DFKI's SC and MT Submissions to IWSLT 2011

David Vilar, Eleftherios Avramidis, Maja Popović and Sabine Hunsicker

German Research Center for Artificial Intelligence (DFKI GmbH)
Language Technology Lab
Berlin & Saarbrücken, Germany
{name.surname}@dfki.de

Abstract

We describe DFKI's submission to the System Combination and Machine Translation tracks of the 2011 IWSLT Evaluation Campaign. We focus on a sentence selection mechanism which chooses the (hopefully) best sentence among a set of candidates. The rationale behind it is to take advantage of the strengths of each system, especially given an heterogeneous dataset like the one in this evaluation campaign, composed of TED Talks of very different topics. We focus on using features that correlate well with human judgement and, while our primary system still focus on optimizing the BLEU score on the development set, our goal is to move towards optimizing directly the correlation with human judgement. This kind of system is still under development and was used as a secondary submission.

1. Introduction

In this paper we describe DFKI's submission to the System Combination and Machine Translation tracks of this years' IWSLT Evaluation [1]. The task consists of TED Talks given by several speakers on varied topics, expectedly with different talking styles. As such, the data is quite heterogeneous and the perspective of combining systems is specially attractive, trying to take advantage of the strengths of each one.

We focused on a sentence selection mechanism based on exploiting features that are expected to correlate well with human judgement. For our MT submission we used this method to boost the performance of several baseline systems trained on different subsets of the available data. We concentrated on the English-to-French translation direction, and, if not noted otherwise, all the results reported on this paper refer to this translation direction.

2. Baseline Systems

An overview of the statistics of the different available training corpora is given in Table 1. The corpora have been tokenized and the words at the beginning of sentences have been converted to the most frequent case. The variability in the available data can already be seen in these statistics. In order to get the perplexity scores, we trained a 4-gram language model with Kneser-Ney smoothing [2] using the SRI toolkit [3] on each corpus and then measured the perplexity on the test2010 corpus. In spite of its small size, the low perplexity of the TED Talk corpus seems to indicate that it is the better suited for this task. This is not a reliable measure, of course, but it can give an early indication of the similarity of the corpora.

Corpus	Sent.	Running	Vocab.	ppl
TED Talk	107K	2.2M	57K	115.1
Europarl	18M	56M	149K	136.6
News	115K	3.4M	63K	159.6
Giga FrEn	22M	800M	3.1M	153.2
UN	12M	402M	682K	207.7

Table 1: Corpus statistics (Sentences, Running Words, Vocabulary and Perplexity) for the English-to-French translation task, after postprocessing. The perplexity figure is computed on the test 2010 french set, using a 4-gram language model trained on the corresponding corpus.

As a starting point and in order to create different MT systems to combine, we trained two freely available machine translation systems on some of the available bilingual corpora for the evaluation (driven partly by the running time needed to train a full system from scratch).

As phrase-based system we used Moses [4], the current standard toolkit for phrase-based translation. We trained the system with a standard setup, using the dev2010 corpus as development set for minimum error rate training.

As a hierarchical phrase-based system we used the Jane toolkit [5], freely available for non-commercial use. The alignments were taken over from the corresponding Moses systems and again a fairly standard setup was used, optimizing on the dev2010 corpus.

Both systems used the same language model: a 4-gram language model trained on the monolingual TED data.

The results[1] of the different baseline setups are summarized in Table 2. It can be seen that the choice of training

[1]The BLEU scores are cases-sensitive and computed used preprocessed references, in the same way as the preprocessing of the original data. As such it may not fully agree with officially calculated results.

System	Training Corpus	BLEU [%]
Moses	TED	29.9
	News	26.0
	Europarl	27.6
	TED+News	30.3
	TED+Europarl	30.0
Jane	TED	29.6
	WMT11News	23.7
	Europarl	24.8
	TED+News	29.1
	TED+Europarl	28.3

Table 2: Results of the baseline systems.

corpus has a critical effect on the performance of the system, and adding more (out-of-domain) data does not necessarily improve the translation quality (particularly for the Jane system). Overall Moses performs better than Jane on this task, which was to be expected as this language direction does not benefit from the additional modelling possibilities of the hierarchical model.

One of the first ideas we considered when planning the evaluation was to see if the performance of the systems *without* the TED training corpus was on-par or near the system trained with this data. If this was the case, the TED corpus could then be used e.g. to tune document-specific models or do some other kind of domain adaptation. However, as the perplexities in Table 1 hinted and the results of Table 2 show, the TED corpus is crucial for obtaining good performance, and this simple approach was thus not practicable.

3. System Combination Through System Selection

The next step in the development of our system was to combine the different systems and try to improve the overall translation quality.

3.1. First Approach: Document Level Combination

Given the nature of the corpus, i.e. the fact that it consists of different talks by diverse speakers on unrelated topics, we first tried a combination on the document level. The goal was to select for each document the system which provides the best translation, using some automatic method akin to text classification. However this strategy showed not to be effective.

Table 3 shows an overview of the BLEU scores of each of the documents composing the test2010 corpus, with the cells shaded to provide a more visual overview of the distribution of the scores. It can be observed that the best system is generally the same for most documents, and the difference in BLEU scores is not very large. Indeed, if we generate a new complete hypothesis by selecting the best system for each document, we hardly get an improvement over the best single system, obtaining a BLEU score of only 30.7%. Note that this is an *oracle* score[2] obtained using the references to guide the selection.

System	BLEU [%]
Worst System	23.7
Best System	30.3
Document Selection	30.7
Sentence Selection	37.1

Table 4: Overview of oracle scores for the system selection approaches.

3.2. Second Approach: Sentence Level Combination

We decided to move towards a finer granularity and investigate the possibilities of combination on the sentence level. In order to get an idea of the possibility for improvement, we performed another oracle experiment, constructing a new hypothesis using the given reference as a guide. We followed a greedy method to construct the hypothesis, in which for every source sentence, we chose the translation candidate with better accumulated BLEU score up to this point, i.e. without taking the still-to-be-visited part of the corpus into account. This method is not guaranteed to find the best possible hypothesis, but it gives us a hint of the performance that is achievable in principle. In this case the improvement is significant, reaching 37.1% BLEU. We therefore decided to continue with this line of work. An overview of these scores can be found in Table 4.

4. Features

In this section we will list the features we compute for each of the systems. How we take advantage of them will be described in Section 5. We have used features that try to focus on characteristics that humans may use to evaluate a system.

4.1. Cross System BLEU

BLEU was introduced in [7] and has shown to have a high correlation with human judgement. In spite of its shortcomings [8], it has been considered the standard automatic measure in the development of SMT systems (with new measures being added upon, but not substituting it).

Of course, the main problem of using the BLEU score as a feature for sentence selection in a real-life scenario is that we do not have the references available. We overcame this issue by generating a custom set of references for each system, using the other systems as gold translations. This

[2]There is evidence [6], that this method does not necessarily produce the best complete hypothesis, but it should be a good enough indicator for our purposes and further discussion (see also 3.2).

Systems

Documents

20.4	18.9	19.3	20.8	21.3	20.6	17.4	17.3	20.2	20.1
34.1	30.1	31.9	35.5	33.8	33.7	27.8	28.9	32.9	32.2
32.1	25.6	26.2	32.1	31.0	30.6	23.0	25.1	31.1	27.9
26.3	23.8	24.3	25.9	26.2	25.5	20.7	21.9	25.4	25.1
29.9	27.1	28.3	30.2	30.8	29.0	25.1	26.0	28.8	28.1
29.9	25.1	28.1	31.0	30.7	29.0	21.2	22.9	27.7	28.2
33.8	27.9	31.6	34.3	33.5	33.9	26.1	26.7	33.9	31.9
34.3	28.0	29.2	34.0	32.2	33.1	25.2	26.7	31.2	31.0
27.4	22.8	24.0	28.4	27.4	27.0	21.6	22.1	27.0	26.5
33.8	32.2	33.9	34.4	33.7	35.0	29.0	31.6	33.6	33.2
26.8	23.8	25.4	25.0	26.8	27.9	20.3	23.6	26.9	26.8

Table 3: Document level scores. The ordering of the systems corresponds to Table 2, the ordering of the documents to the appearance in the test2010 corpus. The shading of the cells has been normalized for each row (i.e. for each document).

is of course inexact, but n-grams that appear on the output of different systems can be expected to be more probable to be correct, and BLEU calculated this way gives us a measure of this agreement. This approach can be considered related to n-gram posteriors [9] or minimum Bayes risk decoding (e.g. [10]) in the context of n-best rescoring, but applied without prior weighting (unavailable directly) and more focused on the evaluation interpretation.

We generated two features based on this idea. The first one is computed at the system level, i.e. it is the same for each sentence produced by a system and serves as a kind of prior weight similar to the one used in other system combination methods (e.g. [11]). The other feature was computed at the sentence level. For this we used the smoothed version of BLEU proposed in [12], again using the output of the rest of the systems as pseudo-reference. As optimization on BLEU often tends to generate short translations, we also include a word penalty feature.

As an additional experiment, we generated a combination hypothesis by a greedy process, similar to the oracle setting described in Section 3.2. In this case, however, we construct both a new hypothesis and the corresponding references at the same time; the translations that have not been chosen as part of the hypothesis enter the reference. Using this simple method alone we achieve a BLEU score of 30.5% (measured using the true reference). This is a modest improvement (and certainly not statistically significant), but consistent among several experiments. We believe that this can be a starting point for further development.

4.2. Error Analysis Features

It is safe to assume that a human judge will try to choose those translations which contain the least amount of errors, both in terms of content and grammaticality. A classification of errors for machine translation systems has been proposed in [13], and [14] presents how to compute a subset of these error categories automatically. The basic idea is to extend the familiar Word Error Rate (WER) and Position independent word Error Rate (PER) measures on word and base-form[3] levels to identify the different kind of errors. For our system we included following features:

Extra Word Errors (EXTer) Extra words in the hypothesis not present in the references.

Inflection Errors (hINFer) Words with wrong inflection. Computed comparing word-level errors and base-form-level errors.

Lexical Errors (hLEXer) Wrong lexical choices in the hypothesis with respect to the references.

Reordering Errors (hRer) Wrong word order in the hypothesis.

Missing Words (MISer) Words present in the reference that are missing in the hypothesis.

All these features are computed using the open source Hjerson[4] tool [15], which also outputs the standard WER metric, which we added as an additional feature.

As was the case in Section 4.1, for computing these measures we do not have a reference available, and thus we use the rest of the systems as pseudo-references. This has the interesting effect that some "errors" are actually beneficial for the performance of the system. For example, it is known that systems optimised on the BLEU metric tend to produce short

[3] Computed using the TreeTagger tool (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/)

[4] The abbreviations for the features are taken over directly from the output of the tool.

hypotheses. In this sense, the extra words considered as errors by the EXTer measure may be actually beneficial for the overall performance of the system (see also the discussion in Section 5).

4.3. Parsing Features

A basic expectation concerning MT quality is that its output should be grammatical. For this purpose, the use of language models has been dominant when building systems, optimizing the output for the highest probability of the consequent n-grams. On the other hand, automatic metrics are also based on n-grams matching with the reference translation. In order to avoid a possible overfitting on n-grams, but also to capture more complex phenomena (such as long distance structures and grammatical fluency) that are still important to quality output and may have been neglected by the statistical systems, we considered including features derived after parsing the systems' output with Probabilistic Context Free Grammars (PCFG). For this the Berkeley Parser [16] was used.

PCFG parsing allows the generation of n-best lists of trees, scored probabilistically, leading to the selection of the tree with the highest score. From this process, we extracted the number of distinct parsing trees of the sentence, after having allowed the generation of an n-best list of size $n = 1000$. A smaller number of trees could mean that there are less possible tree derivations, i.e. less parsing ambiguity.

Parsing statistics are not only an indicator of the grammaticality of the sentence, but also of how complicated it is, assuming it is fully grammatical. Therefore, we relied on the assumption of isomorphism; i.e. the complexity of the parse of the translated sentence (if fully grammatical) would be proportional to the complexity of the parse of the source sentence (which is expected to be fully grammatical). For this reason, we parsed both the source input and the system outputs and computed the source to target ratio for such scores.

As an additional feature, we included counts and source to target ratios of verb phrases, given the same isomorphism assumption and the fact that a possible "loss" of verb (not explicitly handled by a language model) would radically decrease sentence quality. Further parsing features are subject of future work.

4.4. IBM1 Scores

IBM1-like scores on the sentence level are known to perform well for the rescoring of n-best lists from a single system (see e.g. [17]). Additionally, they have been shown in [18] to correlate well with human judgement for evaluation purposes. We thus include them as additional features.

4.5. Additional Language Models

For the translation systems we trained ourselves we only included one language model trained on the TED data, as initial experimentation with other language models did not seem to bring clear improvements. We do not know what language models were used for the systems in the system combination shared task. We used all the available monolingual data to build additional language models and compute the corresponding scores.

5. Sentence Selection Mechanism

Two sentence selection mechanisms were tried out for this evaluation. Although our goal is to shift to a selection mechanism geared towards human evaluation, using the data made available in the WMT evaluations [19], this approach is still experimental and in development stage. Therefore we also built a more traditional system based on log-linear models trained on the BLEU score.

5.1. Based on BLEU

Log-linear models are at the heart of most state-of-the-art statistical machine translation systems. They model the translation probability of target sentence e_1^I given source sentence f_1^J directly using the expression

$$p(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(f_1^J, e_1^I)\right)}{\sum_{\tilde{e}_1^I} \exp\left(\sum_{m=1}^M \lambda_m h_m(f_1^J, \tilde{e}_1^I)\right)}, \qquad (1)$$

where the h_m are feature functions as the ones described in Section 4 and the λ_m are the corresponding scaling factors, which we optimize with standard MERT training with respect to the BLEU score. This is also the usual approach used for rescoring n-best lists generated by a single system, and has been used previously for sentence selection purposes (see [20] which uses a very similar approach to our own).

Note that no system dependent features like translation probabilities were computed, as we wanted to keep the system general. In fact, for the system combination task, only the single-best translation was provided, without additional information.

Table 5 gives an overview of the effect of the different features used in this approach.[5] It can be seen that the best performance is obtained when combining all the models. Language model scores alone are not powerful enough to give an improvement over the best single system, and the IBM1 scores even hurt performance (note however that this is a single score). They are however important for the combination with the other models in order to obtain the biggest improvement, as the last rows in Table 5 show.

It is also interesting to analyze the sign of the scaling factors corresponding to the different features, which is shown in Table 6. We explicitly do *not* include the magnitude of the factors because of the different scaling of the feature functions may lead to misinterpretations of the importance of each feature. To be able to interpret this table correctly,

[5] Regretfully, due to practical issues during the evaluation preparation we did not use the same set of features for both approaches.

Model	#Features	BLEU [%]
Worst System	–	23.7
Best System	–	30.3
BLEU	3	31.2
IBM1	1	27.2
Error analysis	6	30.5
LMs	5	30.2
All wo. BLEU	12	31.3
All wo. IBM1	14	31.2
All wo. error analysis	9	30.6
All wo. LMs	10	30.7
All	15	31.5

Table 5: Effect of the different feature models on the test2010 corpus.

Feature	Sign
Cross system BLEU (system level)	−
Cross system BLEU (sentence level)	−
Word penalty	−
EXTer	−
hINFer	−
hLEXer	+
hRer	+
MISer	+
WER	+
IBM1	−
LM (Europarl)	−
LM (Giga FrEn)	−
LM (monolingual TED)	+
LM (UN)	+
LM (News)	+

Table 6: Sign of the scaling factors corresponding to the features.

we must take into account that the system minimizes costs (proportional to negative log-probabilities).

The features associated with the BLEU score all get a negative scaling factor, as such higher values are good for the performance of the system, as expected. This also includes the word penalty, i.e. longer sentences are favored.

The error analysis features mostly get positive scaling factors, i.e. the system tries to minimize the number of "errors" of the system (with pseudo-references, see 4.2). One exception are the extra word errors, which, as pointed out before, may help to overcome the tendency towards shorter sentences. Another exception (somehow surprising) are the inflection errors. We do not have a clear explanation for this effect, but it could be that due to the similarity in the construction of the systems, they tend to do the same kind of inflection errors. It can then happen that when one system produces a new inflection it has a higher probability of being right, although this is mere speculation.

The IBM1 score gets a negative scaling factor. As we are working with log-probabilities in this case (not negative!), this indicates that higher IBM1 scores are beneficial for the system. For the language models, the picture is mixed, the system tries to maximize the probability of some of them, but to minimize it for others.

5.2. Based on human ranking

We considered employing a supervised machine learning approach trained over sentences evaluated by human judges, made available by the WMT evaluations. The system was trained based on human rankings of MT output and consequently used to replicate ranking for our sentence-level translation alternatives.

According to this approach [21], ranking is decomposed into a set of pairwise decisions, where each translation output gets compared with each one of the other alternatives. For this purpose, a binary classifier is trained to learn to compare the output quality. The sentence that wins most of the pairwise comparisons (ranked first) is selected for the system combination output.

A Naïve Bayes classifier was trained by estimating $p(C)$ out of relative frequencies of the pairwise examples[6] as following:

$$p(C, F_1, \ldots, F_n) = p(C) \prod_{n}^{i=1} p(F_i|C) \qquad (2)$$

where C is the binary class value and F_i the features. We experimented with several combinations of features, including language model probabilities, parsing features, IBM1 scores and a count of unknown words. Since our features had continuous values (mainly probabilistic scores) $p(F_i|C)$ in Bayes' training was estimated using locally weighed linear regression [24]).

The best results of the various experiments with different feature combinations are shown in Table 7. We report the segment-level correlation coefficient Kendall's tau (τ) [25] and the accuracy of the classifier succesfully selecting the best ranked output[7].

Although these numbers are helpful for selecting the optimal feature set, they already show a rather limited performance of the method. A first issue, which was considered at the end of the development cycle, was that contradictory judgements had been included in the human evaluation set. We gathered all the sentences that had more than one judgment and applied majority voting, in order to get only one

[6]Human judgments for training were obtained from the data of the Shared Evaluation Tasks for WMT08, 09 and 11 [22, 23, 19]

[7]Human judgments for testing and development were obtained from the WMT10 Shared Evaluation Task [26]

Features	τ	acc[%]
basic, VPs, trees, $\text{trees}_{\text{ratio}}$, ibm1	0.077	37.5
basic, VPs, trees, $\text{trees}_{\text{ratio}}$	0.057	35.1
basic, $\text{VPs}_{\text{ratio}}$, trees, $\text{trees}_{\text{ratio}}$	0.049	34.4
basic, $\text{VPs}_{\text{ratio}}$, trees, $\text{trees}_{\text{ratio}}$, ibm1	0.056	33.8
basic, VPs, trees, $\text{trees}_{\text{ratio}}$, ibm1, maj.v	0.078	38.6

τ	segment-level Kendall tau coefficient.
acc	"select-best" accuracy
basic	count of unknown words, language model score, ratio of source to target sentence length.
trees	count of n-best PCFG trees up to 1000

Table 7: Overview classification/select-best performance by using the Bayes classifier

System	% Best	BLEU [%]
BLEU based	43	31.5
Human rank based	26	26.4

Table 8: Comparison of the two approaches for sentence selection

rank per sentence. This lead to the correlation results denoted with "+maj.v." in Table 7.

Another drawback of the pairwise idea occurs when two or more systems collect the same winnings of comparisons, so they are all ranked best. Forceful selection of one of them (e.g. the first appearing in the list), lead to a decrease of about 10% on accuracy. Further research (which may also be related to more informative features) to overcome this problem of uncertainty would have some potential. Additionally, the fact that this selection mechanism is trained on human judgments of news-data, but applied on spoken language topics, creates an out-of-domain situation which may be unpredictable.

5.3. Comparison of Both Approaches

Table 8 shows a short comparison of both approaches for sentence selection. As a comparison based only on the BLEU score would not be fair, as the one based on human ranking is not optimized on this measure, we also performed a binary system comparison as described in [13] on 100 sentences[8]. On 43 sentences the BLEU based system was judged better than the other system, for 26 the contrary was true (for the rest no significant difference was found between the systems). Given this fact and the corresponding BLEU scores we decided to submit the BLEU based system as our primary submission.

[8]The human judge was not a French native speaker.

System	BLEU [%]
Worst System	25.1
Best System	33.5
BLEU Based combination	34.4
Ranks Based combination	29.5

Table 9: Automatic scores on the test2011 corpus.

6. Results

In this section we will analyze the official evaluation results. Note that, contrary to the previous reported scores where we used our internal preprocessing of the references, the scores reported here are calculated in a standardized way, either by the organizers or by the evaluation server made available. Thus they are directly comparable to the results obtained by other groups.

6.1. MT Track

Table 9 includes the results of our submissions on the test2011 corpus. As can be seen, the BLEU based sentence selection mechanism is able to achieve nearly 1% BLEU improvement over the best single system. The rank-based selection again does not obtain good results in terms of this automatic measure, but it should be stressed that it is not optimized with respect to it.

6.2. SC Track

For the system combination track, some additional practical issues have to be commented on. The organizers asked the participants in the translation tracks to provide translation on the development corpora in order to have data to train the SC systems on. However, and probably due to the relatively long time before the final submission and the deadline for providing development data, few groups submitted the translated development sets. As an example, for the English-to-French translation direction, only three groups submitted translations of the development data, while eight groups participated in the final MT track evaluation. We ourselves are the first to blame, as we did not submit any translation of the development data due to not having it ready at the time of the deadline. This has a negative effect at the time of optimizing the systems, as the conditions for the development data do not match the conditions for the test data. There is also the issue that the version of the systems used for the development data is probably not so up-to-date like the final version, but this is expected to have a less critical effect.

Taking this into consideration we optimized the BLEU-based system on the provided development data but also submitted a contrastive system with the scaling factors optimized on our own set of systems. It turns out that this last system outperformed the primary submission. On the devel-

System	BLEU [%]
Worst System	34.4
Best System	37.7
BLEU Based combination	37.5
BLEU Based combination (contrastive)	37.9
Rank Based combi	36.0

Table 10: Automatic scores on the test2011 corpus (SC track).

System	BLEU [%]
Worst System	19.6
Best System	26.3
BLEU Based combi	23.6

Table 11: Automatic scores on the test2011 corpus (SC track, Arabic-to-English direction).

opment data the newly optimized system performed better, but that was probably an overfitting effect, as for the test2011 data it was not able to outperform the best performing single system. The automatic scores are shown on Table 10 for the English-to-French direction and on Table 11 for the Arabic-to-English translation direction, where we also participated, although without success. For the Chinese-to-English translation direction we were not able to obtain improvements even on the development data.

7. Conclusions

We have described DFKI's submission to the IWSLT 2011 evaluation campaign. Our main focus has been put on the development of a sentence selection mechanism which aims to take advantage of features designed to correlate well with human judgements, like error analysis of the translations or BLEU scores. We use the output of the systems to generate pseudo-references for those features that need them. The system combination is able take advantage of the different strengths of each system and achieves nearly 1% improvement in BLEU, boosting the performance of baseline systems and bringing them on-par with most submissions by other participants.

In spite of the good results obtained with sentence selection applied for the internally trained systems, the good performance did not carry over to the combination of systems trained by other participants, where we only obtained a very modest gain. We feel this was mainly due to a mismatch between the development and the testing conditions.

We have also initiated the shift towards a selection mechanism guided by human judgement instead of purely automatic measures. Although the system does not yet achieve state-of-the-art performance, we feel that this is the way to go in further development of machine translation systems, specially if we want to improve more strict (or human driven) quality measures.

8. Acknowledgements

This work was done with the support of the TaraXŰ Project[9], financed by TSB Technologiestiftung Berlin-Zukunftsfonds Berlin, co-financed by the European Union-European fund for regional development.

9. References

[1] M. Federico, L. Bentivogli, M. Paul, and S. Stueker, "Overview of the IWSLT 2011 Evaluation Campaign," in *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, USA, dec 2011.

[2] R. Kneser and H. Ney, "Improved backing-off for M-gram language modeling," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processingw*, vol. 1, May 1995, pp. 181–184.

[3] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proc. of the Seventh International Conference on Spoken Language Processing*. ISCA, Sep. 2002, pp. 901–904.

[4] 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 *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, Jun. 2007, pp. 177–180.

[5] D. Vilar, D. Stein, M. Huck, and H. Ney, "Jane: Open Source Hierarchical Translation, Extended with Reordering and Lexicon Models," in *Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 262–270.

[6] D. Chiang, S. DeNeefe, Y. S. Chan, and H. T. Ng, "Decomposability of translation metrics for improved evaluation and efficient algorithms," in *Proc. of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, October 2008, pp. 610–619.

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

[9] http://taraxu.dfki.de

[8] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the Role of Bleu in Machine Translation Research," in *Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, Apr. 2006, pp. 249–256.

[9] R. Zens and H. Ney, "N-gram Posterior Probabilities for Statistical Machine Translation," in *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), Workshop on Statistical Machine Translation*, New York City, Jun. 2006, pp. 72–77.

[10] N. Ehling, R. Zens, and H. Ney, "Minimum Bayes risk decoding for BLEU," in *Annual Meeting of the Assoc. for Computational Linguistics*, Prague, Czech Republic, Jun. 2007, pp. 101–104.

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

[12] C.-Y. Lin and F. J. Och, "ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation," in *Proc. of the 20th international conference on Computational Linguistics*, ser. COLING '04, Geneva, Switzerland, 2004.

[13] D. Vilar, J. Xu, L. F. D'Haro, and H. Ney, "Error Analysis of Machine Translation Output," in *International Conference on Language Resources and Evaluation*, Genoa, Italy, may 2006, pp. 697–702.

[14] M. Popović and H. Ney, "Towards Automatic Error Analysis of Machine Translation Output," *Computational Linguistics*, vol. 37, no. 4, Dec 2011, to appear.

[15] M. Popović, "Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output," *The Prague Bulletin of Mathematical Linguistics*, pp. 59–68, 2011.

[16] S. Petrov, L. Barrett, R. Thibaux, and D. Klein, "Learning accurate, compact, and interpretable tree annotation," in *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia: Association for Computational Linguistics, July 2006, pp. 433–440.

[17] S. Hasan, R. Zens, and H. Ney, "Are very large N-best lists useful for SMT?" in *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting*. Rochester, NY: Association for Computational Linguistics, Apr. 2007, pp. 57–60.

[18] M. Popovic, D. Vilar, E. Avramidis, and A. Burchardt, "Evaluation without references: Ibm1 scores as evaluation metrics," in *Proc. of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 7 2011, pp. 99–103.

[19] C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan, "Findings of the 2011 workshop on statistical machine translation," in *Proc. of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, July 2011, pp. 22–64.

[20] A. Hildebrand and S. Vogel, "Combination of machine translation systems via hypothesis selection from combined n-best lists," in *MT at work: Proc. of the Eighth Conference of the Association for Machine Translation in the Americas*. Citeseer, 2008, pp. 254–261.

[21] E. Avramidis, M. Popović, D. Vilar, and A. Burchardt, "Evaluate with Confidence Estimation: Machine ranking of translation outputs using grammatical features," in *Proc. of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, July 2011, pp. 65–70.

[22] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "Further meta-evaluation of machine translation," in *Proc. of the Third Workshop on Statistical Machine Translation*. Columbus, Ohio: Association for Computational Linguistics, June 2008, pp. 70–106.

[23] C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder, "Findings of the 2009 Workshop on Statistical Machine Translation," in *Proc. of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece: Association for Computational Linguistics, March 2009, pp. 1–28.

[24] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American statistical association*, vol. 74, no. 368, pp. 829–836, 1979.

[25] M. G. Kendall, "A New Measure of Rank Correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938.

[26] C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan, "Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation," in *Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 17–53.