ENGLISH-LATVIAN SMT: THE CHALLENGE OF TRANSLATING INTO A FREE WORD ORDER LANGUAGE

Maxim Khalilov[1], José A. R. Fonollosa[2], Inguna Skadiņa[3], Edgars Brālītis[3], and Lauma Pretkalniņa[3]*

[1] Institute for Logic, Language and Computation
Universiteit van Amsterdam
Amsterdam, The Netherlands

[2] Centre de Recerca TALP
Universitat Politècnica de Catalunya
Barcelona, Spain

[3] Institute of Mathematics and Computer Science
University of Latvia
Riga, Latvia

ABSTRACT

This paper presents a comparative study of two approaches to statistical machine translation (SMT) and their application to a task of English-to-Latvian translation, which is still an open research line in the field of automatic translation.

We consider a state-of-the-art phrase-based SMT and an alternative N-gram-based SMT systems. The major differences between these two approaches lie in the distinct representations of bilingual units, which are the components of the bilingual model driving translation process and in the statistical modeling of the translation context.

Latvian being a rather free word order language implies additional difficulties to the translation process. We contrast different reordering models and investigate how well they deal with the word ordering issue.

Moving beyond automatic scores of translation quality that are classically presented in MT research papers, we contribute presenting a manual error analysis of MT systems output that helps to shed light on advantages and disadvantages of the SMT systems under consideration and identify the most prominent source of errors typical for both SMT systems.

***Index Terms*—** Natural languages, finite state machines, language processing, statistical machine translation.

1. INTRODUCTION

Translation into languages with relatively free word order has received a lot less attention than translation into fixed word order languages (English), or into analytical languages (Chinese). Free word order languages differ crucially from the languages that follow a restrictive word order scheme, first of all, in the way how the pragmatic information is conveyed. In fixed word order languages (like, German, English, or Spanish) the order of syntactic constituents varies negligibly (or does not vary at all) and the emotional component of the message is usually transmitted through intonation variation[1]. In contrast to them, the free word order languages (like, Latvian, Russian, or Greek) often rely on the order of constituents to convey the topicalization or focus of the sentence.

Latvian language is the target language in the experiments that we report in this paper. There are about 1.5 million native Latvian speakers around the world: 1.38 million are in Latvia, while others are spread in USA, Russia, Sweden, and some other countries. Also Latvian language is second language for about 0.5 million inhabitants of Latvia and several tens of thousands from neighbor countries, especially Lithuania[2].

Latvian is one of two living Baltic languages and it is characterized by rich morphology, relatively complex pre- and postposition structures and high level of morpho-syntactic ambiguity. Despite that it descends from the same ancestor language as Germanic languages, it differs from them significantly and the experience gained from machine translation into German or English can hardly be transferred to the English-to-Latvian translation task.

Nowadays, scientific community is starting to express doubts that the models working pretty well for fixed word order languages are still efficient for free word order languages (for example, construction of an English-to-Czech SMT system taking into consideration very rich morphology

*The bulk of the work presented in this paper was done during the first author's Ph.D studies in Centre de Recerca TALP, Universitat Politècnica de Catalunya, Barcelona (Spain).

[1] There are some exceptions to the general rule, for example, when it is necessary to emphasize the object of the sentence (*I agree* ***with you*** -> ***With you*** *I agree*), or in question sentences.

[2] Source: State Language Agency `http://www.valoda.lv/lv/latviesuval`

and relatively free word order of Czech is one of the goals of the Euromatrix(plus) project[3]). A thorough discussion of the appropriate word ordering strategy (using contextual information) for English-to-Turkish rule-based machine translation can be found in [1]; in [2], the authors concentrate on SMT of indigenous Australian languages (one of the two languages under consideration is a prototypical non-configurational language).

However, translation from Latvian into English and vice versa has not received much attention in the SMT community: the first and only study on Latvian-to-English SMT, to our knowledge, was dated to 2007 [3], that is much later than SMT systems for popular language pairs.

In this paper, we study some aspects of English-to-Latvian MT. First, we compare the outputs of two SMT systems following two different approaches to MT and reporting results in terms of automatic evaluation metrics. We consider a "de facto" standard phrase-based Moses[4] system [4] and an N-gram-based SMT system [5]. We then study two alternative word reordering techniques for each translation system and measure how effective they are translating from English into a non-configurational Latvian language.

The paper concludes with human error analysis performed in order to identify the major strengths and weaknesses of the Moses and N-gram-based SMT systems when translating into Latvian.

The rest of this paper is organized as follows. Section 2 briefly describes phrase- and N-gram-based SMT system configurations, Section 3 outlines the experimental setup, Section 4 details the results of automatic translation quality evaluation, along with the results of human evaluation and error analysis, while Section 5 presents the conclusions drawn from the study.

2. TWO APPROACHES TO SMT

SMT is based on the principle of translating a source sentence ($f_1^J = f_1, f_2, ..., f_J$) into a sentence in the target language ($e_1^I = e_1, e_2, ..., e_I$). The problem is formulated in terms of source and target languages; it is defined according to equation (1) and can be reformulated as selecting a translation with the highest probability from a set of target sentences (2):

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ p(e_1^I \mid f_1^J) \right\} = \quad (1)$$

$$= \arg\max_{e_1^I} \left\{ p(f_1^J \mid e_1^I) \cdot p(e_1^I) \right\} \quad (2)$$

where I and J represent the number of words in the target and source languages, respectively.

Modern state-of-the-art SMT systems operate with the bilingual units extracted from the parallel corpus based on word-to-word alignment. They are enhanced by the *maximum entropy approach* and the posterior probability is calculated as a *log-linear combination* of a set of feature functions [6]. Using this technique, the additional models are combined to determine the translation hypothesis $\hat{e}_1^I$ that maximizes a log-linear combination of these feature models [7], as shown in (3):

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \quad (3)$$

where the feature functions h_m refer to the system models and the set of λ_m refers to the weights corresponding to these models.

There have been a bunch of publications that investigate the source of the possible improvements and degradations in translation quality when using translation systems underlined by different statistical models. For example, in [8], the N-gram-based system is contrasted with a state-of-the-art phrase-based framework, while in [9], the authors seek to estimate the advantages, weakest points, and possible overlap between syntax-augmented MT and N-gram-based SMT. In [10] the comparison of phrase-based, hirearchical, and syntax-based SMT systems is provided.

In this section we discuss the translation models compared in this work.

2.1. Phrase-based SMT

Most of modern state-of-the-art SMT systems follow the phrase-based approach to translation. The basic idea of this approach is to segment the given source word sequence into monolingual phrases, afterwards translate them and compose the target sentence [6].

A phrase-based translation is considered as a three step algorithm: (1) the source sequence of words is segmented in phrases, (2) each phrase is translated into target language using translation table, (3) the target phrases are reordered to be inherent in the target language.

A bilingual phrase (which in the context of SMT do not necessarily coincide with their linguistic analogies) is any aligned pair of m source words and n target words that satisfies two basic constraints: (1) words are consecutive along both sides of the bilingual phrase and (2) no word on either side of the phrase is aligned to a word outside the phrase [11]. The probability of the phrases is estimated by relative frequencies of their appearance in the training corpus.

The system built for the English-to-Latvian translation experiments is implemented within the open-source MOSES toolkit [12]. Standard training and weights tuning procedures which were used to build our system are explained in details on the MOSES web page: `http://www.statmt.org/moses/`. Two word reordering methods are considered: a

[3] `http://www.euromatrix.net/`

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

distance-based distortion model (see 2.1.1) and lexicalized MSD block-oriented model (see 2.1.2).

2.1.1. Distance-based

A simple distance-based reordering model default for Moses system is the first reordering technique under consideration. This model provides the decoder with a cost linear to the distance between words that should be reordered.

2.1.2. MSD

A lexicalized block-oriented data-driven MSD reordering model [13] considers three different orientation types: monotone (M), swap (S), and discontinuous(D). MSD model conditions reordering probabilities on the word context of each phrase pair and considers decoding process a block sequence generation process with the possibility of swapping a pair of word blocks. Notice that in the experiments conducted within the framework of this study a MSD model was used together with a distance-based reordering model.

2.2. N-gram-based SMT system

Alternative approach to SMT is the N-gram-based approach [5], which regards translation as a stochastic process that maximizes the joint probability $p(s,t)$, leading to a decomposition based on bilingual n-grams, typically implemented by means of a Finite-State Transducer [14].

The core part of the system constructed in this way is a translation model (TM), which is based on bilingual units, called tuples, that are extracted from a word alignment according to certain constraints. A bilingual TM actually constitutes an n-gram LM of tuples, which approximates the joint probability between the languages under consideration and can be seen here as a LM, where the language is composed of tuples.

The tuple-based approach is considered monotonous because the model is based on the sequential order of tuples during training. However, for a great number of translation tasks, a certain reordering strategy is required. In the framework of this study we consider two reordering models: a non-deterministic reordering method (see 2.2.2) and a deterministic version of the statistical machine reordering (SMR) algorithm (see 2.2.3).

2.2.1. Additional features

The N-gram translation system implements a log-linear combination of five additional models:

- *an n-gram target LM;*
- *a target LM of Part-of-Speech (POS) tags*;
- *a word penalty model* that is used to compensate for the system's preference for short output sentences;
- *source-to-target and target-to-source lexicon models* as shown in [15]).

2.2.2. Extended word reordering

An extended monotone distortion model based on the automatically learned reordering rules was implemented as described in [16]. Based on the word-to-word alignment, tuples were extracted by an *unfolding* technique. As a result, the tuples were broken into smaller tuples, and these were sequenced in the order of the target words.

The reordering strategy is additionally supported by a 4-gram LM of reordered source POS tags. In training, POS tags are reordered according to the extracted reordering patterns and word-to-word links. The resulting sequence of source POS tags is used to train the n-gram LM.

2.2.3. Statistical machine reordering

A SMR technique is described in details in [17]. Here, reordering is thought as a first-pass translation performed on the source corpus, which converts it into an intermediate representation, in which source-language words are presented in an order that more closely matches that of the target language. A monotone sequence of source words is translated into the reordered sequence using SMT techniques: SMR and SMT are performed using the same modeling tools as N-gram-based systems but using different statistical log-linear models.

Statistical word classes are used to introduce generalization power to the reordering model.

2.2.4. Decoding and optimization

The open-source MARIE[5] decoder was used as a search engine for the translation system. Details can be found in [18]. The decoder implements a beam-search algorithm with pruning capabilities. All the additional feature models were taken into account during the decoding process. Given the development set and references, the log-linear combination of weights was adjusted using a *simplex* optimization method and an n-best re-ranking as described in `http://www.statmt.org/jhuws/`.

3. EXPERIMENTS

3.1. Data

We used JRC Acquis 2.2 parallel corpus [19] of about 270K parallel sentences. Development set contained of 500 sentences randomly extracted from the bilingual corpus, test corpus size was 1,000 lines. Both the datasets were provided

[5] `http://gps-tsc.upc.es/veu/soft/soft/marie/`

with 1 reference translation. Basic statistics of the bilingual corpus can be found in Table 1.

	Latvian	English
Training		
Sentences	269.98 K	269.98 K
Words	5.40 M	6.65 M
Vocabulary	101.25 K	60.47 K
Development		
Sentences	0.50 K	0.50 K
Words	9.90 K	12.36 K
Vocabulary	3.08 K	2.30 K
Test		
Sentences	1.00 K	1.00 K
Words	20.18 K	24.64 K
Vocabulary	4.98 K	3.49 K

Table 1: Basic statistics of the JRC-Acquis corpus.

3.2. Experimental details

Word alignments were estimated with GIZA++ tool[6] assuming 4 iterations of the IBM2 model, 5 HMM model iterations, 4 iterations of the IBM4 model, and 50 statistical word classes (estimated with the mkcls tool[7]).

Phrase-based experiments were conducted following the guidelines provided on the Moses site[4]. We used the 2008 version of Moses decoder. As an alternative to a traditional (unfactored) model (*PB-u*), we considered a factored phrase-based SMT (*PB-f*) that constructed translation/generation models on the basis of the factorized corpus (preface words, POS tags, and lemmas for English and Latvian).

[6] http://code.google.com/p/giza-pp/
[7] http://www.fjoch.com/mkcls.html

A 4-gram target LM with unmodified Kneser-Ney backoff discounting was generated using the SRI Language Modeling Toolkit [20] and was used in all the experiments.

The following MSD reordering system configuration was used: (*msd-bidirectional-fe* configuration).

The SMR experiments were carried out using 50 classes in the reordering step.

4. RESULTS

4.1. System configurations and evaluation

Two SMT systems (*PB-u* - unfactored and *PB-f* - factored) were contrasted considering the set of experiments carried out on the phrase-based system. Within each system configuration we considered two reordering models: a distance-based model alone (as described in 2.1.1) and a distance-based model operating together with a MSD model (see 2.1.2).

N-gram-based SMT system was enhanced with two alternative reordering models: SMR (see 2.2.3) and an extended input graph model (details can be found in 2.2.2).

We considered four evaluation metrics:

- The BLEU score [21] that accounts for evaluation of the translation quality, by measuring the distance between a given translation and the set of reference translations using an n-gram LM (a 4-gram in this study);
- The NIST score [22] which is based on the BLEU score, but weights n-grams in order to provide less informative n-grams with higher weights;
- The WER score [23] which calculates the minimum word-level Levenshtein distance between a translation system output and a reference translation;
- The PER score [24] which is a variation of WER metric, alleviating the effect of a possibly different word order between an acceptable translation hypothesis and reference translation.

System	Reordering	Dev	Test			
			BLEU	NIST	PER	WER
Phrase-based SMT (Moses)						
PB-u	distance	42.38	43.87	78.80	38.34	51.12
	distance + MSD	42.69	43.95	78.91	38.48	50.47
PB-f	distance	42.11	42.96	78.68	38.71	51.75
	distance + MSD	42.40	43.80	78.63	38.63	50.93
N-gram-based SMT (TALP)						
NB	SMR	43.20	44.64	82.03	35.01	47.98
	Input graph	43.52	45.11	82.40	35.05	47.97

Table 2: English-to-Latvian experimental results.

Automatic evaluation was case sensitive and punctuation marks were considered.

4.2. Automatic evaluation

The results of automatic evaluation of translation quality are shown in Table 2. Best scores are placed in cells filled with grey (within phrase-based and N-gram-based experimental sets).

The major conclusion that can be drawn from the results is that the N-gram-based translation model significantly outperforms the phrase-base system for the English-Latvian language pair. The absolute difference in BLEU score of the best ranked *NB* (namely, *NB* with input graph reordering model) and *PB* (namely, *PB-u* with distance-based and MSD reordering models) systems is about 1.15 BLEU points (that accounts for ≈2.6% in a relative scale). This difference is statistically significant for a 95% confidence interval and 1000 resamples [25][8].

Another important observation is that both "distance+MSD" *PB* models (factored and unfactored) are comparable in terms of automatically evaluated accuracy and both outperform their "distance-based only" versions. The difference between *PB-u* and *PB-f* "distance+MSD" systems is not statistically significant. We speculate that a reordering model plays more important role than a translation model factorization when translating into free word order languages.

The *NB* system enhanced with an input graph POS reordering model achieves better MT performance than the SMR version of this system and this difference is statistically significant.

The difference between "distance-based only" and "distance+MSD" versions of the phrase-based SMT systems is not statistically significant in case of the unfactored TM and it is significant in case of the factored model.

According to the PER metric, the introduction of the MSD model does not introduce any significant improvement. At the same time, the performance of the "distance+MSD" configurations expressed in the WER score is about 0.6-0.8 points better[9] than the performance shown by the distance-based reordering models. As a rough approximation, these results can be interpreted as that the MSD model implies an important improvement in word ordering within a sentence and outperforms the distance-based model applied alone.

4.3. Human evaluation and error analysis

Human analysis of translation output allows going beyond automatic scores and, in the general case, provides a comprehensive comparison of multiple translation systems.

[8]Hereafter, statistical significance test is carried out on the BLEU score measured on the test dataset.

[9]For the WER and PER metrics the lower the score, the better the performance of a SMT system.

Two best systems according to automatic scores were chosen from the phrase-based and N-gram-based experiment sets for human evaluation (*PB-u* with distance-based and MSD reordering models, and *NB* with input word graph model). Every non-repetitive test line from the output of these systems was presented to the judge, who was instructed to decide that the two translations were of equal quality, or that one translation was better than the other. The results of the standard systems comparison can be found in Table 3 and demonstrate that the *NB* system outperforms the *PB* one.

	PB-u +distance +MSD	NB +input graph	Equal
Preference	58	193	539

Table 3: Human evaluation results (standard systems).

In addition, we performed error analysis on 100 first sentences from the test data. The analysis of typical errors generated by each system was done following the error classification scheme proposed in [26] by contrasting the systems output with the reference translation. Table 4 presents the comparative statistics of errors generated by the *PB-u* system enhanced with distance-based and MSD reorderings and the *NB* system with input graph reordering model.

Evaluation of the word order correctness for free word order languages is not a trivial task. We considered equally all admissible word order combinations for the Latvian translations. The clumps are marked erroneous only if the word order is not acceptable in Latvian. In this sence, error analysis gives a more complete and fair view of translation quality than automatic scores which just compare a translation output with a reference translation.

The most prominent source of errors generated by the *PB-u* system, in comparison wit the *NB* system, is related to missing words found in the translation output. We explain it by a high analytical inflection of the Balto-Slavic languages that is modeled better by the N-gram-based system since it involves surrounding context not only for phrase reordering, but conditions translation decisions on previous translation decisions.

However, the aforementioned feature of the N-gram-based architecture turns to be a weakness when dealing with local word reordering, that is reflected in the high number of reordering errors produced by the *NB* system. Experimental results show that internal phrase-based reordering enhanced with the distance-based and MSD block-oriented reordering models (viewing translation as a monotone block sequence generation process) outperforms the POS-based word graph reordering model used in N-gram-based experiments (22 local word/phrase order errors coming from the *Pb-u* system vs. 37 errors of this type produced by the *NB* system).

At the same time, long-range word dependencies are modeled by *PB-u* and *NB* with comparable performance. For clar-

Type	Sub-type	PB-u + MSD	NB + input graph
Missing words		**64**	**16**
	Content words	52	10
	Filler words	12	6
Word order		**35**	**58**
	Local word order	11	23
	Local phrase order	11	14
	Long range word order	6	7
	Long range phrase order	7	14
Incorrect words		**128**	**82**
	Wrong lexical choice	25	20
	Incorrect disambiguation	10	4
	Incorrect form	51	46
	Extra words	34	9
	Style	8	2
	Idioms	0	1
Unknown words		**4**	**8**
Punctuation		**20**	**18**
Total		**250**	**182**

Table 4: Error statistics for a 100-line representative test set.

ity's sake, it is important to notice that the English-to-Latvian translation task is not characterized by the high number of long-range reordering dependencies.

Other important sources of errors of the *PB-u* system are extra words embedded into translated sentences (34 for the *PB-u* vs. 9 for the *NB*). We explain it by the key difference in internal representation of translation units between phrase-based and N-gram-based SMT systems.

5. CONCLUSIONS AND FUTURE WORK

In this paper two alternative SMT systems are compared: the standard phrase-based and the N-gram-based SMT systems. Both translation systems include modern reordering models in final configuration. The comparison was created to be as fair as possible, using the same training material and the same tools on the preprocessing, word-to-word alignment, and language modeling steps.

The results shows that the N-gram-based SMT outperforms Moses-based translation system for the English-to-Latvian translation task in terms of automatic scores (the difference is $\approx$1.15 BLEU points) and human "best/worse" evaluation (the output of the N-gram-based system was ranked higher than the one of the phrase-based system in 193 sentences, while the opposite occurred in 58 cases).

Human error analysis clarifies advantages and disadvantages of the systems under consideration and reveals the most important sources of errors for both systems. The phrase-based system suffers from the missing words problem, while, in case of N-gram-based SMT, the most frequent errors are caused by weak word reordering on the local level.

Findings of this study, along with the robust error analysis of the SMT system outputs can be a very important step on the way of the translation quality improvement when dealing with free word order languages.

A study on introducing of a feature intending to reflect a free word order scheme of the Latvian language is an interesting research topic to be done in the future. Another appealing research topic can be to modify the standard evaluation metrics used for the automatic assessment of translation quality such that they could consider multiple addmisible word permutations within a sentence to express the same message typical for the non-configurational languages.

6. ACKNOWLEDGEMENTS

Work partially supported by the Spanish Ministerio de Educación y Ciencia (TIN2006-12767), by the Spanish Government under grant TEC2006-13964-C03 (AVIVAVOZ project), and by the Latvian Council of Science (project "Application of Factorized methods in English-Latvian Statistical Machine Translation System"). The authors want to thank Khalil Sima'an (Universiteit van Amsterdam) for his valuable discussions and suggestions.

7. REFERENCES

[1] B. Hoffman, "Translating into free word order languages," in *Proceedings of COLING'96*, Copenhagen, Denmark, August 1996, pp. 556–561.

[2] S. Zwarts and M. Dras, "Statistical machine translation of australian aboriginal languages: Morphological analysis with languages of differing morphological richness," in *Proceedings of the Australasian Language Technology Workshop*, Melbourne, Australia, December 2007, pp. 134–142.

[3] I. Skadiņa and E. Brālītis, "Experimental statistical machine translation system for Latvian," in *Proceedings of the 3rd Baltic Conference on HLT*, 2008, pp. 281–286.

[4] Ph. 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*, 2007, pp. 177–180.

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

[6] F. Och and H. Ney, "Discriminative Training and Maximum Entropy Models for Statistical Machine Translation," in *Proceedings of ACL 2002*, 2002, pp. 295–302.

[7] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J.D. Lafferty, R. Mercer, and P.S. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[8] J. M. Crego, M. R. Costa-jussà, J. B. Mariño, and J. A. R. Fonollosa, "Ngram-based versus phrase-based statistical machine translation," in *Proceedings of the 2nd Int. Workshop on Spoken Language Translation (IWSLT'05)*, 2005, pp. 177–184.

[9] M. Khalilov M. and J. A. R. Fonollosa, "N-gram-based statistical machine translation versus syntax augmented machine translation: comparison and system combination," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*, Athens, Greece, April 2009, pp. 424–432.

[10] A. Zollmann, A. Venugopal, F. Och, and J. Ponte, "A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT," in *Proceedings of Coling 2008*, Manchester, August 2008, pp. 1145–1152.

[11] F. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol. 3, no. 4, pp. 417–449, December 2004.

[12] Ph. 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*, 2007, pp. 177–180.

[13] C. Tillman, "A unigram orientation model for statistical machine translation," in *Proceedings of HLT-NAACL'04*, 2004.

[14] F. Casacuberta, E. Vidal, and J. M. Vilar, "Architectures for speech-to-speech translation using finite-state models," in *Proceedings of the Workshop on Speech-to-Speech Translation: Algorithms and Systems*, 2002, pp. 39–44.

[15] F. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, 2004.

[16] J. M. Crego and J. B. Mariño, "Improving statistical MT by coupling reordering and decoding," *Machine Translation*, vol. 20, no. 3, pp. 199–215, 2006.

[17] M. R. Costa-jussà and J. A. R. Fonollosa, "Statistical machine reordering," in *Proceedings of the HLT/EMNLP 2006*, Sydney, Australia, 2006, pp. 70–76.

[18] J. M. Crego, J. B. Mariño, and A. de Gispert, "An ngram-based statistical machine translation decoder," in *Proceedings of INTERSPEECH05*, 2005.

[19] S. Ralf, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga, "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages," in *Proceedings of LREC'2006*, Genoa, Italy, May 2006.

[20] A. Stolcke, "SRILM: an extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, 2002, pp. 901–904.

[21] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of ACL 2002*, 2002, pp. 311–318.

[22] G. Doddington, "Automatic evaluation of machine translation quality using n-grams co-occurrence statistics," in *Proceedings of HLT 2002*, 2002, pp. 128–132.

[23] I. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard, "On the use of information retrieval measures for speech recognition evaluation," IDIAP-RR 73, IDIAP, Martigny, Switzerland, 2004.

[24] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf, "Accelerated DP based search for statistical translation," in *Proceedings of EUROSPEECH 1997*, Rhodes, Greece, September 1997, pp. 2667–2670.

[25] Ph. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of EMNLP 2004*, 2004, pp. 388–395.

[26] D. Vilar, J. Xu, L. F. D'Haro, and H. Ney, "Error Analysis of Machine Translation Output," in *Proceedings of LREC 2006*, 2006, pp. 697–702.