АБУМАТРАН
ABUMATRAN

Abu-MaTran

Automatic building of Machine Translation

PIAP- GA-2012-324414

D5.1b Evaluation of the MT systems deployed in the second development cycle

Dissemination level	Public
Delivery date	2014/12/31
Status and version	Final, v1.0
Authors and affiliation	Mikel L. Forcada (UA), Tommi Pirinen (DCU), Raphaël Rubino (Prompsit), Antonio Toral (DCU)

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

Table of Contents

Executive Summary

This deliverable reports on the evaluation of the MT systems described in deliverable D4.1b (Forcada et al. 2014), which have been developed between the first milestone (month 7) and the second milestone (month 24) of project Abu-MaTran. Substantial improvements over the results of milestone 1 are observed for generic English→Croatian and Croatian→English systems. Results are also shown for systems built for a specific-domain system (tourism): these systems clearly outperform their generic counterparts as well as general-purpose systems like Google Translate. The use of morph segmentation for Croatian does not seem to improve performance according to the usual automatic measures, but shows clear potential as it drastically lowers the out-of-vocabulary rate.

The deliverable also reports on the the results of our submission to the WMT14 shared task, where one of our systems was ranked first in terms of manual evaluation, and on the clear benefit of corpus cleaning when training a system on the OpenSubtitles corpus in both translation directions (English to Croatian and viceversa), as shown by an increase of +10 absolute BLEU points.

1 Introduction

This deliverable covers the evaluation of the machine translation (MT) systems built during the second development cycle of the project (as described in Deliverable 4.1b, Forcada et al. 2014a).

Each of the next sections cover the evaluation of each of the systems developed in this cycle. For ease of reading, they follow the same order as they were presented in Deliverable 4.1b (e.g. Section 2.1 below covers the evaluation of the MT system described in Section 2.1 of Deliverable 4.1b).

Unless noted otherwise, each system is evaluated with three state-of-the-art evaluation metrics: BLEU, METEOR and TER. These metrics were introduced in Deliverable 5.1a (Toral et al. 2013b, Section 2).

2 Milestone 2 MT Systems

2.1 General MT System

Appendix A in Deliverable D5.1a (Toral et al. 2013a) shows the results obtained by our best Milestone 1 MT systems, as well as Google, using the three test sets described in Deliverable D5.1a Section 3, namely SETimes, Tatoeba and WMT13. From these three test sets, we decide to focus the Milestone 2 evaluation on the latter one: WMT13 test set, described in more details in Deliverable D3.1a (Toral et al. 2013a). The selection of this particular test set amongst the three initial ones is explained in Deliverable D4.1b (Forcada et al. 2014a).

The results reported for Milestone 1 MT systems on the WMT13 test set do not show a consistent leading system compared to the others. We decide to compare Milestone 2 systems to the best direct translator of Milestone 1 (*Milestone 1 SMT*), but also to the best results regardless of which MT system obtains it (*Milestone 1 Best Scores*) and to two third-party on-line MT systems[1] (Google Translate[2] and Yandex).[3] We report the results obtained by our Milestone 2 SMT systems in Table 1 for the English to Croatian translation direction and in Table 2 for Croatian to English direction.

System	BLEU	TER	METEOR
DGT Translation Memory	0.1607	0.7279	0.1690
hrenWaC	0.2015	0.6689	0.1909
JRC Acquis	0.1736	0.7237	0.1728
Open Subtitles 2013	0.2245	0.6370	0.2053
SETimes	0.2232	0.6559	0.2055
TED Talks	0.2010	0.6848	0.1922
Milestone 1 SMT	0.2248	0.6508	0.2087
Milestone 1 Best Scores	0.2248	0.6494	0.2093
Yandex	0.198	0.690	0.196
Google	0.268	0.603	0.234
Concatenation - no Open Subtitles 2013	**0.2449**	0.6228	**0.2165**
Concatenation	0.2408	**0.6171**	0.2152

Table 1: Scores for Milestone 2 MT systems (English to Croatian)

The translation results from English to Croatian show that our best system is the one trained on the concatenation of all the parallel training data without Open Subtitles 2013. This system is

[1] Translations of the test set were obtained with these third-party systems as of August 22th 2014.
[2] http://translate.google.com
[3] https://translate.yandex.com/

outperformed by Google by +2.2 absolute BLEU points.[4] We notice a +1 absolute BLEU point improvement of the Google system on the same test set compared to the same evaluation conducted for the first milestone of the project described in Deliverable 5.1a. The current Milestone 2 SMT system outperforms our best Milestone 1 system by +2 absolute BLEU points. We also outperform the online Yandex translator by +4 absolute BLEU points.

System	BLEU	TER	METEOR
DGT Translation Memory	0.2240	0.6428	0.2267
hrenWaC	0.2919	0.5737	0.2675
JRC Acquis	0.2517	0.6157	0.2463
Open Subtitles 2013	0.3082	0.5466	0.2774
SETimes	0.3134	0.5510	0.2830
TED Talks	0.2642	0.5997	0.2531
Milestone 1 SMT	0.3375	0.5289	0.3013
Milestone 1 Best Scores	0.3375	0.5274	0.3013
Yandex	0.330	0.546	0.293
Google	0.405	0.465	0.343
Concatenation - no Open Subtitles 2013	**0.3543**	0.5172	**0.3086**
Concatenation	0.3463	**0.5150**	0.3013

Table 2: Scores for Milestone 2 MT systems (Croatian to English)

The translation results for the Croatian to English direction show the same trend observed for the other direction. Our best system is the concatenation except Open Subtitles 2013, outperformed by Google. However, the best TER scores are obtained by the system trained on the concatenation of all the data, for both translation directions. Our Milestone 1 system results are improved by nearly +2 absolute BLEU points. The results are consistent for the two translation directions. The difference in terms of BLEU score between our best Milestone 2 SMT system and the Yandex online translator is smaller for this translation direction (+2 absolute BLEU points) compared to the English to Croatian direction.

The results obtained with parallel corpora used to train individual SMT systems show that SETimes and Open Subtitles 2013 lead to the best results compared to the other corpora. Open Subtitles 2013 leads to the lowest TER scores, a word-level evaluation metric, indicating a better coverage of the test set vocabulary due mainly to the size of the corpus (more than 17,000,000 sentence pairs). On the other hand, SETimes contains more than 200,000 sentence pairs and leads to the best BLEU scores when translating the test set from Croatian to English, and second best after Open Subtitles 2013 on the other translation direction. This hints to a relative proximity between SETimes and the test set compared to the other corpora.

A comparison of the results obtained with individual parallel corpora indicates which of these corpora is bringing more useful information according to our current test set. For instance, the DGT Translation Memory contains more sentence pairs and tokens than the SETimes corpus, but the use of the latter one as training data leads to better results according to the automatic metrics. The difference between the two systems is large, +6 absolute BLEU points from English to Croatian and +6 absolute BLEU points from Croatian to English. This indicates possible noise in the DGT Translation Memory, but also possible thematic disparity between this training corpus and the test set.

This difference between corpora is even more significant when we compare the results obtained with SETimes with those obtained with the JRC Acquis corpus. The amount of sentence pairs in the

[4] A "BLEU point" is 0.01 BLEU.

latter corpus is three times higher and leads to a drop of at least -5 absolute BLEU points compared to the former corpus. We assume that noisy or out-of-topic elements are present in the corpora leading to low scores according to automatic metrics when used individually to train SMT systems. However, a deeper analysis of each training corpus is required to measure the proximity with the test set. This phenomenon may then also be present when concatenating all the parallel corpora to train an SMT system. Thus, we assume that a filtering process would allow us to improve over the best results reported in this deliverable.

2.2 Tourism MT System

In this section we report on the results obtained for the tourism domain. Section 2.2. of Deliverable 4.1b (Forcada et al., 2014) provides details on the MT systems built for this task, while Section 3.1 in Deliverable 3.1b (Esplà-Gomis et al., 2014) provides the context of this task as well as details regarding the data sources acquired.

We evaluated each MT system again according to three evaluation metrics (BLEU, TER, METEOR) as well as according to the percentage of out-of-vocabulary words (OOV). MT systems built solely on specific data outperform the generic baseline even if the specific datasets are much smaller in size compared to the generic datasets. Finally, the best result, both in terms of evaluation metrics and coverage (OOVs) was obtained when we combined the best performing specific datasets and the generic dataset through linear interpolation.

This system leads to a 16.3pts BLEU improvement over the generic SMT system, 13.4pts BLEU improvement over the results obtained with the publicly available Google online translator (as measured in March 2014) and 4.2pts BLEU improvement over the best domain-specific system trained using the data crawled with the best crawler and the best configuration. More details about the evaluation and the differences between the results are given in Toral et al., (2014).

3 Morph Segmentation

In this section we report on preliminary results regarding experimentation for morphologically motivated modelling of MT. We evaluated the system using automatically generated morph segmentations by two systems: *Morfessor* (Morfessor 2.0 Baseline) and *Flatcat* (Morfessor Flatcat). The measurements are made using the held-off part of SETimes (i.e. in-domain) and WMT 2013 (out of domain) test set described in Deliverable D3.1a. The results are shown in Table 3 for English to Croatian translation and Table 4 for Croatian to English translation. The tables show results of BLEU, TER and METEOR tests as measured using the packages mteval13a.pl, tercom-0.7.25.jar, meteor-1.5.jar and oov.pl that come with the Moses decoder, version 2.1.1.

English to Croatian								
	SETimes				**WMT'13**			
systems	**BLEU**	**TER**	**MET**	**OOV**	**BLEU**	**TER**	**MET**	**OOV**
Moses	0.3370	0.5789	0.2563	1.00%	0.1575	0.7271	0.1732	3.10%
Morfessor	0.3095	0.6345	0.2367	1.00%	0.1434	0.7665	0.1586	3.10%
Flatcat	0.2891	0.6623	0.2395	1.00%	0.1367	0.7905	0.1625	3.10%

Table 3: Translation quality translating from English into Croatian using word-forms (Moses baseline) and two different unsupervised morph segmentation schemes (Morfessor 2.0 baseline and Morfessor Flatcat). The values for BLEU, TER and METEOR as given by the scripts rounded to four decimal points and OOV in percent units. Note that out-of-vocabulary rates are not affected by the use of morph segmentation, because it is applied to the target language.

Croatian to English								
	SETimes				WMT'13			
systems	BLEU	TER	MET	OOV	BLEU	TER	MET	OOV
Moses	0.3971	0.5191	0.4056	2.10%	0.2577	0.6112	0.3232	7.30%
Morfessor	0.3816	0.5444	0.4034	0.50%	0.2181	0.7117	0.3124	0.40%
Flatcat	0.3872	0.5366	0.3978	1.60%	0.2350	0.6560	0.3088	2.20%

Table 4: Translation quality translating from Croatian into English using word-forms (Moses baseline) and two different unsupervised morph segmentation schemes (Morfessor 2.0 baseline and Morfessor Flatcat). The values for BLEU, TER and METEOR as given by the scripts rounded to four decimal points and OOV in percent units.

In the preliminary results, the evaluation scores (BLEU, TER, METEOR) are slightly worse than with word-form-based using the baseline Moses setup, which is in line with other language experiments that have been documented using morfessor for SMT (e.g. Virpioja et al. (2007), Fishel (2010)). As the authors of the mentioned articles, we believe that this does not directly indicate a failure of the method but it requires further analysis of the MT outputs to get an insight of the strengths and weaknesses of morph-segmented SMT. Such analysis will be conducted in Year 3.

The promising point in the results is the increase in coverage (decrease in out-of-vocabulary rate), shown in the OOV column, for the Croatian to English task, as this is one of the main potential sources of improvement for morph-based model. Assuming that any of the newly acquired words as combination of morphs is translated correctly, they will naturally contribute positively to the translation quality.

Other factors contributing to the low evaluation score of the morph-based system at the moment is some of the problems with interactions between pre-processing by the MT pipeline versus pre-processing by the morph-segmenting pipeline and the failures in the morph de-segmenting in post-processing, both of which we aim to improve upon for Milestone 3.

4 Participation in the WMT14 Shared Task

This section reports on the results obtained by our systems in the WMT14 shared task (cf. Section 4 in Deliverable 4.1b, Forcada et al. 2014, for a description of these systems). Our submitted test sets were automatically and manually evaluated following the shared task guidelines (please see Bojar et al., 2014) for more details). We report the results obtained by the Abu-MaTran participants during the WMT14 translation task in Table 5. Our participation in the French to English translation direction was limited to the automatic evaluation. For both translation directions, we entered the constrained part of the task.[5]

According to the manual evaluation, our system obtained the top rank for constrained systems (in a draw with two other submissions, by Edinburgh and Karlsruhe). According to the automatic evaluation metrics used in the shared task (BLEU and TER), our systems ranked second into French and fourth into English.

[5] The constrained systems are restricted to the data provided by the shared task organisers.

	Manual			Automatic		
	Rank	Range	Score	Rank	BLEU	TER
English→French	1	2 - 5	0.185	2	0.349	0.547
French→English	NA	NA	NA	4	0.340	0.531

Table 5: Results obtained by the Abu-MaTran SMT systems during the WMT14 shared translation task on the French–English language pair in both translation directions.

5 OpenSubs Cleaning

We present in Section 5 of Deliverable D4.1b (Forcada et al. 2014) a cleaning process allowing us to benefit from noisy parallel corpora, as well as the experimental setup in order to evaluate the impact of our cleaning procedure on MT performance. We train two SMT systems per language direction, one with an uncleaned (original) version of the OpenSubtitles 2013 parallel corpus, and one with a cleaned version of this corpus using our cleaning process. Results for the Croatian to English direction are reported in Table 6, while results for the English to Croatian direction are presented in Table 7.

	BLEU	TER	METEOR
Open Subtitles 2013 - No Cleaning	0.2249	0.6541	0.2196
Open Subtitles 2013 - Cleaning	0.3082	0.5466	0.2774

Table 6: Croatian to English

	BLEU	TER	METEOR
Open Subtitles 2013 - No Cleaning	0.0915	0.8305	0.1125
Open Subtitles 2013 - Cleaning	0.2245	0.6370	0.2053

Table 7: English to Croatian

A strong impact of the cleaning process on the performances of the SMT systems trained with the Open Subtitles 2013 parallel corpus is observed for both translation directions. More than +9 absolute BLEU points are gained when cleaning the corpus and translating from Croatian to English. For the other direction, the cleaning procedure leads to a gain of +13 absolute BLEU points. These results are consistent over the three automatic metrics, showing a larger gain for the English to Croatian direction compared to the other translation direction.

6 Conclusions

This deliverable has reported the evaluation of the MT systems developed during the period of the second milestone of the project (M7–M24).

Our generic MT system for milestone 2 outperforms the generic system we built for milestone 1 for both translation directions (English to Croatian and viceversa). In both cases, the improvement is substantial at around 2 absolute BLEU points. As for the MT system for a specific domain (tourism) we have developed, it outperforms a generic system as well as Google Translate by a wide margin. We have then evaluated our experimental system that uses morph segmentation. While the results are slightly lower than the baseline in terms of evaluation metrics, this system also bring some benefits as it results in higher coverage.

Apart from these, we have also reported on the results of our submission to the WMT14 shared task, where we had a top ranked system in terms of manual evaluation. Finally, we have reported scores on systems built with clean and unclean versions of OpenSubtitles, in order to have an

extrinsic evaluation of our cleaning procedure. For both translation directions (English to Croatian and viceversa), a MT system built with the clean data outperforms a system built with the unclean data by around +10 absolute BLEU points.

Bibliography

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., Tamchyna, A. (2014) "Findings of the 2014 Workshop on Statistical Machine Translation", in Proceedings of the 9th Workshop on Statistical Machine Translation (WMT), pp. 12–58.

Esplà-Gomis, M., Forcada, M.L., Ljubešić, N., Papavassiliou, V., Prokopidis, P. Ortiz-Rojas, S., Rubino, R., Sánchez-Cartagena, V.M., Toral, A. (2014) "Deliverable D3.1b: Acquisition for cycle 2", version 1.0, available from http://www.abumatran.eu/?page_id=59

Forcada, M.L., Pirinen, T., Rubino, R., Toral. A. (2014) "Deliverable D4.1b: MT systems for the second development cycle", version 1.0, available from http://www.abumatran.eu/?page_id=59

Fishel, M., and Harri Kirik. "Linguistically Motivated Unsupervised Segmentation for Machine Translation" in Proceedings of LREC. 2010, p. 1741–1745.

Rubino, R., Toral, A., Sánchez-Cartagena, V.M., Ferrández-Tordera, J., Ortiz-Rojas, S. Ramírez-Sánchez, G., Sánchez-Martínez, F., Way, A. (2014), "Abu-MaTran at WMT 2014 Translation Task: Two-step Data Selection and RBMT-Style Synthetic Rules". In Proceedings of the 9th Workshop on Statistical Machine Translation (WMT), pp. 171–177.

Toral, A. Rubino, R., Esplà-Gomis, M., Pirinen, T., Way, A., Ramírez-Sánchez, G. (2014), "Extrinsic Evaluation of Web-Crawlers in Machine Translation: a Case Study on Croatian–English for the Tourism Domain". In Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT), pp. 221–224.

Toral, A., Cortés-Vaíllo, S., Ramírez-Sánchez, G., Forcada, M.L., Ljubešić, N. (2013a) "Deliverable D3.1a: Acquisition for the first development cycle", version 1.0, available from http://www.abumatran.eu/?page_id=59

Toral, A., Cortés-Vaíllo, S., Ramírez-Sánchez, G., Klubička, F., Ljubešić, N. (2013b) "Deliverable D5.1a: Evaluation of the MT systems deployed in the first development cycle", version 1.0, available from http://www.abumatran.eu/?page_id=59

Virpioja, S., Väyrynen, J. J., Creutz, M., and Sadeniemi, M. (2007). Morphology-Aware Statistical Machine Translation Based on Morphs Induced in an Unsupervised Manner. In Proceedings of Machine Translation Summit XI, Copenhagen, Denmark, 10-14 September, 2007, pp. 491-498.