АБУМАТРАН
ABUMATRAN

Abu-MaTran

Automatic building of Machine Translation

PIAP- GA-2012-324414

D5.1a Evaluation of the MT systems deployed in the first development cycle

Dissemination level	Public
Delivery date	2013/08/31
Status and version	Final, 1.0
Authors and affiliation	Antonio Toral (DCU), Santiago Cortés-Vaíllo (Prompsit), Gema Ramírez-Sánchez (Prompsit), Filip Klubička (UZ), Nikola Ljubešić (UZ)

Project funded by the European Community under the Seventh Framework Programme for Research and Technological Development

Table of Contents

Executive Summary

One of the goals of project Abu-MaTran is to rapidly build a machine translation system from English to Croatian. Indeed, after collecting the necessary corpora (described in deliverable D3.1a) a number of systems were built (described in deliverable D4.1a) and the best one (according to the evaluation described in this deliverable) was made available through http://translator.abumatran.eu/ on July 1, simultaneously with the accession of Croatia to the European Union.

This deliverable describes the automatic evaluation of the English–Croatian MT systems built in the first development cycle.

1 Introduction

This deliverable describes the automatic evaluation of the English–Croatian machine translation (MT) systems built in the first development cycle (as described in Deliverable 4.1a). In the next sections we describe the metrics used to evaluate the MT systems, we present the evaluation data sets and we show and discuss the results.

2 Evaluation Metrics

To evaluate the MT systems we have used a set of representative state-of-the-art automatic metrics. Given an evaluation dataset (made up of a set of human-translated source and target sentences), these metrics compare the output produced by an MT system to the reference and provide a score (typically in the range 0 to 1). Thus, all these metrics require a reference translation.

These are the three metrics we have used in the evaluation procedure:

- BLEU (Papineni et al., 2002), which could be considered the *de facto* standard automatic metric to evaluate MT systems nowadays, at least in the statistical research community. It is a string-based metric and its critics argue that it tends to favour statistical systems over rule-based ones (Callison-Burch et al., 2006). We have used version 13a from the mteval toolkit.[1]
- METEOR[2] (Banerjee and Lavie, 2005) is also a string-based metric but uses additional linguistic information (stemming and synonyms) to provide a more fine-grained evaluation at the lexical level. A drawback of this metric is that it is partly language-dependent as it requires a stemmer and WordNet.[3] Therefore its degree of applicability depends on the resources availables for each language (from full application for languages for which there are synonymy and paraphrases to basic application for languages for which none of these modules is available). METEOR-NEXT (Denkowski and Lavie, 2010) is an updated version of the same metric. We have used version 1.4.
- TER (Snover et al., 2006) is an error-based metric. It adopts a different approach, in that it computes the number of substitutions, insertions, deletions and shifts that are required to modify the output translation so that it completely matches the reference translation(s) . The rationale behind this evaluation metric is quite simple to understand for people who are not MT experts, as it provides an estimation of the amount of post-editing effort needed by an end-user. We have used the version of the TER metric included in TER-Plus version 0.1.[4]

Finally, we also report the coverage of the MT systems evaluated. For this we consider out-of-vocabulary (OOV) words, i.e. the percentage of words in the input that are unknown to the MT system. OOV percentages are calculated with the script included in the Moses toolkit.

1 ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl
2 https://www.cs.cmu.edu/~alavie/METEOR/
3 http://wordnet.princeton.edu/
4 http://www.umiacs.umd.edu/~snover/terp/

3 Evaluation Datasets

We have used several sources of English–Croatian parallel data covering different domains and styles for the evaluation:

1. SETimes, a corpus from content published on the SETimes.com news portal. While most of this corpus is used for training, we left aside 2,000 sentence pairs, which are used for the evaluation.
2. Tatoeba,[5] a corpus from "a large database of example sentences translated into several languages". It could be considered as language-learners text. This source contains 691 sentence pairs.
3. WMT, a corpus in the news domain from the WMT 2013 shared task.[6] It is available in a number of European languages but not Croatian. We translated part of this corpus (1,000 sentences) into Croatian in order to use it in the evaluation.

4 Results

This section presents the results obtained by the different types of systems built: direct, pivot and direct with synthetic back-off. Finally we compare the results of the our best system to those of a general purpose on-line system.

In the remainder of this section we report results in term of BLEU scores. For the sake of clarity we omit results with the other metrics that we have used (TER and METEOR) since the trends are similar across all the metrics. Complete results including all the three metrics are provided in Appendix A.

4.1 Direct systems

Figures 1 and 2 show the BLEU scores for the direct systems for English to Croatian (systems EN-HR1 and EN-HR2) and Croatian to English (systems HR-EN1 and HR-EN2), respectively.

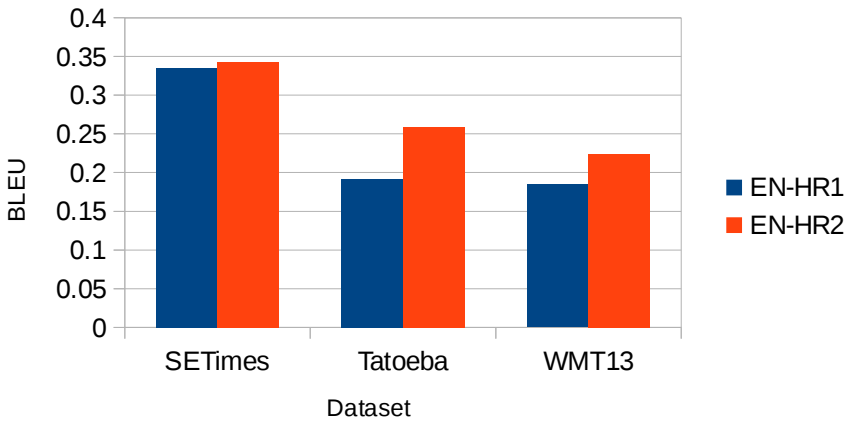

Figure 1: BLEU scores for direct systems (English→Croatian)

5 http://opus.lingfil.uu.se/Tatoeba.php
6 http://www.statmt.org/wmt13/translation-task.html

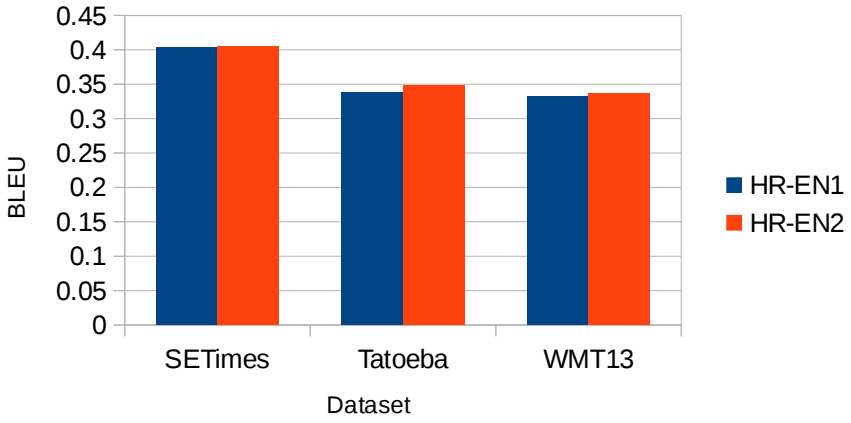

Figure 2: BLEU scores for direct systems (Croatian→English)

The difference between systems EN-HR1 and EN-HR2 (and between HR-EN1 and HR-EN2) is that EN-HR2 (and HR-EN2) add a considerably bigger language model built on monolingual data sources. The addition of such a language model results in notable improvements for both language directions and for all the three evaluation data sets.

4.2 Pivot-based systems

Figure 3 shows the BLEU scores obtained by the pivot-based systems for English to Croatian.[7] Our best direct system (EN-HR2) is also included as a baseline.

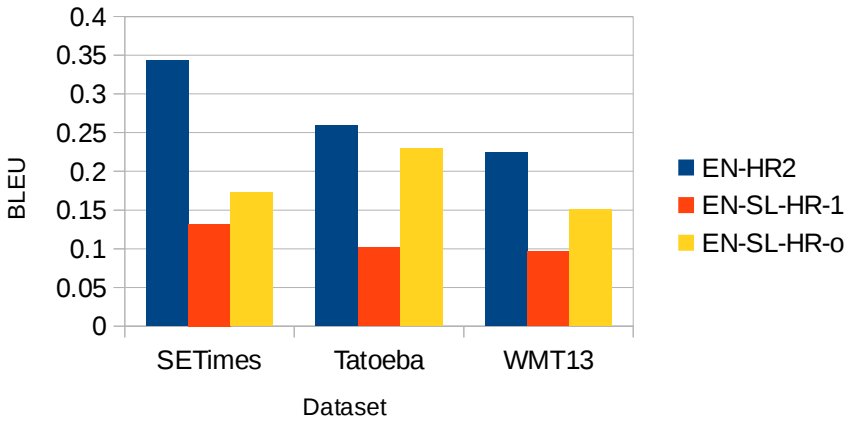

Figure 3: BLEU scores for pivot-based systems (English→Croatian)

7 Pivot systems are not run in the opposite direction (Croatian to English) as the framework we use (as explained in deliverable D4.1a) relies on a two-stage system (source to pivot and pivot to target) where the first system should provide *n* best outputs. In our use case, the first system in the English to Croatian direction, the English→Slovene system (SMT-based), can provide *n* translations, but the first system in the other direction, Croatian→Slovene (rule-based), provides only one translation.

EN-SL-HR-1 can be thought as the lower-bound attainable in the pivot-based approach as only the 1-best translation from the English to Slovene system is used. On the other hand, EN-SL-HR-o can be thought as the upper-bound as we consider *n* equals 2,000 and we provide an oracle (the Slovene→Croatian system translates the 2,000 best sentences obtained from the English→Slovene system and the best one, according to sentence-level BLEU score, is output).

For all the three evaluation data sets the pivot upper-bound system (EN-SL-HR-o) falls clearly behind the best direct system (EN-HR2). Despite the availability of far more data for the English–Slovene pair compared to the amount of data available for English–Croatian, our direct system outperforms the indirect system.

4.3 Direct systems with synthetic back-off

Next we evaluate a system that combines the direct and indirect approaches, by using the latter as an additional SMT phrase table for back-off (cf. Section 3.3. of D4.1a).

Figures 4 and 5 show the BLEU scores of the direct systems with synthetic back-off for English to Croatian (systems EN-HR2-bf and EN-HR2-b) and for Croatian to English (systems HR-EN2-bf and HR-EN2-b), respectively. In these systems "b" stands for back-off and "f" stands for filtered (only the subset of synthetic sentence pairs without unknown words are used). Our best direct systems (EN-HR2 and HR-EN2) are also included as baselines.

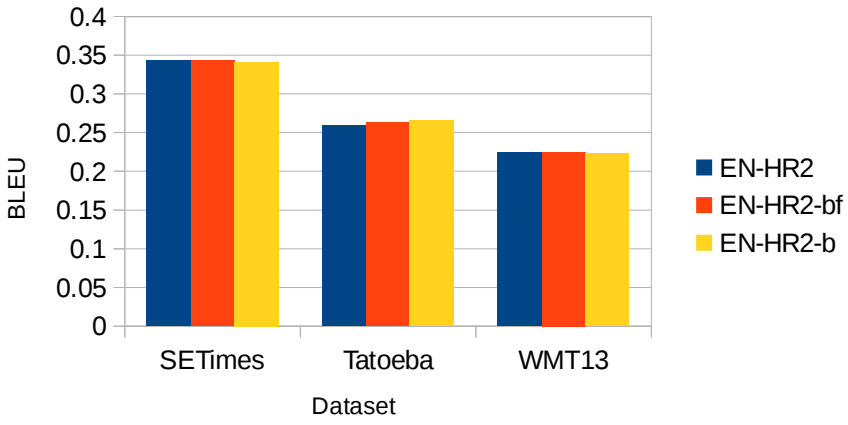

Figure 4: BLEU scores for direct systems with synthetic back-off (English→Croatian)

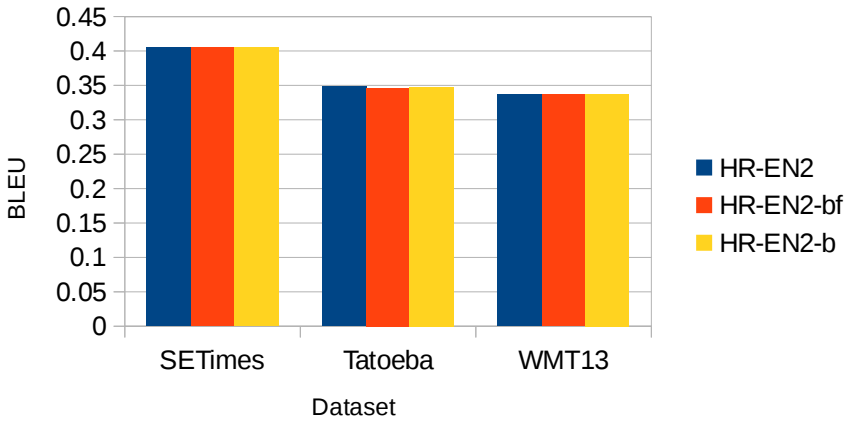

Figure 5: BLEU scores for direct systems with synthetic back-off (Croatian→English)

In all cases (for both language pairs and for all the three evaluation data sets), the differences in terms of BLEU are negligible. We analyse further the performance of these systems by looking at their coverage. Figures 6 and 7 show the percentage of OOVs for the direct systems with synthetic back-off and the best direct systems.

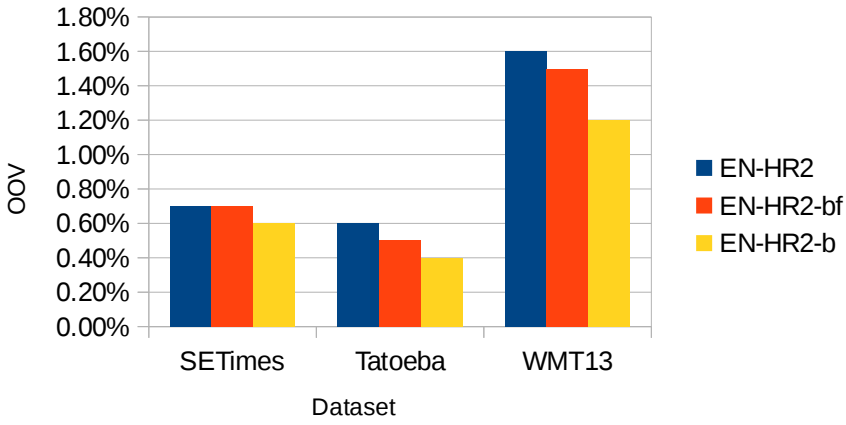

Figure 6: OOV scores for direct systems with synthetic back-off (English→Croatian)

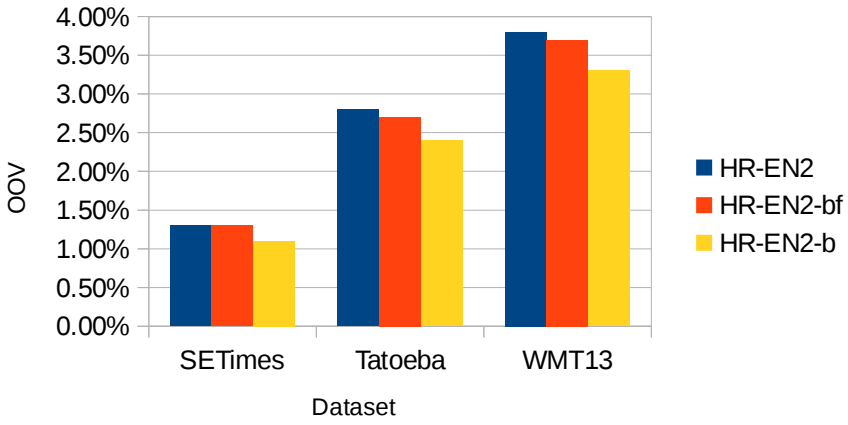

Figure 7: OOV scores for direct systems with synthetic back-off (Croatian→English)

While all the systems are very similar in terms of BLEU scores, there are clear differences when we look at OOVs. In all cases the direct systems with synthetic back-off provide better coverage than the direct system. Within the systems with synthetic back-off, the one that uses the full synthetic data provides better coverage than the one that uses filtered synthetic data.

4.4 Comparison with on-line systems

Finally we compare our best direct systems (EN-HR2 and HR-EN2) to a general purpose on-line system (Google Translate). Figures 8 and 9 show the BLEU scores for these systems for English to Croatian and Croatian to English, respectively.

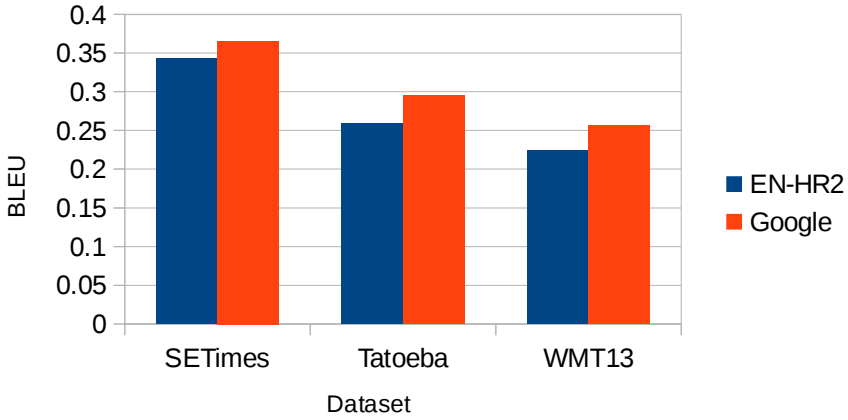

Figure 8: BLEU scores for our best system and Google (English→Croatian)

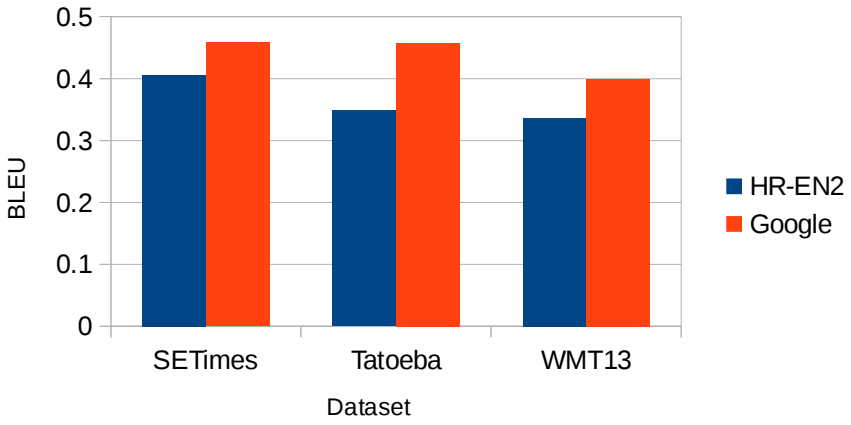

Figure 9: BLEU scores for our best system and Google (Croatian→English)

The on-line system obtains slightly better results for English to Croatian (2 to 3 points gap, depending on the data set), while its advantage widens for the opposite direction (6 to 10 points gap). This wider gap for translating into English is probably due to this third party system using a huge language model for English.

5 Conclusions and Future Work

This document has covered the evaluation of the MT systems that have been built for the English–Croatian language pair (in both directions), as described in D4.1a. The document has introduced the metrics to perform the evaluation as well as the evaluation datasets.

The best of these systems was made available on July 1, 2013 (the EU accession date for Croatia) through http://translator.abumatran.eu/.

It should be emphasised that we have built our systems for the first milestone in a very short period of time (month 6 of the project) using available open-source software tools and publicly available data sets for training and evaluation. However, the performance achieved by our systems is already close to that of general-purpose on-line MT systems. In this regard, our comparison with Google shows that our best system is quite close to Google's score for English→Croatian (2 to 3 points gap in terms of BLEU) while Google's advantage widens for the opposite direction, Croatian→English (6 to 10 points gap in terms of BLEU).

For the next milestone we plan to build upon these systems in two directions:

1. We will consider specific domains, crawling and building MT systems for domains that are relevant for the English – Croatian language pair (e.g. tourism, given the importance of this sector in Croatia's economy).
2. We will explore adding a layer of linguistic processing tailored for this language pair. E.g. Croatian being a considerably more morphologically rich language than English, techniques such as morph segmentation should be useful to improve the performance of MT for this language pair.

Bibliography

Banerjee, S. and A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics. Ann Arbor, Michigan, USA. 65-72.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. Proceedings of EACL-2006: 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy. 249-256.

Denkowski, M. and A. Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages. Proceedings of the ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR. Uppsala, Sweden. 339-342.

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A method for automatic evaluation of Machine Translation. In 40th Annual Meeting of the Association of Computational Linguistics. Philadelphia, PA, USA. 311-318.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of the Conference of the Association for Machine Translation in the Americas Conference. 223-231.

Appendix A

This appendix contains the detailed results of the MT evaluation. We show these results in two tables, covering the results for English to Croatian and Croatian to English, respectively. For each system we report its scores according to the metrics considered in the evaluation (BLEU, TER, METEOR and OOV) for each of the test sets (SETimes, Tatoeba and WMT13).

	SETimes				Tatoeba				WMT13			
System	**BLEU**	**TER**	**MET**	**OOV**	**BLEU**	**TER**	**MET**	**OOV**	**BLEU**	**TER**	**MET**	**OOV**
EN-HR1	0.3345	0.5766	0.2547	0.70%	0.1918	0.7057	0.1887	0.60%	0.1848	0.6885	0.1876	1.60%
EN-HR2	**0.3431**	0.5675	**0.2593**	0.70%	0.2591	0.6283	0.2285	0.60%	0.2242	0.6508	**0.2093**	1.60%
EN-SL-HR-1*	0.1321	0.7955	0.1346	2.60%	0.1016	0.8135	0.1245	0.80%	0.0967	0.8026	0.1312	2.00%
EN-SL-HR-o*	0.1728	0.6691	0.2007	2.60%	0.2299	0.8481	0.1458	0.80%	0.1510	0.9338	0.1497	2.00%
EN-HR2-bf	0.3429	**0.5668**	0.2591	0.70%	0.2630	**0.6247**	**0.2293**	0.50%	**0.2248**	0.6508	0.2087	1.50%
EN-HR2-b	0.3416	0.5669	0.2581	**0.60%**	**0.2655**	0.6250	0.2275	**0.40%**	0.2229	**0.6494**	0.2070	**1.20%**
Google	0.3657	0.5340	0.2727	NA	0.2955	0.5879	0.2555	NA	0.2570	0.6120	0.2284	NA

Table 1: Machine translation results for English to Croatian systems. For each dataset and metric we show in bold our best score.

**The OOV value reported for these systems is the percentage of tokens in the source side of the test set that are not present in the source side of the training data (English – Slovene). Some tokens that are covered in by the English → Slovene SMT system may become OOVs in the second stage system (Slovene → Croatian RBMT system).*

	SETimes				Tatoeba				WMT13			
System	**BLEU**	**TER**	**MET**	**OOV**	**BLEU**	**TER**	**MET**	**OOV**	**BLEU**	**TER**	**MET**	**OOV**
HR-EN1	0.4035	0.5059	0.3373	1.40%	0.3379	0.5453	0.2756	4.00%	0.3324	0.5301	0.2991	4.50%
HR-EN2	0.4052	**0.5026**	0.3424	1.30%	**0.3486**	0.5328	**0.2823**	2.80%	0.3364	**0.5274**	0.3006	3.80%
HR-EN2-bf	**0.4053**	0.5030	0.3423	1.30%	0.3462	0.5330	0.2814	2.70%	0.3365	0.5289	0.3002	3.70%
HR-EN2-b	0.4052	**0.5026**	**0.3428**	**1.10%**	0.3474	**0.5323**	0.2811	**2.40%**	**0.3375**	0.5289	**0.3013**	**3.30%**
Google	0.4578	0.4489	0.3723	NA	0.4569	0.4223	0.3449	NA	0.3994	0.4706	0.3391	NA

Table 2: Machine translation results for Croatian to English systems. For each dataset and metric we show in bold our best score.