For experimenting with our approach, we built translation systems to serve as direct or pivot systems using a phrase-based MT engine for several language pairs of the Europarl corpus [20], which is available in 11 languages: Danish (da), German (de), English (en) , Spanish (es), Finnish (fi), French (fr), Greek (el), Italian (it), Dutch (nl), Portuguese (pt) and Swedish (sv). We also decided to study three source–target language pairs, two for which translation accuracy, as measured by automatic metrics, is moderate, (de–en) and (fr–de), and one for which translation accuracy, is much higher. (fr–en).

This allowed us to check whether the improvements provided by our method carry over even in situations where the baseline is high; conversely, it also allows us to assess whether the proposed techniques are applicable when the baseline is average and poor.

In order to measure the contribution of each of the auxiliary languages we decided to use a training corpus common for all language pairs. We used the English side as the bridge language to collect exactly the same sentences for each language pair, collecting up to 320,304 sentence pairs in all language pairs. Some statistics on the used data are shown in Table 1.

Development and test data for the this condition were obtained by leaving out respectively 500 and 1000 sentences from the common subset (same sentences for all languages).

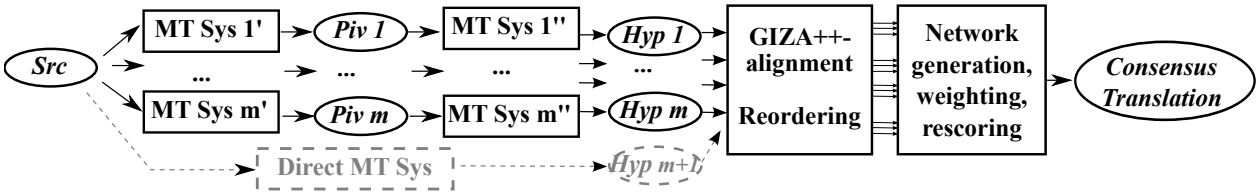

Figure 1: Structure of the multipivot system

5.2. Translation engines

The translation engine for these experiments implements the n-gram-based approach to statistical machine translation detailed by Marino et al. [21]. The overall translation accuracy is comparable to state-of-the-art phrase-based translation engines such as the MOSES system [22].

In a nutshell, the translation model is implemented as a stochastic finite-state transducer trained using a n-gram language model of *(source,target)* pairs [23]. Training such a model requires to reorder source sentences so as to match the target word order. Reordering hypotheses are computed before decoding takes place via a stochastic finite-state automaton that builds a lattice with the most promising hypotheses according to a set of rewrite rules previously collected from the training bi-texts using the word alignments.

In addition to the bilingual n-gram model, our SMT system implements eight additional models which are linearly combined following a discriminative modeling framework [24]: two *lexicalized reordering* models [25], which attempt to model the orientation of the current translation unit according to the previous as well as the ordering of the next unit with respect to the current unit, a *target-language model* which provides information about the target language structure and fluency; two *lexicon models*, which constitute complementary translation models computed for each given tuple; a 'weak' distance-based *distortion model*; and finally a *word-bonus model* and a *tuple-bonus model* which are used in order to compensate for the system preference for short translations. For this study, we used 3-gram bilingual tuple and 3-gram target language models built using Kneser-Ney smoothing [26]; training was performed with the SRI language modeling toolkit [27].

After preprocessing the corpora with standard tokenization tools, word-to-word GIZA++ [19] alignments are performed in both directions, followed by the *grow-diag-final-and* heuristic [28].

5.3. Experiments

The two principal research questions we wanted to answer with the experiments for this paper were: Can we use the multi-pivot approach *instead* of a direct source–target system, with comparable translation scores? And secondly, can we use the multi-pivot approach to *improve* the output of an existing direct system?

The former question is most relevant for the "matrix" scenario sketched in Section 3, where we have a large number of possible source and target languages, and do not want to build separate translation systems for each individual pair – either because we want to save on time and space resources, or because we do not have large enough amounts of training data for all pairs. Our training data represents more the first rationale, because we can generally expect data rather to exist for two of our source languages (en and fr) than for most of the pivot languages.

The latter question targets more on scenarios where bilingual training data is limited (at least for the source–target pair itself), and where we seek to improve translation results by exploiting all data we have. Here, one might expect that the selected training data we used is not likely to show improvements: Since the source part is the same for all source–pivot training corpora, and so is the target part of all pivot–target corpora, no pivot systems should have been able to learn phrases which are "new" compared to a direct system. Almost all improvements we might see will come from disambiguation of ambiguous phrases, improved reordering, or other effects from pivot language modelling.

In the experiments, we first generated translations of the dev and test set directly from the source to the target language for each of the language pairs de–en, fr–

Table 2: Results on TEST of multi-pivot translation with and without direct system for FR–EN.

	single		pivot only		direct + pivot	
system	BLEU	TER	BLEU	TER	BLEU	TER
direct	29.60	54.69	—			
via ES	27.85	57.18				
via PT	27.36	56.91			29.54	54.09
via EL	25.41	60.08	28.72	55.38	29.97	53.87
via IT	25.88	56.54	29.01	54.20	30.05	53.79
via DA	25.58	60.08	29.46	54.30	30.87	53.37
via NL	25.25	59.38	29.92	53.89	30.48	53.47
via DE	23.61	60.48	29.58	53.55	30.41	53.30
via SV	23.84	57.18	29.75	53.47	30.57	53.27
via FI	19.25	69.16	29.78	53.57	30.78	53.26

Table 3: Results on TEST of multi-pivot translation with and without direct system for DE–EN.

	single		pivot only		direct + pivot	
system	BLEU	TER	BLEU	TER	BLEU	TER
direct	24.76	58.70	—			
via NL	22.74	61.59				
via DA	22.83	63.40			24.63	57.98
via PT	22.02	63.63	23.60	59.58	25.36	57.04
via FR	21.64	62.95	24.47	58.50	25.51	56.74
via ES	21.34	62.55	24.43	58.04	25.37	56.84
via EL	20.96	63.71	24.66	57.50	25.30	56.84
via SV	21.44	61.13	25.05	57.04	25.54	56.68
via FI	18.12	68.00	24.86	57.33	25.33	56.75
via IT	18.19	61.32	25.20	57.76	25.32	56.67

en, and fr–de. Next, we generated the pivot part – translations from the source to each individual pivot language, and from there to the common target language. We then calculated the BLEU and TER scores for each pivot and direct translation in the target language, and ordered the systems by their BLEU score. For the first scenario, we combined the topmost 3, 4, ..., 9 pivot translations with each other[2]. For the second scenario, we combined the direct translation with the topmost 2, 3, ..., 9 pivot translations. All combination parameters were tuned for an optimum (TER-BLEU) score.

6. Experimental results

Tables 2, 3, and 4 show the TER and BLEU scores for the experiments on the three language pairs. All scores are case insensitive, and have been calculated on the blind test set.

The first two columns show the scores for the direct system (first line), and each individual pivot language (i.e., the score from the full translation source–pivot–target). We see that the best pivot system is only between 1.2 and 2.0 BLEU points worse than the direct system, and between 0.4 and 0.9 points in TER.

The results from the pivot-only experiments – which correspond to the "instead of direct" or "matrix" scenario – can be found in the two center columns of these tables. Depending on the language pair, we find that a combination of four (fr–en) to six (de–en) pivot languages leads to translation scores which are on par with or better than those of the original system. Adding more pivot languages improves the translation results even more, even though a maximum is reached at six systems for fr–en. For all three language pairs, already a combination of four pivot systems shows a BLEU score not too far away from the direct system, at a significantly better TER score.

The two rightmost columns then show the results from the "improvement" scenario, i.e. the combination of the direct and several pivot systems. Here, already the addition of two pivot systems improves the translation results over the direct system, both in BLEU and TER. Adding more pivot system improves the scores even more, up to a peak of six to eight systems (fr–en, de–en). The maximum improvement is +1.3/-1.3 abs. in BLEU/TER for fr–en, +0.8/-2.0 for de–en, and +1.4/-4.0 for fr–de. For both sets of experiments, we see relative improvements for all three language pairs, independent of the translation accuracy of the baseline

Table 4: Results on TEST of multi-pivot translation with and without direct system for FR–DE.

	single		pivot only		direct + pivot	
system	BLEU	TER	BLEU	TER	BLEU	TER
direct	18.20	68.70	—			
via ES	16.99	71.00				
via NL	16.98	69.09			18.85	66.87
via PT	16.55	69.68	18.01	66.28	18.86	65.40
via IT	16.72	70.56	18.40	65.69	18.82	65.08
via EN	15.91	71.46	18.48	64.96	19.15	65.89
via DA	16.49	70.10	18.71	65.93	19.30	65.14
via EL	16.09	70.53	18.97	64.31	19.59	64.74
via SV	13.99	73.53	19.29	64.63	19.63	64.45
via FI	11.57	82.82	19.41	64.52	19.63	64.16

[2]Note that we need at least three different translations to apply our combination approach.

Source
les réflexions étranges de ceux qui trouvent que ceux qui ne pratiquent pas d'enrichissement devraient recevoir des droits de plantation supplémentaires sont quand même complètement débiles!

Reference translation
and the strange idea some people have that wine growers not using enrichment should be given additional planting rights is simply crazy.

Direct translation fr–en
the strange ideas of those who find that those who do not practise should receive additional planting rights are still completely débiles!

Single pivot translation fr–(es)–en
the comments of those who are those who are not being enrichment should receive additional planting rights are completely mental anyway!

Multi-pivot translation fr–(es+pt+el+it+da+nl)–en
the strange of those who think that those who do not practise enrichment should receive additional planting rights are débiles!

Multi-pivot plus direct translation fr–(en+es+pt+el+it+da)–en
the strange ideas of those who think that those who do not practise enrichment should receive additional planting rights are completely débiles!

Figure 2: Example of a translation from French to English using the different direct, pivot, and multi-pivot systems.

system. This indicates that our approach might show improvements even for larger data sets.

Figure 2 shows an example from the French to English test set. Listed are source, reference translation, and translations from the different MT/pivot systems. We see that the direct system is lacking correct translations for enrichissement/enrichment and débiles/crazy. The pivot system translates both correctly, but fails to build the proper sentence structure, and translates réflexions/ideas wrongly. Multi-pivot-only translation, using six pivot languages here, fails again at réflexions and on débiles, but gets the sentence structure correctly. Finally, multi-pivot plus direct translation gets everything right, except of débiles – obviously, the correct translation mental had been voted down by the majority of the systems.

7. Conclusions and Outlook

In this work, we have presented a novel method to improve machine translation on settings with many possible source and target languages, or for language pairs with scarce bilingual resources. Intermediate or pivot translations in several different languages are generated, and then translated separately into the target language. These translations are then combined using a confusion-network based system combination technique. Experimental results with up to ten different pivot languages were performed for three source–target language pairs. The MT systems were trained on multi-parallel data from the Europarl corpus. The experiments confirmed that a combination of about 6 pivot languages (depending on the corpus pair) can replace a "direct", i.e. non-pivot MT system in terms of translation quality. This would allow to save separate MT systems for several language pairs in the scenario above.

We further showed that the output of an existing "direct" translation system can even be improved if we combine the output of additional pivot systems with it. For this, even the addition of only two pivot systems shows an improvement.

As our presented system combination method is completely independent of the upstream MT engines, a possible extension we are planning to investigate next is whether the translation results can improved further by having more than one translation engine per language pair, or more than a single best translation in the pivot language. Then, fewer pivot languages could be required.

One caveat of our method that we have not addressed so far is that while it saves the setup of a full MT engine for each individual language pair, it still requires an individual tuning of the system combination

weights and parameters for each pair. Further research will be needed to investigate whether it is possible to determine a set of parameters which can be kept fixed at least per target language, and thus reduce the number of tuning runs from quadratic to linear in the number of source/target languages.

Finally, our work focussed on frequent languages as source and target language, and some less frequent languages as pivot languages – even though this did not matter in our case, because all training corpora had the same size and cope for all MT systems. For scenarios where pivoting is used because of a lack of bilingual resources, a repetition of these experiments the other way round might be more appropriate, with high-volume languages as pivot, and rare languages as source and/or target. Especially data sets which are not orthogonal in the sense that there are no sentence pairs which are unique to just a subset of languages need to be investigated. In addition, we are interested to identify methods for the selection of the best pivot languages for a given training data matrix and/or language pair.

8. Acknowledgments

Part of this research was conducted in the scope of the European Associated Laboratories IMMI-Labs. This work was partly realized as part of the Quaero Programme, funded by OSEO, the French State agency for innovation. This work was partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-08-C-0110.

9. References

[1] M. Utiyama and H. Isahara, "A comparison of pivot methods for phrase-based statistical machine translation," in *Proceedings of the conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (NAACL-HLT)*, 2007, pp. 484–491.

[2] M. Kay, "The proper place of men and machines in language translation," *Machine Translation*, vol. 12, no. 1-2, pp. 3–23, 1997, first appeared as a Xerox PARC working paper in 1980.

[3] F. J. Och and H. Ney, "Statistical multi-source translation," in *Proc. of the MT Summit VIII*, 2001, pp. 253–258.

[4] E. Matusov, N. Ueffing, and H. Ney, "Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment," in *Proc. of the European chapter of the Association for Computational Linguistics (EACL)*, 2006, pp. 33–40.

[5] J. Schroeder, T. Cohn, and P. Koehn, "Word lattices for multi-source translation," in *Proc. of the 12th Conference of the European chapter of the Association for Computational Linguistics (EACL)*, March 2009, pp. 719–727.

[6] G. Leusch, E. Matusov, and H. Ney, "The RWTH system combination system for WMT 2009," in *Proc. of the Fourth Workshop on Statistical Machine Translation (WMT)*. Association for Computational Linguistics, Mar. 2009, pp. 56–60.

[7] P. Koehn, A. Birch, and R. Steinberger, "462 Machine Translation Systems for Europe," in *Proc. of the MT Summit XII*, August 2009, pp. 65–72.

[8] S. Kumar, F. J. Och, and W. Macherey, "Improving word alignment with bridge languages," in *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, June 2007, pp. 42–50.

[9] C. Callison-Burch, P. Koehn, and M. Osborne, "Improved statistical machine translation using paraphrases," in *Proceedings of the conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (NAACL-HLT)*, 2006, pp. 17–24.

[10] A. Max, "Sub-sentencial paraphrasing by contextual pivot translation," in *Proceedings of the 2009 Workshop on Applied Textual Inference*, Suntec, Singapore, 2009, pp. 18–26.

[11] ——, "Example-based paraphrasing for improved phrase-based statistical machine translation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, 2010, pp. 656–666.

[12] T. Cohn and M. Lapata, "Machine translation by triangulation: Making effective use of multi-parallel corpora," in *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 45, no. 1, 2007, p. 728.

[13] H. Wu and H. Wang, "Revisiting pivot language approach for machine translation," in *Proc. of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, August 2009, pp. 154–162.

[14] L. Schwartz, "Multi-source translation methods," in *Proc. of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA)*, October 2008.

[15] A. Eisele, "Parallel corpora and phrase-based statistical machine translation for new language pairs via multiple intermediaries," in *Proc. of the Sixth International Conference on Language Resources and Evaluation (LREC)*, May 2006, pp. 845–848.

[16] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

[17] S. Bangalore, G. Bordel, and G. Riccardi, "Computing consensus translation from multiple machine translation systems," in *Proc. of the 2001 Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2001, pp. 351–354.

[18] E. Matusov, G. Leusch, R. E. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y.-S. Lee, J. B. Marino, M. s. Paulik, S. Roukos, H. Schwenk, and H. Ney, "System combination for machine translation of spoken and written language," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 7, pp. 1222–1237, Sept. 2008.

[19] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, March 2003.

[20] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proc. of the MT Summit X*, September 2005.

[21] J. Marino, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. Fonollosa, and M. Costa-jussà, "N-gram based machine translation," *Computational Linguistics*, vol. 32, no. 4, pp. 527–549, 2006.

[22] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 45, June 2007, p. 2.

[23] F. Casacuberta and E. Vidal, "Machine translation with inferred stochastic finite-state transducers," *Computational Linguistics*, vol. 30, no. 3, pp. 205–225, 2004.

[24] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, July 2002, pp. 295–302.

[25] C. Tillmann, "A unigram orientation model for statistical machine translation," in *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Boston, MA, May 2004, pp. 101–104.

[26] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)*. Morristown, NJ, USA: Association for Computational Linguistics, 1996, pp. 310–318.

[27] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. of the International Conference on Spoken Langage Processing (ICSLP)*, vol. 2, Denver, CO, 2002, pp. 901–904.

[28] P. Koehn, A. Axelrod, A. Birch, C. Callison-Burch, M. Osborne, and D. Talbot, "Edinburgh system description for the 2005 IWSLT speech translation evaluation," in *Proc. of the International Workshop on Spoken Language Translation 2005 (IWSLT)*, Pittsburgh, PA, October 2005.