Results of the WMT16 Tuning Shared Task

Bushra Jawaid, Amir Kamran, Miloš Stanojević
University of Amsterdam
ILLC
{initial.last}@uva.nl

Ondřej Bojar
Charles University in Prague
MFF ÚFAL
bojar@ufal.mff.cuni.cz

Abstract

This paper presents the results of the WMT16 Tuning Shared Task. We provided the participants of this task with a complete machine translation system and asked them to tune its internal parameters (feature weights). The tuned systems were used to translate the test set and the outputs were manually ranked for translation quality. We received 4 submissions in the Czech-English and 8 in the English-Czech translation direction. In addition, we ran 2 baseline setups, tuning the parameters with standard optimizers for BLEU score. In contrast to previous years, the tuned systems in 2016 rely on large data.

1 New Introduction

The standard phrase based and hierarchical statistical machine translation (SMT) systems rely on several models that predict the hypothesis quality. Some of them are taking care that the translations are lexically correct (translation models), some that it is fluent (language models), some that it is not too long (word and phrase penalty) etc. The list of features can go from a dozen to a more than million of sparse features.

Clearly, not all of these features are equally important. For this reason they are combined in a linear model in which each one of the features is assigned a weight that scales its contribution to the total score of the hypothesis.

Estimating these weights has been an important part of MT research for many years. Different learning algorithms have been published, some helpful features proposed and many evaluation metrics considered as alternative objectives for optimization. In search for the best combination of proposed components of weight estimation, we organize this task in which the potential solutions can compete in a controlled setting: a fixed system to be optimized and a fixed tuning and test set. Everything else is up to the participants.

This way of evaluation of the tuning algorithms and objectives can settle some of the dilemmas that existed in the community. For example, is KBMIRA better than MERT? The choice is usually based on recommendations between researchers or by their comparison on BLEU score which is not always the best way to compare two systems. In this task, we compare the systems based on how humans judge the output of these systems.

Another very common design choice is which objective to optimize. The evaluation metrics are usually designed to correlate well with human judgments of translation quality, see Bojar et al. (2016c) and the previous papers summarizing WMT metrics tasks. However, a metric that correlates well with humans on final output quality may not be usable in weight optimization for various technical reasons. Many metrics that have very high correlation with human judgment achieve that by using complex models that are very slow so they might present a bottle-neck in the tuning process when the chosen evaluation metric needs to evaluate a huge number of translations in the n-best lists.

BLEU (Papineni et al., 2002) was shown to be very hard to surpass (Cer et al., 2010) as a tuning metric and this is also confirmed by the previous WMT15 Tuning Task results (Stanojević et al., 2015) and by the results of the invitation-only WMT11 Tunable Metrics Task (Callison-Burch et al., 2010)[1]. Note however, that some metrics have been successfully used for system tuning (Liu et al., 2011; Beloucif et al., 2014).

The aim of the WMT16 Tuning Task[2] is (just like in WMT15 Tuning Task) to attract attention

[1] http://www.statmt.org/wmt11/tunable-metrics-task.html

[2] http://www.statmt.org/wmt16/tuning-task/

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 232–238,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

		Sentences		Tokens		Types	
	Source	cs	en	cs	en	cs	en
LM corpora	Europarl v7, News Commentary v11, News Crawl (2007-15), News Discussion v1	54M	206M	900M	4409M	2.1M	3.2M
TM corpora	CzEng 1.6pre for WMT16	44M		501M	587M	1.8M	1.2M
Dev set	newstest2015	2656		46K	54K	12.9K	7.7K
Test set	newstest2016	2999		56.9K	65.3K	15.1K	8.8K

Table 1: Data used in the WMT16 tuning task.

	Dev		Test	
Direction	Token	Type	Token	Type
en-cs	391	314	644	486
cs-en	289	199	507	331

Table 2: Out of vocabulary word counts

to the exploration of all the three aspects of model optimization: (1) the set of features in the model, (2) optimization algorithm, and (3) MT quality metric used in optimization.

For (1), we provide a fixed set of "dense" features and also allow participants to add additional "sparse" features. For (2), the optimization algorithm, task participants are free to use one of the available algorithms for direct loss optimization (Och, 2003; Zhao and Chen, 2009), which are usually capable of optimizing only a dozen of features, or one of the optimizers handling also very large sets of features (Cherry and Foster, 2012; Hopkins and May, 2011), or a custom algorithm. And finally for (3), participants can use any established evaluation metric or a custom one.

1.1 Tuning Task Assignment

The way the tuning task is organized is the same as in the previous WMT15 tuning task (Stanojević et al., 2015). Tuning task participants were given a complete model for the phrase-based variant of the machine translation system Moses (Koehn et al., 2007) and the development set (newstest2015), i.e. the source and reference translations. No "dev test" set was provided, since we expected that participants will internally evaluate various variants of their method by manually judging MT outputs. In fact, we offered to evaluate a certain number of translations into Czech for free to ease the participation for teams without any access to speakers of Czech.

A complete model consists of a phrase table extracted from the parallel corpus, two lexicalized reordering tables and the two language model extracted from the monolingual data. As such, this defines a fixed set of dense features which is bigger than last year both in the number of additional models and in the size of the models themselves (language models are trained on much bigger datasets). The participants were allowed to add any sparse features implemented in Moses Release 3.0 (corresponds to Github commit 2d6f616) and/or to use any optimization algorithm and evaluation metric.

Each submission in the tuning task consisted of the configuration of the MT system, i.e. the additional sparse features (if any) and the values of all the feature weights.

2 Details of Systems Tuned

The systems that were distributed for tuning are based on Moses (Koehn et al., 2007) implementation of phrase-based model. The language models were 5-gram models built using KenLM (Heafield et al., 2013) with modified Kneser-Ney smoothing (James, 2000) without pruning. For word alignments, we used fast-align toolkit (Dyer et al., 2013). Alignments are computed in both directions and symmetrized using *grow-diag-final-and* heuristic.

We use CzEng 1.6pre[3] (Bojar et al., 2016b) parallel data for the extraction of translation models. We train two language models for each translation direction: the first model is trained on CzEng 1.6pre target data and the second model is trained on concatenation of Europarl v7, News Commentary data (`parallel-nc-v11`), news data (2007-2013, 2014-v2, 2015) and additionally news discussion v1 (for English language model only), as released for WMT16[4]. We excluded CommonCrawl data because we wanted to avoid data without a clear match with the news domain.

Besides the translation tables and language models we also provided two lexicalised reordering models for each direction. Both reordering

[3] http://ufal.mff.cuni.cz/czeng/czeng16pre

[4] http://www.statmt.org/wmt16/translation-task.html

System	Participant
BLEU-MIRA, BLEU-MERT	baselines
AFRL	United States Air Force Research Laboratory (Gwinnup et al., 2016)
DCU	Dublin City University (Li et al., 2015)
FJFI-PSO	Czech Technical University in Prague (Kocur and Bojar, 2016)
ILLC-UvA-BEER	ILLC – University of Amsterdam (Stanojević and Sima'an, 2015)
NRC-MEANT, NRC-NNBLEU	National Research Council Canada (Lo et al., 2015)
USAAR	Saarland University (Liling Tan; no corresponding paper)

Table 3: Participants of WMT16 Tuning Shared Task

models were extracted using code readily available in Moses. One of the models is word-based (Koehn et al., 2005) and the other is hierarchical (Galley and Manning, 2008). Both reordering models use *msd* orientation in both forward and backward direction, with model conditioned on both the source and target languages (msd-bidirectional-fe).

Before any further processing, the data was pre-tokenized and tokenized (using standard Moses scripts) and lowercased. We also removed parallel sentences longer than 60 words or shorter than 4 words, no data cleaning was performed for monolingual data. Table 1 summarizes the final dataset sizes and Table 2 provides details on out-of-vocabulary items.

Aside from the dev set provided, the participants were free to use any other data for tuning (making their submission "unconstrained"), but no participant decided to do that. All tuning task submissions are therefore also constraint in terms of the WMT16 Translation Task (Bojar et al., 2016a).

We leave all decoder settings (n-best list size, pruning limits etc.) at their default values. While the participants may have used different limits during tuning, the final test run was performed at our site with the default values. It is indeed only the feature weights that differ.

3 Tuning Task Participants

The list of participants and the names of the submitted systems are shown in Table 3.

We provide a brief summary of each evaluated optimization method in the rest of this section, concluding with baseline approaches (Section 3.7).

3.1 ILLC-UvA-BEER

ILLC-UvA-BEER (Stanojević and Sima'an, 2015) was tuned using PRO (Hopkins and May, 2011) learning algorithm with new version of BEER evaluation metric. The authors claim that common trained evaluation metrics learn to give too much importance to recall and thus lead to overly long translations in tuning. For that reason, they modify the training of BEER to value recall and precision equally. This modified version of BEER is used to train the MT system.

3.2 NRC-MEANT and NRC-NNBLEU

NRC-MEANT is a system tuned against MEANT (Lo et al., 2015) using batch MIRA with an additional length penalty to avoid semantic parsing unreasonably long MT output. Due to the additional huge language model in this year's baseline, the MT system would generate unreasonably long MT output in the second iteration of the tuning cycle. This severely affects the running time of MEANT because running automatic semantic parser on long sentences is costly. Therefore, a length penalty is implemented in MEANT: for MT output that is 2 times or 15 word tokens longer than the reference, MEANT does not run SRL on it and falls back to the backoff bag-of-word phrasal similarity. This could be one of the reasons why MEANT-tuned system is not performing as competitive as last year.

NRC-NNBLEU is a system tuned against a new metric that replaces the n-gram exact match in BLEU with n-gram word embeddings cosine similarity.

3.3 DCU

DCU (Li et al., 2015) is tuned with RED, an evaluation metric based on matching of dependency n-grams. As tuning algorithm the authors have used KBMIRA.

3.4 AFRL1 and AFRL2

As in the previous year's submissions (Erdmann and Gwinnup, 2015), the AFRL systems used Drem, which is a derivative-free optimization algorithm that interpolates n-best lists returned by the decoder. Methodology for the current tuning task is nearly identical, since recent changes to Drem mostly relate to improving treatment of n-best list rescoring techniques (Gwinnup et al., 2016). The objective function used within Drem is the same for cs-en AFRL and en-cs system AFRL1:

$$0.045 \cdot \text{NIST} + 0.45 \cdot \text{Meteor} + 0.1 \cdot \text{Kendall's } \tau.$$

The en-cs system AFRL2 uses the following objective, which tests the sensitivity of the result to the metric and the suitability of the metric chrF3 (Popović, 2015) as a tuning metric:

$$0.045 \cdot \text{NIST} + 0.45 \cdot \text{chrF3} + 0.1 \cdot \text{Kendall's } \tau.$$

The practice of regularizing each metric by using expected (i.e., soft-max) sufficient statistics is maintained as before (Erdmann and Gwinnup, 2015).

3.5 USAAR-*

USAAR submissions are similar to the ones from last year. They use both KBMIRA and MERT for tuning and combine them in different ways. USAAR-HMM trains with KBMIRA and MERT independently and then combines the weights of the final iterations by using harmonic mean. USAAR-HMM-MIRA is the same as USAAR-HMM except that after the harmonic mean is computed, the tuning is continued with KBMIRA for additional 25 iterations. USAAR-HMM-MERT is the same as USAAR-HMM-MIRA except that MERT is used instead of KBMIRA for continuing the training after harmonic averaging.

3.6 FJFI-PSO

FJFI-PSO (Kocur and Bojar, 2016) replaces the "inner optimization loop" in Moses MERT with Particle Swarm Optimization, an algorithm that lends itself easily to parallelization. Everything else in Moses MERT is unchanged and FJFI-PSO optimizes to the default BLEU.

3.7 Baseline Methods

In addition to the systems submitted, we provided three baselines:

- BLEU-MERT-DENSE – MERT tuning with BLEU without additional features
- BLEU-MIRA-DENSE – KBMIRA tuning with BLEU without additional features

Since all the submissions including the baselines were subject to manual evaluation, we did not run the MERT or MIRA optimizations more than once (as is the common practice for estimating variance due to optimizer instability). We simply used the default settings and stopping criteria and picked the weights that performed best on the dev set according to BLEU.

4 Results

We used the submitted `moses.ini` and (optionally) sparse `weights` files to translate the test set. The test set was not available to the participants at the time of their submission (not even the source side). We used the Moses recaser trained on the target side of the parallel corpus to recase the outputs of all the models.

Finally, the recased outputs were manually evaluated, jointly with regular translation task submissions of WMT (Bojar et al., 2016a). Monitoring the results of the tuning task already during the manual evaluation period, we observed that tuning systems perform very similarly. When most of the evaluated language pairs collected sufficient number of manual judgements, we asked the organizers of the translation task evaluation to reopen annotation interface for tuning systems, hoping for better separation of the submissions. The WMT16 evaluation data thus contain a number of annotation items where all ranked translation correspond to output of a tuning system. This subset of annotations may be of special interest, e.g. to analyze the behaviour of annotators when all candidate translations are very similar.

The resulting human rankings were used to compute the overall manual score using the TrueSkill method, same as for the main translation task (Bojar et al., 2016a).[5]

Tables 4 and 5 contain the results of the submitted systems sorted by their manual scores.

The horizontal lines represent separation between clusters of systems that perform similarly. Cluster boundaries are established by the same method as for the main translation task.

[5] As in previous year, we also checked TrueSkill scores when only tuning systems would be considered. Since this non-standard evaluation leads to the same clusters of similarly-performing systems as the official TrueSkill does, we do not report it here.

System Name	TrueSkill Score	BLEU
BLEU-MIRA	0.114	22.73
AFRL	0.095	22.90
NRC-NNBLEU	0.090	23.10
NRC-MEANT	0.073	22.60
ILLC-UvA-BEER	0.032	22.46
BLEU-MERT	0.000	22.51

Table 4: Results on Czech-English tuning

System Name	TrueSkill Score	BLEU
BLEU-MIRA	0.160	15.12
ILLC-UvA-BEER	0.152	14.69
BLEU-MERT	0.151	14.93
AFRL2	0.139	14.84
AFRL1	0.136	15.02
DCU	0.134	14.34
FJFI-PSO	0.127	14.68
USAAR-HMM-MERT	-0.433	7.95
USAAR-HMM-MIRA	-1.133	0.82
USAAR-HMM	-1.327	0.20

Table 5: Results on English-Czech tuning

5 Discussion

We see that manual evaluation of tuning systems can draw only very few clear division lines. Czech-to-English has only two clusters of significantly differing quality and English-to-Czech is even less discerning: all except USAAR-* systems fall into the same cluster. The low number of clusters was obtained also last year, but this year, we believe that the situation is worsened by the large-scale setup of the tuned systems.

There are a few observations that can be made about the baseline results.

Just like last year, KBMIRA turns out to consistently be better than MERT even for the system with small number of features. The difference is especially big for Czech-English where system tuned with MERT ended up as the worst and system tuned with KBMIRA as the best.

In fact, KBMIRA tuning for BLEU is not only better than MERT but better than any other tuning system for both language pairs. This baseline is a clear winner of this task. Some systems that did well last year did not repeat their success this year. For example, the last year's winner for English-Czech DCU was unfortunately worse than both baselines and three other systems.

Except for the winning baseline, the results do not generalize much over translation direction. ILLC-UvA-BEER is second best in English-Czech but second worst in Czech-English. Its success on English-Czech can probably be explained by character-level scoring that is important for morphologically rich language such as Czech.

The submitted systems used different combinations of tuning algorithms (MERT, KBMIRA, PRO, Drem or combinations of MERT and KBMIRA) and different metrics (BEER, BLEU, RED, MEANT and combinations of chrF, NIST, METEOR and Kendall τ) so it is difficult to see which aspect of the system contributed most to its results. Systems that we can compare directly are for example AFRL1 and AFRL2 where the main difference was that AFRL2 uses chrF3 in its mixture of metrics instead of METEOR. This particular variation has contributed to slight improvement in human score, but it degraded the BLEU score.

Optimizing for BLEU does not seem to be always beneficial. Even though the systems tuned for BLEU did well in the task, the systems that got the best BLEU scores are not the winning systems. For Czech-English, NRC-NNBLEU got the best BLEU score result, but it ended up third. Also, tuning for BEER with PRO consistently outperforms tuning for BLEU with MERT. It is difficult to say whether this is because PRO is a better learning algorithm or because BEER is a better metric. However, if we use KBMIRA instead of MERT then evaluation with BLEU seems to be sufficient to outperform all the other systems.

6 Conclusion

We presented the results of WMT16 Tuning Task, a shared task in optimizing parameters of a given phrase-based system when translating from English to Czech and vice versa.

This year, the tuned system was a large-scale one, trained on almost all of the available data in the constrained translation task. All the tuning task submissions were thus on the scale of a standard WMT system, validating the applicability of proposed methods from practical point of view. Given that the number of submitted systems was very similar to last year, we conclude that the participants succeeded in this challenge.

Overall, six teams took part in one or both directions, sticking to the constrained setting.

The submitted configurations were manually evaluated jointly with the systems of the main WMT translation task.

The results confirm that KBMIRA with the standard (dense) features optimized towards BLEU should be preferred over MERT. The clear winner of the task was KBMIRA system tuned for BLEU score, although the quality of most submitted systems is hard to distinguish manually.

Acknowledgments

We would like to thank to Christian Federmann for support during manual evaluation and to Matt Post for providing us with various types of outputs from the manual WMT ranking. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements n^{o} 645452 (QT21) and n^{o} 644402 (HimL). The work on this project was also supported by the Dutch organisation for scientific research STW grant nr. 12271.

References

Meriem Beloucif, Chi-kiu Lo, and Dekai Wu. 2014. Improving MEANT Based Semantically Tuned SMT. In *Proc. of 11th International Workshop on Spoken Language Translation (IWSLT 2014)*, pages 34–41, Lake Tahoe, California.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016a. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, August. Association for Computational Linguistics.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016b. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings*. Springer Verlag, September 12-16. In press.

Ondřej Bojar, Yvette Graham, , and Amir Kamran Miloš Stanojević. 2016c. Results of the WMT16 Metrics Shared Task . In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, August. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July. Association for Computational Linguistics. Revised August 2010.

Daniel Cer, Christopher D. Manning, and Daniel Jurafsky. 2010. The Best Lexical Metric for Phrase-based Statistical MT System Optimization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 555–563, Stroudsburg, PA, USA. Association for Computational Linguistics.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 427–436, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June. Association for Computational Linguistics.

Grant Erdmann and Jeremy Gwinnup. 2015. Drem: The AFRL Submission to the WMT15 Tuning Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 848–856, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jeremy Gwinnup, Tim Anderson, Grant Erdmann, Katherine Young, Michaeel Kazi, Elizabeth Salesky, and Brian Thompson. 2016. The AFRL-MITLL WMT16 News-Translation Task Systems. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, August. Association for Computational Linguistics.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

Mark Hopkins and Jonathan May. 2011. Tuning As Ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1352–1362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Frankie James. 2000. Modified Kneser-Ney Smoothing of N-gram Models. Technical report.

Viktor Kocur and Ondřej Bojar. 2016. Particle Swarm Optimization Submission for WMT16 Tuning Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, August. Association for Computational Linguistics.

Philipp Koehn, Amittai Axelrod, Alexandra B. Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *Proceedings of International Workshop on Spoken Language Translation*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris 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.

Liangyou Li, Hui Yu, and Qun Liu. 2015. MT Tuning on RED: A Dependency-Based Evaluation Metric. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2011. Better Evaluation Metrics Lead to Better Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 375–384, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chi-kiu Lo, Philipp Dowling, and Dekai Wu. 2015. Improving evaluation and optimization of MT systems against MEANT. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, 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 on Association for Computational Linguistics*, ACL '02, pages 311–318.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.

Miloš Stanojević and Khalil Sima'an. 2015. BEER 1.1: ILLC UvA submission to metrics and tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.

Miloš Stanojević, Amir Kamran, and Ondřej Bojar. 2015. Results of the WMT15 Tuning Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.

Bing Zhao and Shengyuan Chen. 2009. A Simplex Armijo Downhill Algorithm for Optimizing Statistical Machine Translation Decoding Parameters. In *HLT-NAACL (Short Papers)*, pages 21–24. The Association for Computational Linguistics.