DFKI's experimental hybrid MT system for WMT 2015

Eleftherios Avramidis, Maja Popović* and Aljoscha Burchardt
German Research Center for Artificial intelligence (DFKI)
Language Technology Lab
firstname.lastname@dfki.de
∗ Humboldt University of Berlin
maja.popovic@hu-berlin.de

Abstract

DFKI participated in the shared translation task of WMT 2015 with the German-English language pair in each translation direction. The submissions were generated using an experimental hybrid system based on three systems: a statistical Moses system, a commercial rule-based system, and a serial coupling of the two where the output of the rule-based system is further translated by Moses trained on parallel text consisting of the rule-based output and the original target language. The outputs of three systems are combined using two methods: (a) an empirical selection mechanism based on grammatical features (primary submission) and (b) IBM1 models based on POS 4-grams (contrastive submission).

1 Introduction

The system architecture we will describe has been developed within the QTLEAP project.[1] The goal of the project is to explore different combinations of shallow and deep processing for improving MT quality. The system presented in this paper is the first of a series of MT system prototypes developed in the project. Figure 1 shows the overall architecture that includes:

- A statistical Moses system,
- the commercial transfer-based system Lucy,
- their serial combination ("LucyMoses"), and
- an informed selection mechanism ("ranker").

The components of this hybrid system will be detailed in the sections below.

[1] http://qtleap.eu/

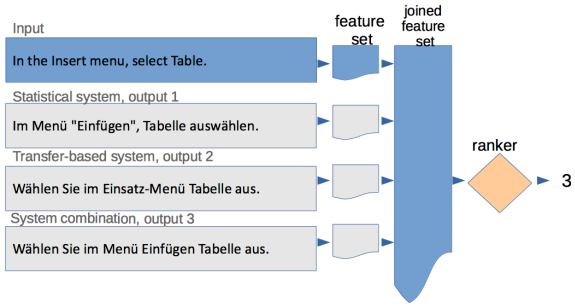

Figure 1: Architecture of System Combination.

2 Translation systems

Moses

Our statistical machine translation system was based on a vanilla phrase-based system built with Moses (Koehn et al., 2007) trained on the corpora Europarl ver. 7, News Commentary ver. 9 (Bojar et al., 2014), Commoncrawl (Smith et al., 2013) and MultiUN (Eisele and Chen, 2010). Language models of order 5 have been built and interpolated with SRILM (Stolcke, 2002) and KenLM (Heafield, 2011). For German to English, we also experimented with the method of pre-ordering the source side based on the target-side grammar (Popović and Ney, 2006). As a tuning set we used the *news-test 2013*.

Lucy

The transfer-based Lucy system (Alonso and Thurmair, 2003) includes the results of long linguistic efforts over the last decades and that has been used in previous projects including EUROMATRIX, EUROMATRIX+ and QTLAUNCHPAD, while relevant hybrid systems have been submitted to WMT (Chen et al., 2007; Federmann et al., 2010; Hunsicker et al., 2012). The transfer-based approach has shown good results that compete with pure statistical systems, whereas it focuses on translating according to linguistic struc-

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 66–73,
Lisboa, Portugal, 17-18 September 2015.

tures. Its functionality is based on hand-written linguistic rules and there are no major empirical components. Translations are processed on three phases:

- the **analysis phase**, where the source-language text is parsed and a tree of the source language is constructed
- the **transfer phase**, where the analysis tree is used for the transfer phase, where canonical forms and categories of the source are transferred into similar representations of the target language
- the **generation phase**, where the target sentence is formed out of the transfered representations by employing inflection and agreement rules.

LucyMoses

As an alternative way of automatic post-editing of the transfer-based system, a serial transfer+SMT system combination is used, as described in (Simard et al., 2007). For building it, the first stage is translation of the source language part of the training corpus by the transfer-based system. In the second stage, an SMT system is trained using the transfer-based translation output as a source language and the target language part as a target language. Later, the test set is first translated by the transfer-based system, and the obtained translation is translated by the SMT system. In previous experiments, however, the method on its own could not outperform Moses trained on a large parallel corpus. The example in Figure 1 (taken from the QTLEAP corpus used in the project) nicely illustrates how the serial coupling operates. While the SMT output used the right terminology ("Menü Einfügen" – "insert menu"), the instruction is not formulated in a very polite manner. In contrast, the output of the transfer-based system is formulated politely, yet mistranslating the menu type. The serial system combination produces a perfect translation. In this particular case, the machine translation is even better than the human reference ("Wählen Sie im Einfügen Menü die Tabelle aus.") as the latter is introducing a determiner for "table", which is not justified by the source.

2.1 Sentence level selection

We present two methods for performing sentence level selection, one with pairwise classifier and one based on POS 4-gram IBM1 models.

2.1.1 Empirical machine learning classifier (primary submission)

The machine learning (ML) selection mechanism is based on encouraging results of previous projects including EUROMATRIX+ (Federmann and Hunsicker, 2011), META-NET (Federmann, 2012), QTLAUNCHPAD (Avramidis, 2013; Shah et al., 2013). It has been extended to include several features that can only be generated on a sentence level and would otherwise blatantly increase the complexity of the transfer or decoding algorithm. In the architecture at hand, automatic syntactic and dependency analysis is employed on a sentence level, in order to choose the sentence that fulfills the basic quality aspects of the translation: (a) assert the fluency of the generated sentence, by analyzing the quality of its syntax (b) ensure its adequacy, by comparing the structures of the source with the structures of the generated sentence.

All produced features are used to build a machine-learned ranking mechanism (ranker) against training preference labels. Preference labels are part of the training data and rank different system outputs for a given source sentence based on the translation quality. Preference labels are generated either by automatic reference-based metrics, or derived from human preferences. The ranker was a result of experimenting with various combinations of feature sets and machine learning algorithms and choosing the one that performs best on the development corpus.

The implementation of the selection mechanism is based on the "Qualitative" toolkit that was presented at the MT Marathon, as an open-source contribution by QTLEAP (Avramidis et al., 2014).

Feature sets We experimented with feature sets that performed well in previous experiments. In particular:

- Basic syntax-based feature set: unknown words, count of tokens, count of alternative parse trees, count of verb phrases, PCFG parse log likelihood. The parsing was performed with the Berkeley Parser (Petrov and Klein, 2007) and features were extracted from both source and target. This feature set has performed well as a metric in WMT-11 metrics task (Avramidis et al., 2011).
- Basic feature set + 17 QuEst baseline features: this feature set combines the basic syntax-based feature set described above

with the baseline feature set of the QuEst toolkit (Specia et al., 2013) as per WMT-13 (Bojar et al., 2013). This feature set combination got the best result in WMT-13 quality estimation task (Avramidis and Popović, 2013). The 17 features set includes shallow features such as the number of tokens, LM probabilities, number of occurences of the target work within the target probability, average numbers of translations per source word in the sentence, percentages of unigrams, bigrams and trigrams in quartiles 1 and 4 of frequency of source words in a source language corpus and the count of punctuation marks.

Machine Learning As explained above, the core of the selection mechanism is a ranker which reproduces ranking by aggregating pairwise decisions by a binary classifier (Avramidis, 2013). Such a classifier is trained on binary comparisons in order to select the best out of two different MT outputs given one source sentence at a time. As a training material, we used the evaluation dataset of the WMT shared tasks (years 2008-2014), where each source sentence was translated by many systems and their outputs were consequently ranked by human annotators. These preference labels provided the binary pairwise comparisons for training the classifiers. Additionally to the human labels, we also experimented on training the classifiers against automatically generated preference labels, after ranking the outputs with METEOR (Banerjee and Lavie, 2005). In each translation direction, we chose the label type (human vs. METEOR) which maximizes if possible all automatic scores on our development set, including document-level BLEU.

We exhaustively tested all suggested feature sets with many machine learning methods, including Support Vector Machines (with both RBF and linear kernel), Logistic Regression, Extra/Decision Trees, k-neighbors, Gaussian Naive Bayes, Linear and Quadratic Discriminant Analysis, Random Forest and Adaboost ensemble over Decision Trees. The binary classifiers were wrapped into rankers using the *soft pairwise recomposition* (Avramidis, 2013) to avoid ties between the systems. When ties occurred, the system selected based on a predefined system priority (Lucy, Moses, LucyMoses). The predefined priority was defined manually based on preliminary observations in order to prioritize the transfer-based system, due to its tension to achieve better grammaticality. Further analysis on this aspect may be required.

Best combination The optimal systems are using:

1. the *Basic feature set + 17 QuEst baseline features* for German$rightarrow$English, trained with Suppor Vector Machines (Basak et al., 2007) against human ranking labels.

2. the *basic syntax-based feature set* for English→German, trained with Support Vector Machines against METEOR scores. METEOR was chosen since for this language pair, the empirical mechanism trained on human judgments had very low performance in term of correlation with humans.

2.1.2 POS 4-gram IBM1 models (contrastive submission)

Using the IBM1 scores (Brown et al., 1993) for automatic evaluation of MT outputs without reference translations has been proposed in Popović et al. (2011), and the best variant in terms of correlation with human ranking was the target-from-source direction based on POS 4-grams. Therefore, we investigated this variant for our sentence selection, and we submitted the obtained translation outputs as contrastive.

The IBM1 scores are defined in the following way:

$$\text{IBM1} = \frac{1}{(S+1)^H} \prod_{i=1}^{H} \sum_{j=0}^{S} p(h_i|s_j) \qquad (1)$$

where s_j are the POS 4-grams of the source language sentence, S is the POS 4-gram length of this sentence, h_i are the POS 4-grams of the target language translation output (hypothesis), and H is the POS 4-gram length of this hypothesis.

A parallel bilingual corpus for the desired language pair and a tool for training the IBM1 model are required in order to obtain IBM1 probabilities $p(h_i|s_j)$. For the POS n-gram scores, appropriate POS taggers for each of the languages are necessary. The POS tags cannot be only basic but must have all details (e.g. verb tenses, cases, number, gender, etc.).

The bilingual IBM1 probabilities used in our experiments are learnt from the German-English part of the WMT 2010 News Commentary bilingual corpora. Both German and English POS tags were produced using TreeTagger (Schmid, 1994).

3 Experimental results

Table 1 presents BLEU scores (Papineni et al., 2002), word F-scores and POS F-scores (Popović, 2011) for all individual systems and system combinations for both translation directions. The following interesting tendencies can be observed:

- German→English:
  - Moses and LucyMoses are comparable on the word level (BLEU and WORDF)
  - LucyMoses is best on the syntactic (POS) level
  - LucyMoses achieves better scores than both its components
  - using all three systems with a selection mechanism is the best option
- English→German:
  - Lucy is comparable with Moses on the word level and better on syntactic level
  - LucyMoses improves all scores
  - LucyMoses+Moses (LM+M) is the best combination for word level scores
  - Lucy+LucyMoses (L+LM) is comparable with the combination of all three systems (L+LM+M) for the syntactic oriented POSF score

We submitted the combination of all three systems for both selection mechanisms and for both translation directions. It should be noted that the ML classifier is used for the project's first official prototype, whereas the IBM1 classifier has been investigated only recently in the framework of the project – therefore the primary submission for the shared task is the ML classifier although it yielded lower automatic scores than the IBM1 classifier.

In order to estimate the limits of the classifiers for the given three MT systems, upper bound scores are presented in the last two rows, when selecting criteria were the WORDF and POSF scores themselves. It can be seen that there is a room for improvement for both selection methods. Further investigation, tuning and extension of the selection mechanisms will provide more insights and has potential for future improvements of the selection itself as well as of the MT systems.

Preliminary results concerning analysis of differences between the systems and behaviour of classifiers are shown in the following section.

3.1 Analysis of the results

First step towards better understanding of the selection mechanisms is to investigate the contribution of each of the individual systems in the final translation output. The results are presented in Table 2 in the form of percentage of sentences selected from each system. It is notable that:

- the ML classifier mostly favors the transfer-based output;
- for the English→German translation, the same holds for the IBM1 classifier; for the other translation direction, Lucy is selected very rarely – for less than 2% sentences;
- upper bound selection yields a more or less uniform distribution, however WORDF is clearly biased towards LucyMoses and POSF towards Lucy.

First indication is that the deep features of the ML classifier are active and therefore this classifier has a bias towards the transfer-based output. Furthermore, system contributions of upper bound selection methods indicate that the transfer-based outputs are more grammatical and thus favored by the syntax-oriented POSF score, whereas the LucyMoses system, which can be seen as a lexical reparation of a grammatical output, is favored by the lexical WORDF score. Nevertheless, these first hypotheses need to be confirmed by further studies that are planned.

Table 3 shows examples of differences between the selection methods as well as between the three individual MT systems. The sentences are taken from the WMT-15 test set. First column denotes the selection method which choose the particular translation output. Sentence 1 illustrates the differences between two classifiers as well as between two F-scores; POSF score and ML classifier opt for the transfer-based translation, whereas IBM1 choses Moses and WORDF score prefers LucyMoses. Sentences 2-4 show the discrepance between the ML classifier and the automatic scores; the IBM1 score selection differs from the upper bound selections only for the sentence 4. Such sentences are the most probable reason for lower overall MLC performance in terms of automatic scores. The last sentence shows an example where both classifiers agree, but they disagree with both F-scores.

(a) De→En

German→English			BLEU	WORDF	POSF
individual systems		Lucy (L)	20.8	25.9	42.6
		Moses (M)	23.2	28.2	42.7
		LucyMoses (LM)	23.2	27.9	44.2
selection mechanism	ML classifier	L+LM+M	22.6	27.4	43.6
	POS 4-gram IBM1	L+M	23.2	28.2	42.8
		L+LM	23.2	27.9	44.2
		LM+M	23.7	28.6	44.5
		L+LM+M	23.7	28.6	44.5
upper bounds	max(WORDF)	L+LM+M	**26.9**	**30.8**	46.8
	max(POSF)	L+LM+M	25.6	30.7	**48.6**

(b) En→De

English→German			BLEU	WORDF	POSF
individual systems		Lucy (L)	17.3	22.9	44.5
		Moses (M)	17.1	23.1	41.9
		LucyMoses (LM)	18.9	24.4	45.3
selection mechanism	ML classifier	L+LM+M	18.1	23.7	44.4
	POS 4-gram IBM1	L+M	18.2	23.6	44.7
		L+LM	18.6	24.0	45.7
		LM+M	19.1	24.4	45.1
		L+LM+M	18.9	24.1	45.4
upper bounds	max(WORDF)	L+LM+M	**22.4**	**26.6**	47.1
	max(POSF)	L+LM+M	21.0	26.1	**49.4**

Table 1: Translation results [%] for the German-English language pair.

(a) De→En

German→English		Lucy	Moses	LucyMoses
ML classifier		42.1	36.6	21.3
POS 4-gram IBM1	L+M	2.8	97.2	/
	L+LM	2.5	/	97.5
	LM+M	/	42.4	57.6
	L+LM+M	1.7	56.0	42.3
WORDF	L+LM+M	29.3	31.8	38.9
POSF	L+LM+M	34.5	33.7	31.8

(b) En→De

English→German		Lucy	Moses	LucyMoses
ML classifier		44.0	8.0	48.0
POS 4-gram IBM1	L+M	56.5	43.5	/
	L+LM	63.3	/	36.7
	LM+M	/	45.5	54.5
	L+LM+M	41.5	22.1	36.3
WORDF	L+LM+M	34.2	29.4	36.3
POSF	L+LM+M	42.3	27.1	30.5

Table 2: Percentage of selected sentences from each individual system.

The table also illustrates advantages of the serial LucyMoses system – this system produces the best translation output for all presented sentences except for sentence 3.

4 Summary and outlook

We described a hybrid MT system based on three different individual systems where the final translation output is produced by a sentence level selection mechanism, with the possibility to include deep linguistic and grammatical features. Preliminary analysis suggests that various improvements are possible, starting from improvements on the transfer-based system (handling of lexical items such as terminology, MWEs, OOVs and robustness of parsing), the serial combination (e.g., improved disambiguation), up to more detailed analysis and testing and improvement of the selection mechanism (e.g., integrating more "deep" information from external parsing).

Acknowledgments

This paper has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 610516 (QTLeap: Quality Translation by Deep Language Engineering Approaches). We are grateful to the anonymous reviewers for their valuable feedback.

References

Juan A. Alonso and Gregor Thurmair. 2003. The comprendium translator system. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, LA, September.

Eleftherios Avramidis and Maja Popović. 2013. Machine learning methods for comparative and time-oriented Quality Estimation of Machine Translation output. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 329–336, Sofia, Bulgaria, August. Association for Computational Linguistics.

Eleftherios Avramidis, Maja Popović, David Vilar, and Aljoscha Burchardt. 2011. Evaluate with Confidence Estimation : Machine ranking of translation outputs using grammatical features. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 65–70, Edinburgh, Scotland, July. Association for Computational Linguistics.

Eleftherios Avramidis, Lukas Poustka, and Sven Schmeier. 2014. Qualitative: Open source python tool for quality estimation over multiple machine translation outputs. *The Prague Bulletin of Mathematical Linguistics*, 102(1):5–16.

Eleftherios Avramidis. 2013. Sentence-level ranking with quality estimation. *Machine Translation (MT)*, 28(Special issue on Quality Estimation):1–20.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. In *Proceedings of the ACL 05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI, June.

Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. 2007. Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10):203–224.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *8th Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Yu Chen, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus, and Silke Theison. 2007. Multi-Engine Machine Translation with an Open-Source (SMT) Decoder. In *Proceedings of the Second Workshop on Statistical Machine Translation.*, pages 193–196. Association for Computational Linguistics.

Andreas Eisele and Yu Chen. 2010. MultiUN: A Multilingual Corpus from United Nation Documents. In Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odjik, Joseph Mariani, Bente Maegaard, Khalid Choukri, and Nicoletta Calzolari (Conference Chair), editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-2010), May 19-21, La Valletta, Malta*, pages 2868–2872. European Language Resources Association (ELRA).

1)	src:	Die Geschichte erinnert sich, und das sollten wir auch tun.
	ref:	History remembers, as should we.
POSF, MLC	Lucy:	The history remembers, and we should also do that.
IBM1	Moses:	Remembers the history, and we should do this.
WORDF	LucyMoses:	History remembers, and we should do the same.
2)	src:	Eine neue Runde indirekter Gespräche wird voraussichtlich noch in diesem Monat in Ägypten beginnen .
	ref:	A new round of indirect talks is expected to begin later this month in Egypt.
MLC	Lucy:	A new round of indirect conversations will probably still begin in this month in Egypt.
	Moses:	A new round of indirect talks is likely to begin in this month in Egypt.
WORDF, POSF, IBM1	LucyMoses:	A new round of indirect talks is likely to begin this month in Egypt.
3)	src:	Ich denke schon.
	ref:	I think so.
	Lucy:	I already think.
WORDF, POSF, IBM1	Moses:	I think so.
MLC	LucyMoses:	I have already think.
4)	src:	Über mehrere Jahre hatte niemand in dem Haus gelebt.
	ref:	No one had lived in the house for several years.
WORDF, POSF	Lucy:	Over several years nobody had lived in the house.
IBM1	Moses:	No one had over several years lived in the House.
MLC	LucyMoses:	For several years, no one had lived in the House.
5)	src:	Mach es nicht schlecht, wenn du nicht weißt, wovon du redest.
	ref:	Don't slag it off if you don't know what you're talking about.
	Lucy:	Do not make it bad if you do not know which you talk about.
MLC, IBM1	Moses:	Do it not bad, if you do not know what they are.
WORDF, POSF	LucyMoses:	Do not make it bad if you do not know what they are talking about.

Table 3: Examples of differences between the selection results and between the three individual systems.

Christian Federmann and Sabine Hunsicker. 2011. Stochastic Parse Tree Selection for an Existing RBMT System. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 351–357, Edinburgh, Scotland, July. Association for Computational Linguistics.

Christian Federmann, Andreas Eisele, Hans Uszkoreit, Yu Chen, Sabine Hunsicker, and Jia Xu. 2010. Further Experiments with Shallow Hybrid MT Systems. In Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan, editors, *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. Workshop on Statistical Machine Translation (WMT-10), located at ACL 2010, July 15-16, Uppsala, Sweden*, pages 77–81, 209 N. Eighth Street Stroudsburg, PA 18360 USA. Association for Computational Linguistics (ACL), ACL.

Christian Federmann. 2012. Can Machine Learning Algorithms Improve Phrase Selection in Hybrid Machine Translation? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation*, pages 113–118, Avignon, France, April. European Chapter of the Association for Computational Linguistics (EACL).

Kenneth Heafield. 2011. KenLM : Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, number 2009, pages 187–197, Edinburgh, Scotland, July. Association for Computational Linguistics.

Sabine Hunsicker, Chen Yu, and Christian Federmann. 2012. Machine Learning for Hybrid Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 312–316, Montréal, Canada, June. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 02)*, pages 311–318, Philadelphia, PA, July.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of the 2007 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York. Association for Computational Linguistics.

Maja Popović and Hermann Ney. 2006. POS-based Word Reorderings for Statistical Machine Translation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 06)*, pages 1278–1283, Genoa, Italy, May.

Maja Popović, David Vilar Torres, Eleftherios Avramidis, and Aljoscha Burchardt. 2011. Evaluation without references: IBM1 scores as evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 99–103, Edinburgh, Scotland, July.

Maja Popović. 2011. Morphemes and POS tags for n-gram based evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 104–107, Edinburgh, Scotland, July.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer.

Kashif Shah, Eleftherios Avramidis, Ergun Biçici, and Lucia Specia. 2013. QuEst: Design, Implementation and Extensions of a Framework for Machine Translation Quality Estimation. *The Prague Bulletin of Mathematical Linguistics (PBML)*, 100:19–30.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Proceedings of The North American Chapter of the Association for Computational Linguistics Conference (NAACL-07)*, pages 508–515, Rochester, NY, April.

Jason R Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt Cheap Web-Scale Parallel Text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria, August. Association for Computational Linguistics.

Lucia Specia, Kashif Shah, José Guilherme Camargo de Souza, and Trevor Cohn. 2013. QuEst - A translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria, August. Association for Computational Linguistics.

Andreas Stolcke. 2002. Srilm — an Extensible Language Modeling Toolkit. In *System*, volume 2, pages 901–904. ISCA, September.