Advances on Spoken Language Translation in the Quaero Program

¶*Karim Boudahmane,* §*Bianka Buschbeck,* †*Eunah Cho,* ‡*Josep Maria Crego,*
**Markus Freitag,* ‡*Thomas Lavergne,* **Hermann Ney,* †*Jan Niehues,*
**Stephan Peitz,* §*Jean Senellart,* ‡*Artem Sokolov,* †*Alex Waibel,*
§*Tonio Wandmacher,* **Joern Wuebker and* ‡*François Yvon*

¶Direction générale de l'armement (DGA), France (firstname.surname@dga.defense.gouv.fr)
†Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany (firstname.surname@kit.edu)
‡LIMSI-CNRS, Orsay, France (firstname.surname@limsi.fr)
*RWTH Aachen University, Aachen, Germany (surname@cs.rwth-aachen.de)
§SYSTRAN Software, Inc. (surname@systran.fr)

Abstract

The Quaero program is an international project promoting research and industrial innovation on technologies for automatic analysis and classification of multimedia and multilingual documents. Within the program framework, research organizations and industrial partners collaborate to develop prototypes of innovating applications and services for access and usage of multimedia data. One of the topics addressed is the translation of spoken language. Each year, a project-internal evaluation is conducted by DGA to monitor the technological advances. This work describes the design and results of the 2011 evaluation campaign. The participating partners were RWTH, KIT, LIMSI and SYSTRAN. Their approaches are compared on both ASR output and reference transcripts of speech data for the translation between French and German. The results show that the developed techniques further the state of the art and improve translation quality.

1. Introduction

Quaero[1] is a research and development program with the goal of developing multimedia and multilingual indexing and management tools for professional and general public applications. German and French public and private organizations collaborate on research and the realisation of advanced demonstrators and prototypes of innovative applications and services for access and usage of multimedia information, such as spoken language, images, video and music. The program facilitates strong synergies between the participating industrial partners and research organisations. Regular evaluations are conducted to evaluate the market readiness and technological maturity of the research and development results.

One of the topics tackled in the Quaero program is spoken language translation (SLT). In this work, the 2011 project-internal evaluation campaign on SLT is described. The campaign focuses on the language pair German-French in both directions, and both human and automatic transcripts of the spoken text are considered as input data. The automatic transcripts were produced by the Rover combination of single-best output of the best submission from each of the three sites participating in the internal 2010 automatic speech recognition (ASR) evaluation, which is described in an accompanying paper [1]. The campaign was designed and conducted by DGA and compares the different approaches taken by the four participating partners RWTH, KIT, LIMSI and SYSTRAN. In addition to publicly available data, monolingual and bilingual corpora collected in the Quaero program were used for training and evaluating the systems.

The approaches to machine translation taken by the partners differ substantially. KIT, LIMSI and RWTH apply statistical techniques to perform the task, whereas SYSTRAN uses their commercial rule-based translation engine. KIT makes use of a phrase-based decoder augmented with part-of-speech (POS) information and bilingual language models. LIMSI applies the n-gram-based approach and rescores the output with a neuronal language model. RWTH performs system combination on several systems, making use of both the phrase-based and the hierarchical paradigm. All partners adapt the speech data within their preprocessing step, in order to be able to apply their usual text translation techniques.

To visualize the improvements over time, previous year's systems are evaluated as well. The results show that the novel techniques developed by the partners within the scope of the program improve the state of the art and lead to better quality of the automatic translations.

The paper is structured as follows. The evaluation framework is specified in Section 2. In Section 3 we describe each participant's translation system. The results of the campaign are discussed in Section 4 and we conclude in Section 5.

[1] http://www.quaero.org

2. Quaero Evaluation Framework

2.1. Description of the Task

Translation of speech is the process of translating the transcription of spoken language in a text document from one natural language to another. Different kinds of text data inputs can be considered according to their closeness to the initial speech: manual transcriptions, automatic transcriptions and final text editions [2]. In our case, both automatic and manual transcriptions has have been used as sources for the translation. The main objective of this evaluation is to measure the performance of the technology and its readiness for integration in innovative projects. This performance has been measured on both directions between the languages French and German. The systems have been evaluated on a mixture of broadcast news and broadcast conversations transcriptions. For each translation direction, two different conditions were considered. They differ on the type of input material. The evaluated conditions were the following: manual transcriptions segmented with a sentence-based segmentation and output of an ASR system with a segmentation generated automatically.

2.2. Data Description

For the statistical systems, two training data sources were available. The partners were allowed to use the well-known data from the *ACL 2010 Joint Fifth Workshop On Statistical Machine Translation*[2] and data, which was collected within the Quaero program. The domain of the collected data is politics-news and UN documents. Both bilingual and monolingual data were provided for the languages German and French. Table 1 shows the statistics for the amount of data released for training. The data was collected from the following individual sources:

- admin.ch
- project-syndicate.org
- bookshop.europa.eu
- presseurop.eu
- arte.tv

The corpora used to evaluate this task have been built from French and German (manual) transcriptions extracted from the test set used in the previous year's Quaero evaluation campaign of ASR [1]. These transcriptions come from recordings of broadcast shows. The transcriptions were re-segmented manually by the human translators into sentences. Indeed the time-based segmentation, traditionally used for ASR purposes, induced translation issues in the previous evaluation. These issues result from the fact that the time-based segments are independent of the semantic units (e.g. units can be split when breathing) and from the difference of syntax between French and German. ASR outputs have also been automatically segmented and aligned with the human generated transcriptions to make possible the use of the same references with the two kinds of sources. The reference transcriptions were translated twice by human translators and their translations have been used as references for both evaluation conditions: translation of manual transcriptions and translation of ASR output. In the first case, only the MT performances are evaluated as the input is a reference transcription, whereas in the second one, the complete processing chain is evaluated as the translation system has to deal with the errors of the speech recognition system. Table 2 summarizes the statistics of the evaluation corpora. Over the years a development set of around 50K words per translation direction has been built from the test sets of the previous years.

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

Table 1: The corpus statistics of the data collected within the Quaero program (bilingual and monolingual)

	German	French
Documents	16 K	
Running words	5.3M	6.3M
	French	
Documents	250 K	
Running words	70 M	
	German	
Documents	69 K	
Running words	25 M	

Table 2: Corpus statistics for the evaluation data sets used in the 2011 Quaero SLT evaluation campaign.

	German-French	French-German
Documents	7	5
Sentences	971	823
Running words	23K	21K

2.3. Metrics and Scoring

The BLEU-4 score [3] and the Translation Edit Rate (TER) [4] were chosen as the evaluation metrics for machine translation in Quaero program. BLEU measures the closeness of a candidate translation to one or several reference translations by counting the number of n-grams in the system output that also occur in the reference translation. TER is an error measure for machine translation that measures the number of edits required to change a system output into one of the references. TER is defined as the minimum number of

edits needed to change a hypothesis so that it matches one of the references, normalized by the average length of the references. The possible edits are the insertion, deletion and substitution of single words and the shifts of word sequences.

In this evaluation two references were used to compute BLEU and TER scores. Both references were produced independently by professional human translators. Scores were calculated in case and punctuation sensitive fashion.

3. System Descriptions

3.1. KIT

The KIT system for the German to French and French to German SLT translation tasks in the Quaero 2011 evaluation campaign is designed as follows.

To adapt our models to the speech translation tasks, we try to match the text-based training data to the text produced by a speech recognizer. After generating the word alignment, we removed all punctuation marks from the source side of the training corpus and mapped the alignment to the new corpus. Then we continued with building the translation models and reordering models on this corpus with the standard techniques for text translation. For the test data, we applied additional smart-casing for all words. That means on encountering an unknown word we check the phrase table for occurrences of that word in a different casing variant and change the case as required to be able to translate it.

Some of the available data contains a lot of noise. The Giga corpus, for example, includes a large amount of noise such as non-standardized HTML characters. Also, the Bookshop and Presseurop corpora contain truncated lines, which do not match its aligned translation sentence. These noisy pairs potentially degrade the quality of the translation model. The special filtering was applied to the Giga corpus and some of the Quaero data. We used a Support Vector Machines classifier to filter the corpus, inspired by the work of [5] on comparable data.

To generate the translation model, we used the MGIZA++ Toolkit to calculate the word alignment for the training corpus. Afterwards, the alignments were combined using the grow-diag-final-and heuristic. Word reordering is addressed using the POS-based reordering model as described in [6] to account for the different word orders in the languages. To cover long-range reorderings, we apply a modified reordering model with non-continuous rules [7]. The part-of-speech tags for the reordering model are obtained using the TreeTagger [8]. The phrase table and the phrases were built with the Moses Toolkit [9] and scored by our inhouse parallel phrase scorer [10]. We used 4-gram language models with Kneser-Ney smoothing, which are generated by using the SRILM toolkit [11]. The system applied a bilingual language model as described in [12] to extend the context of source language words available for translation. Tuning is performed using minimum error rate training against the BLEU score as described in [13]. Translations are generated using our in-house phrase-based decoder [14].

German-French For German to French we applied long-range POS-based reordering rules and lattice phrase extraction. We added a bilingual language model and a POS-based bilingual language model. The part-of-speeches for this model were generated by using the RF tagger for German [15] and the LIA Tagger for French [3]. These taggers produce more fine-grained linguistic information than the TreeTagger, whose output is used for POS-based reordering.

French-German For French to German we also used long-range POS based reordering rules and lattice phrase extraction. Using the POS-based language model led to a big improvement.

3.2. LIMSI

LIMSI's participation in Quaero 2011 evaluation campaign was focused on the translation of German from and into French. The adaptation of our text translation system to speech inputs is mostly performed in preprocessing, aimed at removing dysfluencies, partially recognized or repeated words, etc. The rest of the pipeline is unchanged as compared to text translations.

For translations between German and French we used N-code[4], our in-house statistical machine translation system based on bilingual n-grams.

N-code overview N-code's translation model implements a stochastic finite-state transducer (FST) trained using an n-gram model (source,target) pairs. The training requires source-side sentence reorderings to match the target word order, also performed by a stochastic FST reordering model, which uses POS information to generalize reordering patterns beyond lexical regularities. Complementary to the translation model, ten more features are used in a linear scoring function: a target-language model; four lexicon models; two lexicalized reordering models [16] to predict the orientation of the next translation unit; a weak distance-based distortion model; and finally a word-bonus model and a tuple-bonus model which compensate for the system preference for short translations. The four lexicon models are similar to the standard ones in phrase-based systems: two scores correspond to the relative frequencies of the tuples and two lexical weights, estimated from the automatically generated word alignments. The weights associated to features are found using the minimum error rate training procedure [17] on the development set. The decoding is beam-search-based on top of a dynamic programming algorithm. Reordering hypotheses are computed in a preprocessing step, making use of reordering rules built from the word reorderings introduced in the tuple extraction process. The resulting reordering hypotheses are passed to the decoder as word lattices [18].

German-French Part-of-speech information for German

[3] http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/bechet/download_fred.html

[4] http://www.limsi.fr/Individu/jmcrego/n-code

is computed using in-house CRF-based tagger [19]. All the available data has been preprocessed and word aligned using MGIZA++; these alignments were then used in a standard N-code pipeline. As development set we used the WMT 2010 newstest set; internal tests were conducted on the test data of 2009 and 2011.

LIMSI used the best available text translation system and the preprocessing with tools initially developed and used for our German to English systems [20]. These tools have also been augmented so as to perform a restricted form of long-range reorderings, notably to move separable particles closer to the verbs they depend on [21]. For the reordering models we selected the monotone-swap-discontinuous (MSD) model.

Language models Large 4-gram language models were trained on all the available data as described in [22]. Additionally, SOUL, a neuronal language model was used to rescore the n-best hypotheses. These models were trained following the methodology of [23] and used for rescoring n-best lists. We used 10-gram history size (differences with 6-gram were insignificant). Using the neural language model led to (small but consistent) improvements in all tasks.

3.3. RWTH

This Section describes the RWTH system for the participation in the Quaero 2011 SLT evaluation campaign on both the German to English and English to German task. The adaptation of our text translation system to speech inputs was mostly performed in preprocessing. We deleted all punctuation from the source language of our training and development data. To give the translation system more stability, we inserted on each source sentence one punctuation mark at the end of the sentence. The rest of the pipeline was unchanged as compared to text translations.

For the Quaero 2011 evaluation RWTH utilized state-of-the-art phrase-based and hierarchical translation systems as well as our in-house system combination framework. GIZA [24] was employed to train word alignments, all language models were created with the SRILM toolkit [11] and are standard 4-gram language models with interpolated modified Kneser-Ney smoothing. The phrase-based statistical machine translation system (pbt) used in this work is an in-house implementation of the state-of-the-art machine translation decoder described in [25]. For our hierarchical setups, we employed the open source translation toolkit Jane [26], which has been developed at RWTH and is freely available for non-commercial use. The basic concept of RWTH's approach to machine translation system combination is described in [27, 28].

With both decoders, we did several setups with different amounts of models. Optional additional models are discriminative word lexicon (dwl) models, triplet lexicon models [29] and additionally binary count features. Unless stated otherwise, we optimized the model weights with standard minimum error rate training [17] on 100-best lists on BLEU. With the help of system combination, we combined the hypotheses of all our different setups.

German-French For German to French we did a system combination of the following five systems:

- Jane with standard features
- Jane with additional 26 binary count features
- pbt with standard features
- pbt with additional model dwl
- pbt with additional model triplets

French-German For French to German we did a system combination of the following seven systems:

- Jane with standard features
- Jane with additional 26 binary count features
- Jane with standard features with BLEU $-$TER as optimization criterion
- Jane with additional model triplets
- pbt with standard features
- pbt with additional models dwl and triplets
- pbt with additional model triplets

With the system combination of all different systems, we got an improvement in BLEU and in TER compared to the best single system of both tasks.

3.4. SYSTRAN

The German and French data submitted by SYSTRAN were obtained by the SYSTRAN baseline engine, being traditionally classified as a rule-based system. However, over the decades, its development has always been driven by pragmatic considerations, progressively integrating many of the most efficient MT approaches and techniques. Some of the analysing modules, like the part-of-speech-tagger for example, make use of decision-tree techniques combining linguistic rules with corpus-extracted knowledge. For this reason, it is difficult to categorize the SYSTRAN engine as simply rule- or statistics-based.

An essential component of the SYSTRAN engine are the manually developed linguistic resources, ranging from 100K to 800K entries for each language pair. The dictionaries contain single- and multiword entries as well as complex, customized disambiguation rules. Translation is basically performed in four steps:

1. *Preprocessing*: Normalisation, segmentation, lookup from idiom, stem and compound dictionaries
2. *Analysis*: part-of-speech analysis, homograph resolution, syntactic dependency parsing

Table 3: Results for the German-French translation of manual transcripts.

System	BLEU [%]	TER [%]
KIT	**29.6**	**53.9**
LIMSI	25.9	56.8
RWTH	25.2	60.5
SYSTRAN	20.4	62.7

Table 4: Results for the French-German translation of manual transcripts.

System	BLEU [%]	TER [%]
KIT	20.6	62.3
LIMSI	**26.8**	**58.1**
RWTH	18.6	64.4
SYSTRAN	18.2	68.7

3. *Transfer*: Application of conditional disambiguation rules (based on morpho-syntactic analysis)

4. *Synthesis*: Morphological regeneration (inflection), syntactic rearrangement

A central guiding principle at SYSTRAN for the development of the translation engine is that the output be deterministic and transparent; it ought to be possible to explain the translation results and - if necessary - to modify the rules involved.

4. Results

The results for all four evaluated conditions are summarized in the Tables 3 through 6. On each condition, the results of all four partners are given in BLEU and TER. The best scores according to each metric are in bold face.

From the results it is clear that, especially for the French target language, the statistical systems have advantages over the rule-based engine employed by SYSTRAN when measured with BLEU and TER. This can be expected, as statistical systems are optimized specifically to perform well on these scores. However, rule-based engines are known to often outperform the statistical approach when it comes to acceptance among human evaluators. For the German-French translation of manual transcripts (cf. Table 3), KIT clearly outperforms the other partners in both measures. For the corresponding French-German task (cf. Table 4) LIMSI has a strong advantage over their competitors. When translating the automatic transcripts, the differences between the three statistical systems are much smaller. For both German-French (cf. Table 5) and French-German (cf. Table 6), KIT reaches the best BLEU score, while RWTH has the best TER. In Table 6, we also included scores from the respective evaluation systems of previous years. This shows how the novel techniques developed in the Quaero program in the past years affect translation quality. The best scoring system of 2011 yields improvements of 2.5% BLEU over the best system of 2009. In TER, the improvement between the respective best systems is 2.0%.

Table 5: Results for the German-French translation of automatic transcripts.

System	BLEU [%]	TER [%]
KIT	**18.4**	70.4
LIMSI	13.4	71.0
RWTH	16.1	**69.7**
SYSTRAN	10.0	76.7

Table 6: Results for the French-German translation of automatic transcripts. Previous evaluation systems are included.

System	BLEU [%]	TER [%]
KIT 2009	16.4	67.5
KIT 2010	17.7	66.1
KIT	**18.9**	68.0
LIMSI	17.0	68.7
RWTH 2009	12.0	70.1
RWTH 2010	17.3	66.7
RWTH	17.6	**65.5**
SYSTRAN	16.0	71.5

The results also illustrate the difficulty of translating ASR output as opposed to clean, human generated text. For German-French, the KIT system degrades by 11.2% BLEU and 16.5% TER when moving from reference to automatic transcription as input. On French-German, the difference is 9.8% BLEU and 10.6% TER for the LIMSI system, which performed best on the manual transcripts.

5. Conclusions

In this work, we described the spoken language translation evaluation 2011 of the Quaero research program. It focuses on the German and French languages. As input for the translation engines, both automatic and human generated transcriptions of the speech data was considered. The four partners KIT, LIMSI, RWTH and SYSTRAN make use of substantially different techniques to perform the task, which were compared and evaluated in this campaign conducted by DGA. Both rule-based and statistical approaches are applied. The basic statistical translation engines make use of three different paradigms: phrase-based, hierarchical and n-gram-based. Additionally, each site incorporated specialized techniques developed within the scope of Quaero, that can improve translation quality.

The results show the higher difficulty of the task of translating automatic transcripts rather than clean text. The improvement over time achieved by the research conducted within the Quaero program is shown on the French-German translation of automatic transcripts. The best system of the 2009 evaluation could be improved by 2.5% BLEU.

6. Acknowledgements

This work was achieved as part of the Quaero program, funded by OSEO, French State agency for innovation.

7. References

[1] L. Lamel, J.-L. Gauvain, I. Oparin, V. B. Le, N. T. Vu, T. Schlippe, T. Schultz, F. Kraft, S. Stüker, H. Ney, R. Schlüter, M. Nußbaum-Thom, M. Sundermeyer, J. Despres, Y. Josse, and B. Vieru, "Speech Recognition for Machine Translation in Quaero," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, Dec. 2011.

[2] D. Mostefa, O. Hamon, N. Moreau, and K. Choukri, "Technological showcase and end-to-end evaluation architecture, tc-star project," ELDA, Tech. Rep. Deliverable D30, May 2007.

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, July 2002, pp. 311–318.

[4] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, August 2006, pp. 223–231.

[5] D. S. Munteanu and D. Marcu, "Improving machine translation performance by exploiting non-parallel corpora," *Computational Linguistics*, vol. 31, pp. 477–504, 2005.

[6] K. Rottmann and S. Vogel, "Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model," in *TMI*, Skövde, Sweden, 2007.

[7] J. Niehues and M. Kolss, "A POS-Based Model for Long-Range Reorderings in SMT," in *Fourth Workshop on Statistical Machine Translation (WMT 2009)*, Athens, Greece, 2009.

[8] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," in *International Conference on New Methods in Language Processing*, Manchester, United Kingdom, 1994.

[9] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of ACL 2007, Demonstration Session*, Prague, Czech Republic, 2007.

[10] T. Herrmann, M. Mediani, J. Niehues, and A. Waibel, "The karlsruhe institute of technology translation systems for the wmt 2011," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, July 2011, pp. 379–385. [Online]. Available: http://www.aclweb.org/anthology/W11-2145

[11] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, vol. 2, Denver, CO, Sept. 2002, pp. 901–904.

[12] J. Niehues, T. Herrmann, S. Vogel, and A. Waibel, "Wider context by using bilingual language models in machine translation," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, July 2011, pp. 198–206. [Online]. Available: http://www.aclweb.org/anthology/W11-2124

[13] A. Venugopal, A. Zollman, and A. Waibel, "Training and Evaluation Error Minimization Rules for Statistical Machine Translation," in *Workshop on Data-drive Machine Translation and Beyond (WPT-05)*, Ann Arbor, Michigan, USA, 2005.

[14] S. Vogel, "SMT Decoder Dissected: Word Reordering." in *International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China, 2003.

[15] H. Schmid and F. Laws, "Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging," in *COLING 2008*, Manchester, Great Britain, 2008.

[16] C. Tillmann, "A unigram orientation model for statistical machine translation," in *Proceedings of HLT-NAACL 2004: Short Papers*, ser. HLT-NAACL-Short '04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004, pp. 101–104.

[17] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 2003, pp. 160–167.

[18] J. M. Crego and J. B. Mariño, "Improving statistical mt by coupling reordering and decoding," *Machine Translation*, vol. 20, pp. 199–215, September 2006.

[19] T. Lavergne, O. Cappé, and F. Yvon, "Practical very large scale CRFs," in *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010, pp. 504–513.

[20] I. Durgar El-Kahlout and F. Yvon, "The pay-offs of preprocessing for German-English Statistical Machine Translation," in *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, M. Federico, I. Lane, M. Paul, and F. Yvon, Eds., 2010, pp. 251–258.

[21] S. Nießen and H. Ney, "Improving smt quality with morpho-syntactic analysis," in *Proceedings of COLING*, Saarbrücken, Germany, 2000, pp. 1081–1085.

[22] A. Allauzen, H. Bonneau-Maynard, H.-S. Le, A. Max, G. Wisniewski, F. Yvon, G. Adda, J. M. Crego, A. Lardilleux, T. Lavergne, and A. Sokolov, "LIMSI@WMT11," in *Proceedings of the 6th Workshop on Statistical Machine Translation (WMT)*, 2011, pp. 309–315.

[23] H. S. Le, I. Oparin, A. Allauzen, J.-L. Gauvain, and F. Yvon, "Structured output layer neural network language model," in *ICASSP*, 2011, pp. 5524–5527.

[24] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, Mar. 2003.

[25] R. Zens and H. Ney, "Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation," in *International Workshop on Spoken Language Translation*, Honolulu, Hawaii, Oct. 2008, pp. 195–205.

[26] D. Stein, D. Vilar, S. Peitz, M. Freitag, M. Huck, and H. Ney, "A guide to jane, an open source hierarchical translation toolkit," *The Prague Bulletin of Mathematical Linguistics*, no. 95, pp. 5–18, Apr. 2011.

[27] E. Matusov, N. Ueffing, and H. Ney, "Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment," in *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2006, pp. 33–40.

[28] E. Matusov, G. Leusch, R. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y.-S. Lee, J. Marino, M. 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, 2008.

[29] A. Mauser, S. Hasan, and H. Ney, "Extending statistical machine translation with discriminative and trigger-based lexicon models," in *Conference on Empirical Methods in Natural Language Processing*, Singapore, Aug. 2009, pp. 210–217.