АБУМАТРАН
ABUMATRAN

Abu-MaTran

Automatic Building of Machine Translation

PIAP- GA-2012-324414

D5.1d Evaluation of the MT systems deployed in the fourth development cycle

Dissemination level	Public
Delivery date	2016/12/31
Status and version	Final, v1.0
Authors and affiliation	Filip Klubička (UZ), Víctor Sánchez-Cartagena (Prompsit) and Antonio Toral (DCU)

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

Contents

Executive Summary

This deliverable reports on the evaluation of the MT systems described in deliverable D4.1d, which have been developed between the third milestone (month 36) and the fourth milestone (month 48) of the project. Substantial improvements over the results of milestone 3 are observed. Furthermore, in an error analysis conducted we have gained insights regarding which error types are reduced with the use of linguistic information in statistical MT.

1 Introduction

This deliverable covers the evaluation of the machine translation (MT) systems built during the fourth development cycle of the project (as described in Deliverable D4.1d, (Sánchez-Cartagena et al., 2017)). The automatic evaluation of these systems is covered in section 2. We also describe work done this year on sentence-level ranking (manual evaluation), in Section 3.

While in previous milestones our systems were evaluated by means of overall metrics, either automatic (e.g. BLEU) or manual (e.g. human rankings), in this milestone we go beyond overall system evaluation by carrying out an error analysis of our current best performing statistical system. This is found in section 4.

Finally, due to the emergence of the new neural MT approach, we have conducted a multifaceted analysis of statistical versus neural MT (NMT, cf. Section 5).

2 Automatic evaluation

In this section we report the scores obtained in terms of the BLEU and TER automatic evaluation metrics for the 2 systems developed during the final milestone for our main language direction: English-to-Croatian. These regard the use of factored models in classic statistical MT and the use of the recently introduced neural MT approach. These methods are described in detail in sections 4.1. and 3 of Sánchez-Cartagena et al. (2017), respectively.

Table 1 shows the results of these 2 systems as well as those obtained by the best system that we built during the previous milestone.

System	**BLEU**	**TER**
Factored SMT	0.2700	0.5963
Neural MT	0.3085	0.5552
Milestone 3	0.2544	0.6081

Table 1: Automatic evaluation of the milestone 4 systems

The use of factored models leads to a substantial improvement upon the system built in the previous milestone (6% relative in terms of BLEU). Neural MT on its turn, allows us to obtain a further notable improvement (14%

relative in terms of BLEU compared to the SMT system that uses factored models and 21% compared to the milestone 3 system).

3 Sentence-level rankings

Milestone 3 systems were ranked at sentence level by one annotator as described in Section 3 of Deliverable 5.1c (Forcada et al., 2016). During the current milestone a second annotator ranked these systems. These allowed us to calculate the inter annotator agreement and to have a bigger set of rankings which was useful to establish whether the differences between systems are significant. Details on this can be found in a paper published this year (Toral et al., 2016).

4 Error analysis

In this section we report on the motivation for conducting manual error analysis, as well as describe the framework and overall annotation process, and present the results. This analysis is carried out for the language direction English-to-Croatian.

The fact that Croatian is rich in inflection and has rather free word order, as well as other similar phenomena that English does not, gives rise to specific translation issues. For example, grammatical categories that do not exist in English, like gender or case, may be particularly hard to generate reliably in Croatian. One of the tasks in the project was to take this into account and build a hybrid English–Croatian machine translation system using factored models (for more details, refer to deliverable D4.1d Sánchez-Cartagena et al. (2017), and, for more detail, to Sánchez-Cartagena et al. (2016)), which we expected would address such issues.

Indeed, once the hybrid system was evaluated using automatic metrics, the results did show some improvement, as shown in Section 2. However, the improvement of the factored SMT system is only 6% relative in terms of BLEU, which is certainly significant, but not as much as we might have hoped. Furthermore, and much more importantly, as is the nature of automatic metrics, this scoring method does not indicate whether any of the linguistic problems mentioned earlier have been addressed by employing the factored model. The question of whether the linguistic quality, or rather,

grammaticality of the output has improved has not been answered. Are cases and gender handled better? Is there better agreement? Is the fluency of the translation higher? We thus decide to thoroughly compare the two systems by systematically analyzing their outputs via manual error analysis. In this way we can obtain a more complete idea of what is happening in the translation, which can provide pointers on where to act to obtain further improvements.

4.1 Multidimensional Quality Metrics

After looking into different ways of performing this task, we decided to make use of the Multidimensional Quality Metrics (MQM) developed in the QT-Launchpad project.[1] This is a framework for describing and defining custom translation quality metrics. It provides a flexible vocabulary of quality issue types and a mechanism for applying them to generate quality scores. It does not impose a single metric for all uses, but rather provides a comprehensive catalog of quality issue types, with standardized names and definitions, that can be used to describe particular metrics for specific tasks.

The main reason we chose the MQM framework was the flexibility of the issue types and their granularity — it gave us a reliable methodology for quality assessment, that still allowed us to pick and choose which error tags we wish to use.

In order to carry out the annotations we used translate5,[2] a web-based tool that implements annotations of MT outputs using hierarchical taxonomies, as is the case of MQM.

4.2 Tagset

The MQM guidelines propose a great variety of tags on several annotation layers. However, the full tagset is too comprehensive to be viable for any annotation task, so the process begins with choosing the tags to use in accordance to our research questions. Initially we started off with the core tagset, a default set of evaluation metrics (i.e. error categories) proposed by the MQM guidelines, as seen in Figure 1.

However, given the morphological complexity of Croatian and the level at which we made interventions in the system, we found that these core

[1] `http://www.qt21.eu/mqm-definition/definition-2015-06-16.html`

[2] `http://www.translate5.net/`

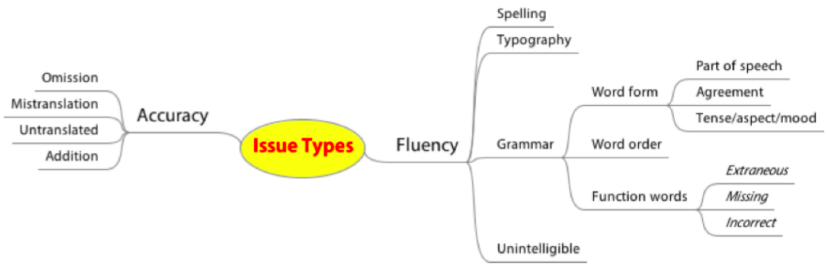

Figure 1: The core error categories proposed by the MQM guidelines

categories were not detailed enough, or rather, did not allow for an analysis of the specific phenomena we were interested in. Some categories that were of interest to us, like specific *Agreement* types, were not present in the tagset, while some errors, like *Typography*, were irrelevant to us. So we created our own set of tags by modifying the core set, rearranging the hierarchy, adding new tags and removing ones that are of little consequence. We call this new tagset the Slavic Tagset, as its expansion allows for identification of grammatical errors which are commonly shared by (South-)Slavic languages. This tagset is outlined in Figure 2.

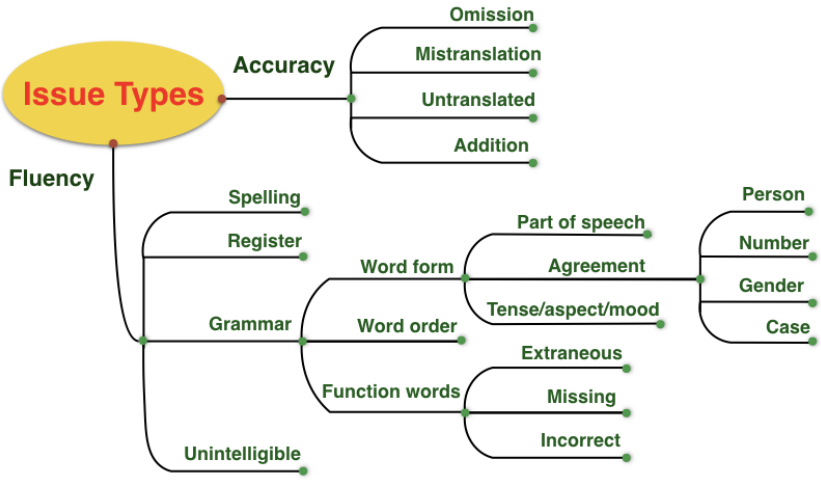

Figure 2: The Slavic tagset, a modified version of the MQM core tagset

4.3 Annotation setup

We had 2 annotators familiar with the MQM framework annotate the data, which consisted of 100 random sentences from a test set developed earlier in the project (manually translated sentences from the news domain, refer to Deliverable D5.1a (Toral et al., 2013) for more details). These sentences were translated by both MT systems (statistical MT with and without factored models), and both translations were then annotated by both our annotators (i.e. each system translated the same 100 sentences, each annotator annotated the 200 translated sentences, making a total of 400 annotated sentences.)

Once the sentences were annotated, the annotation data was extracted, we calculated inter-annotator agreement and analyzed the output to see what the number of error tags can tell us about the performance of each system.

4.4 Inter-Annotator Agreement

Though well thought out and developed, the MQM metrics, and more broadly MT evaluation in general, are notorious for resulting in low inter-annotator agreement scores. This is attested by the body of work that has addressed this issue, most notably Lommel et al. (2014), who worked specifically on MQM, and Callison-Burch et al. (2007), who investigated several tasks. This is why it is important that we also check to what extent our annotators agree on their annotation and whether this is consistent with other work done with MQM so far.

Prior to annotation, both annotators have been thoroughly familiarized themselves with the official annotation guidelines and the decision process[3].

Once the data was annotated, observed agreement was approximated on the level of sentence, and inter-annotator agreement was calculated using Cohen's Kappa (κ) metric (Cohen, 1960).

Agreement was calculated on the annotations of every system separately, as well as on a concatenation of annotations, in order to both see whether there are differences in agreement across systems, as well as to get an idea of overall agreement between annotators. Additionally, Coehn's κ was also calculated for every error type separately. Detailed results can be found in Table 2.

Generally, one can see that our annotators agree better on evaluations of the Milestone 3 system than on the Factored SMT system, and that the over-

[3] `http://www.qt21.eu/downloads/annotatorsGuidelines-2014-06-11.pdf`

Error	**Milestone 3**	**Factored SMT**	**Concatenated**
Mistranslation	0.51	0.48	0.50
Omission	0.34	0.39	0.36
Addition	0.50	0.54	0.52
Untranslated	0.86	0.86	0.86
Unintelligible	0.39	0.32	0.35
Register	0.37	0.20	0.29
Word order	0.56	0.33	0.46
Extraneous	0.56	0.32	0.44
Incorrect	0.37	0.18	0.27
Missing	0.00	0.49	0.40
Tense/aspect/mood	0.44	0.36	0.40
Agreement	0.24	0.41	0.32
Number	0.53	0.55	0.54
Gender	0.46	0.59	0.53
Case	0.53	0.49	0.53
All errors	**0.56**	**0.49**	**0.53**

Table 2: Inter-annotator agreement (Cohen's κ values) for MQM evaluation task

all agreement scores are relatively low, the average total κ is being approximately 0.53, which, according to Cohen, constitutes moderate agreement. As already stated, this is to be expected, given the complexity of the problem and annotation schema. However, this is in fact a much higher score than what has been reported in similar work, most notably Lommel et al. (2014), who achieve much lower MQM annotation κ scores, ranging between 0.25 and 0.34. Though certainly a success on our part, this comparison should be taken with a grain of salt, as our calculations are just an approximation compared to Lommel et al.'s, as our setup was different insomuch that we looked at sentence level agreement, while they calculated agreement on the token level.

Furthermore, when looking at specific error type, the κ scores are relatively consistent across all error types, mostly ranging between 0.35 and 0.55. There is one obvious outlier, however — the Untranslated category. Agreement on this error is extremely high when compared to other error types. This makes a lot of sense, as untranslated text is quite an unambiguous and easily detectable phenomenon, so high agreement among annotators would be expected.

Finally, if there was any doubt that the annotators' judgements are reliable, further analysis of the results shows that their annotations point to the same broad conclusions when considered both separately and together. This is elaborated on in Subsection 4.5.

4.5 Results of annotation

Table 3 presents the sum of raw annotations for every error type, for each system and both annotators, as extracted directly from the translate5 system.

Just by looking at the table one can easily detect that both annotators have judged that the Milestone 3 system contains more errors than the factored system (317 and 264 errors in Milestone 3, while 276 and 199 errors in factored SMT). This trend is consistent across most error categories.

However, even though simply counting the errors can provide a rough idea of which system might perform better, we felt that this approach does not adequately represent our findings, as it does not allow a proper quantification of the quality of the outputs. Certainly, based on data from Table 3 we can claim that the hybrid system produces less errors, but given that the outputs are different, as is the number of tokens in each translation, we felt the need to normalize the data.

	M3 system		Factored system	
Error	**Anno_1**	**Anno_2**	**Anno_1**	**Anno_2**
Accuracy	103	125	92	93
Mistranslation	78	80	64	64
Omission	13	22	11	12
Addition	4	14	10	8
Untranslated	8	9	7	9
Fluency	214	139	184	106
Unintelligible	2	3	2	4
Register	13	6	12	4
Spelling	0	2	0	4
Grammar	197	128	170	94
Word Order	26	16	25	8
Function words	22	10	25	6
Extraneous	4	3	4	2
Incorrect	16	7	18	3
Missing	2	0	3	1
Word Form	149	102	119	80
Part of Speech	10	2	11	4
Tense...	30	23	27	17
Agreement	109	76	80	58
Number	17	12	12	10
Gender	15	9	18	12
Case	71	40	38	23
Person	0	0	0	0
Total error count	**317**	**264**	**276**	**199**

Table 3: Raw annotation data from Milestone 3 and Factored systems: number of errors for each error type

There is not much related work on this, as most papers simply count the number of errors and stop there. After some consideration, we decided on a token-level approach. Instead of counting just error tags produced by each annotator, we count the tokens that these errors are assigned to - tokens that do and tokens that do not have an error annotation. Once these numbers are divided by the total number of tokens, they give a more concrete idea of the ratio of tokens with and without errors. Again, the results of this show that the hybrid system has a smaller error ratio. This is further backed up by a chi-squared (χ^2) statistical significance test, which shows that the difference in the number of tokens with errors is statistically significant, with an average p value lower than 0.0001.

Furthermore, we also wanted to see which error types are the ones making a significant impact on this result. So we repeated these same measurements, but instead of performing them on all error types combined, they were performed for each specific error category. Where values were too small for Pearson's test to handle, Fisher's test for statistical significance was performed instead. The combined results of the calculations and transformations are presented in Table 4.

Several things can be concluded from this table. Firstly, when looking simply at the grand total of tokens with and without errors, the difference between the systems is very statistically significant, i.e. the factored system has significantly less errors than our milestone 3 system.

The separate analysis of specific error types that contribute to this score reveals that only some of the error categories are significantly different between the two systems, or rather, that the difference in errors is statistically significant when it comes to errors in the categories of general *Accuracy*, and more specifically *Mistranslation*, as well as errors in the categories of general *Fluency*, *Grammar*, *Word form* and *Agreement*, as well as most specifically, *Agreement in Case*.

The final point is most interesting, as this was one of our main questions at the beginning - is there a way to better handle agreement when translating to Croatian? We can now confidently say that our factored system produces significantly less agreement errors overall, and given the specific agreement types, the system handles agreement in case significantly better.

However, one should also note the effect size (ϕ coefficient) of these measurements, which is very small. This is consistent with the relatively small increase in BLEU score that was observed during automatic testing. Even so, the differences are undeniably statistically significant, and the factored

	M3		Factored				
Error	**No error**	**Error**	**No error**	**Error**	χ^2	p	ϕ
Accuracy	3467	369	3525	291	9.65	**0.0019**	0.0355
Mistranslation	3547	289	3586	230	6.87	**0.0088**	0.03
Omission	3801	35	3793	23	2.44	0.1183	0.0179
Addition	3814	22	3797	19	0.21	0.6468	0.0052
Untranslated	3813	23	3797	19	0.36	0.5485	0.0069
Fluency	3195	641	3298	518	14.64	**0.0001**	0.0437
Unintelligible	3790	46	3769	47	0.02	0.8875	0.0016
Register	3810	26	3794	22	0.31	0.5777	0.0064
Spelling	3833	3	3812	4	-	0.7257	-
Grammar	3270	566	3371	445	15.97	**<0.0001**	0.0457
Word order	3752	84	3752	64	2.65	0.1035	0.0186
Function words	3801	35	3780	36	0.02	0.8875	0.0016
Extraneous	3829	7	3810	6	0.07	0.7913	0.003
Incorrect	3810	26	3790	26	0	1	0
Missing	3834	2	3812	4	-	0.4518	-
Word form	3389	447	3471	345	14.06	**0.0002**	0.0429
Part of speech	3822	14	3800	16	0.14	0.7083	0.0043
Tense...	3775	61	3765	51	0.85	0.3566	0.0105
Agreement	3466	370	3540	276	14.41	**0.0001**	0.0434
Number	3778	58	3772	44	1.87	0.1715	0.0156
Gender	3788	48	3756	60	1.42	0.2334	0.0136
Case	3614	222	3694	122	29.89	**<0.0001**	0.0625
Total errors	2826	1010	3007	809	27.77	**<0.0001**	0.0602

Table 4: Processed annotation data from both annotators concatenated: total number of tokens with and without errors (for each system), with statistical significance test results (chi-squared (χ^2), p-value and effect size ϕ)

model has brought us a step closer to solving some of the specific issues of translating from English to Croatian. The resulting system produces language that is more fluent and more grammatical, which, among other things, is of help when it comes to the task of post-editing.

5 Neural versus Phrase-based MT Analysis

We have conducted a study (Toral and Sánchez-Cartagena, 2017) whose aim is to shed light on the strengths and weaknesses of the newly introduced neural MT (NMT) paradigm. To do so we compare the translations produced by state-of-the-art neural and phrase-based MT systems (also referred to as classic statistical MT systems) for 9 language directions across a number of dimensions. The main findings are the following:

1. Translations produced by NMT are considerably different from those produced by phrase-based systems. In addition, there is higher inter-system variability in NMT, i.e. outputs by pairs of NMT systems are more different between them than outputs by pairs of phrase-based systems.

2. NMT outputs are more fluent. We corroborate the results of the manual evaluation of fluency at WMT16, which was conducted only for language directions into English, and we show evidence that this finding is true also for directions out of English.

3. NMT systems do more reordering than pure phrase-based ones but less than hierarchical systems. However, NMT reorderings are better than those of both types of phrase-based systems.

4. NMT performs better in terms of inflection and reordering. We confirm that the findings of Bentivogli et al. (2016) apply to a wide range of language directions. Differences regarding lexical errors are negligible.

5. NMT performs rather poorly for long sentences.

6 Conclusions

This deliverable has reported the evaluation of the MT systems developed during the period of the fourth milestone of the project (M37–M48). Our

final system developed during this milestone improves substantialy over the best system built during the previous milestone (relative improvement of 21% in terms of the BLEU automatic evaluation metric).

In this milestone we have gone beyond overall system evaluation by conducting an in-detail error analysis. This study has brought insights as to which error types the use of linguistic information in statistical MT (by means of factored models) can reduce.

Finally, we have conducted a multifaceted evaluation that has contributed to unveil some of the main strengths and weaknesses of the newly introduced neural MT paradigm.

References

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Forcada, M. L., Rubino, R., Pirinen, T., and Toral, A. (2016). Abu-matran deliverable D5.1c: Evaluation for the third development cycle. Technical report.

Lommel, A. R., Popovic, M., and Burchardt, A. (2014). Assessing inter-annotator agreement for translation error annotation. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*.

Sánchez-Cartagena, V. M., Ljubešić, N., and Klubička, F. (2016). Dealing with data sparseness in SMT with factored models and morphological expansion: a case study on croatian. *Baltic Journal of Modern Computing*, 4(2):354.

Sánchez-Cartagena, V. M., Gomis, M. E., Ferrández-Tordera, J., Klubička, F., and Toral, A. (2017). Abu-matran deliverable d4.1d: Mt systems for the fourth development cycle. Technical report.

Toral, A., Cortés-Vaíllo, S., Ramírez-Sánchez, G., Klubička, F., and Ljubešić, N. (2013). Abu-matran deliverable d5.1a evaluation for the first development cycle. Technical report.

Toral, A., Rubino, R., and Ramírez-Sánchez, G. (2016). Re-assessing the impact of SMT techniques with human evaluation: a case study on English–Croatian. *Baltic Journal of Modern Computing*, 4(2):368–375.

Toral, A. and Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (accepted)*.