АБУМАТРАН
ABUMATRAN

Abu-MaTran

Automatic Building of Machine Translation

PIAP- GA-2012-324414

D4.1c MT systems for the third development cycle

Dissemination level	Public
Delivery date	2015/12/31
Status and version	Final, v1.0
Authors and affiliation	Tommi Pirinen (DCU), Raphael Rubino (Prompsit), 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 D4.1c describes work done in the area of machine translation (MT) development and deployment (work package 4) during the period of the third milestone of the project (from month 25 to month 36). As part of this third development cycle, we have extended our generic MT system by building upon the system we prepared for milestone 2; we have extended our use of linguistic knowledge, to approaches such as factored translation and syntax-based translation and improved the morphologically-based processing. We have also used resources from closely related languages to obtain additional data. The deliverable also reports on other related activities carried out within this work package, namely our participation in the WMT15 and TweetMT translation shared task, and in WMT15 quality estimation contest.

1 Introduction

Previous development cycles of the Abu-MaTran project involved the construction of SMT systems for general and specific domains. If the latter was shown to benefit tremendously from Web crawled data, the former could already be handled to a certain extent by publicly available parallel corpora, in addition to monolingual crawled data. For the SMT systems developped for this milestone, we focus on increasing the amount of monolingual and parallel crawled from the Internet, on combining in an optimal way the different corpora for both translation and language models, but also on contrasting the provenance of tuning sets, allowing us to optimise the weights of the SMT components to influence the output of the system. In addition, we explore different ways of augmenting the core SMT systems with linguistic information, namely using morph segmentation, factored models and syntax.

2 General SMT Systems

This section presents the corpora used to train the translation and language models, as well as the method used to combine multiple models trained on different resources or using different settings. We also present in the following subsections the crowdsourced and professional translations of the development set which are used to determine the impact of this particular resource and its origin in the translation process. Finally, a subsection summarises the tools and other resources used in our experiments.

2.1 Translation Models

Publicly available corpora between English and Croatian were already used during the second milestone of the Abu-MaTran project, namely the DGT Translation Memory [1], the JRC Acquis [2], OpenSubtitles 2013, SETimes and Ted talks. These corpora are described in detail in deliverable D4.1b (Forcada et al., 2015, section 2.1.2).

In this milestone we add to the aforementioned corpora an updated version of the HrEnWaC parallel corpus (cf. Section 3.1 of deliverable D3.1c),

[1] `https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory`

[2] `http://tinyurl.com/CroatianAcquis`

which is roughly 7 times the size of the previous version, used in the previous milestone (698,097 segment pairs vs. 99,001). In addition, we use the Serbian–English SrEnWaC corpus, where its Serbian side is machine-translated to Croatian (cf. Section 4).

Due to the high amount of data as well as the heterogeneity of sources (e.g. from translations of the European Commission to subtitles produced by amateur translators), we explored different ways to combine these corpora. Our baseline in this respect is the simple concatenation of all the corpora. Then we will explore different alternatives, namely (i) linear interpolation of phrase tables (building one phrase table per individual corpus) and cross-entropy (Moore and Lewis, 2010) based data selection using the development set.

In the systems developed for the current milestone, we also experiment with three recent statistical MT techniques, namely hierarchical reordering models (Galley and Manning, 2008), operation sequence models (OSM) (Durrani et al., 2011) and bilingual neural language models (BiNLM) (Devlin et al., 2014).

Another novelty in the systems built for the current milestone is that we use a development set with three reference translations (two crowdsourced and one professional, cf. Section 4 of deliverable D3.1c) for tuning the weights of the different components.

2.2 Language Models

Language models for the milestone 3 systems are built in the same way as in the previous milestone and use the following corpora:

- For Croatian. The updated hrWaC corpus, cf. deliverable D4.1b (Forcada et al., 2015, section 2.1.2). For some systems, the language model also includes the target-side of the parallel training data.
- For English. The language model used in our submission to the WMT15 shared task (cf. Section 5.1).

3 Linguistically-augmented SMT Systems

In deliverable D4.1b of the previous milestone we introduced the concept of unsupervised morphological segmentation (Forcada et al., 2015, section 3) to

the overall MT pipeline. In the current milestone, this is carried on further. First, we aim to empirically determine the optimal parameters for our target language: Croatian. In addition, we extend the morphological segmentation methodology by using a rule-based, linguist-written morphological analyser to induce linguistically motivated morph boundaries.

On top of continuing the work around morph segmentation, in the current milestone we experiment with two other approaches to build linguistically-augmented SMT systems: factored models and tree-based models.

The rest of the section is organised as follows: in section 3.1 we describe the new additions to morphological segmentation scheme we have experimented, in section 3.2 we describe our factored models and in section 3.3 we describe our syntax-augmented models.

3.1 Morph Segmentation

In the previous milestone we rolled out the first version of our segmented SMT models. In this version we have used additional segmentation methods by leveraging a rule-based morphological analyser into a segmentation model and tuning the parameters of existing unsupervised segmentation systems.

3.1.1 Unsupervised Morph Segmentation

In order to improve the quality of the morph segmentation system from the previous milestone, described in Deliverable D4.1b (Forcada et al., 2015), we have empirically tried to determine a better selection of hyper-parameters for the segmentation. We have chosen to tune the two parameters from Morfessor (Creutz et al., 2005) (the unsupervised segmenter we use in our experiments): the *frequency dampening* and the *perplexity threshold*. These are the two main parameters in the unsupervised setup, that require language- and corpus-size adjustments.

The differences between the automatic evaluation scores of MT systems built with different values for Morfessor parameters are not large, and therefore we have averaged the results of each experiment (i.e. pair of values) over three separate runs of the tuning algorithm, which is unstable in that it contains a randomised component, and determined the best individual parameter and system combinations from the averaged results.

In the statistical morph segmentation methods used for pre-processing the Croatian data we used two variants of Morfessor segmentation: Base-

line 2.0 (Virpioja et al., 2013)[3] and flatcat (Grönroos et al., 2014)[4]. The segmentation can be fine-tuned by transforming the frequencies used in the likelihood calculations. By default the absolute frequencies of word-forms in corpora are used, making common words and their substrings more prominent in the training. Using logarithms of frequencies (*log* in tables) or turning all frequencies into 1 (*ones* in tables) will discount the common word-forms. In flatcat training, the re-segmentation is done iteratively to minimise the *perplexity* of the morph segmentation, the variable perplexity threshold is used to control when the training continues iterations, and was set to 100 in our previous experiments, in this experiment we empirically find the best value from 10, 50 and 100. To measure the effects of these parameters, we performed the evaluation of machine translation quality as described in Deliverable D5.1c.

3.1.2 Rule-based Morph Segmentation

To obtain rule-based, linguistic morph segmentation models, we have used the morphological analyser produced by the Apertium project in collaboration with Abu-Matran.[5]

The analyser itself does not contain morph segment boundary information, so we have created a method to extract potential dictionary automatically from the structure of the language description. The method makes use of the fact that typical morphological descriptions are composed of item-and-arrangement style progression of morphs. In Apertium dictionaries, this is realised as XML-based lists of word-roots (`sections`) and morph sets arranged into paradigm definitions (`pardef`). We assume that in each paradigm definition, entry boundary (`e`) is also a morph boundary and create segmentation points there. The assumption is not totally accurate in all dictionaries, for example, the Serbo-Croatian macrolanguage is implemented by having word-internal variation between individual languages defined as paradigms. However, the resulting segmentation model is fit, without further processing, for the purpose of segmenting morphs for statistical MT. A script that was used in this experiment is included in github.[6]

[3] `http://github.com/aalto-speech/morfessor`

[4] `http://github.com/aalto-speech/flatcat`

[5] `https://svn.code.sf.net/p/apertium/svn/languages/apertium-hbs`

[6] `https://github.com/flammie/autostuff-moses-smt/blob/master/apertium-segment.bash`

3.2 Factored Models

Factored models have been built based on state-of-the-art statistical taggers for Croatian using a HMM-based Hunpos (Halácsy et al., 2007)[7] and a CRF-based CRF suite.[8] We use the language models by University of Zagreb[9] (Agić and Merkler, 2013). The rationale to experiment on two systems performing on the same task is again to find out the technology that results in the best performance for SMT; the expectation is that CRF, typically leading to higher accuracy for PoS tagging, should lead to higher quality translations too.

Both models produce one tag per word drawn from the multext-east tag set (Erjavec, 2004). The produced analyses are in various CONLL-style formats, one word-form per line, we use a series of simple scripts for the conversion, some of the scripts from the Moses code baseand further additions in our repositories.[10]

We train models based on the singular complex tags and the surface words as factors for Croatian, and the surface forms as the sole factor English. The Croatian language model for the part-of-speech factor is trained in the same way as other language models, using the HRWaC 2.0 data.

3.3 Hierarchical and syntax-enhanced models

We developed further experiments based on syntactical processing to assess the effect of syntax on the MT output.

The following Moses-based systems were tested: hierarchical (unsupervised syntax), string-to-tree and tree-to-string. The experiments with linguistic syntax were based on the Croatian dependency parser from Zagreb [11] and the minimum-spanning tree parser MSTparser[12] for syntax on the Croatian side, and the Berkeley parser [13] for syntax on the English side.

The conversion from MST dependency syntax to Moses' phrase structure syntax was performed using tools from the moses-MBOT project;[14] the

[7] http://code.google.com/p/hunpos/
[8] https://github.com/chokkan/crfsuite
[9] http://nlp.ffzg.hr/resources/models/tagging/
[10] https://github.com/flammie/autostuff-moses-smt/
[11] http://nlp.ffzg.hr/resources/models/tagging/
[12] https://sourceforge.net/projects/mstparser/
[13] https://github.com/slavpetrov/berkeleyparser
[14] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/

conversion consists of the following steps: MST to CoNLL-X, CoNLL-X to CoNLL-09 and CoNLL-09 to the XML format used in Moses. The formats produced by these tools are not exactly in the same conll format expected by the next tool in the pipeline—an issue caused by possible unclarity in some conll format specs—so each step contains additional cleanups: CoNLL-X format produced by MSTParser's converter script has additional fields for sentence length that MBOT's converter cannot parse and numbers of fields not relevant to dependency parsing do not match.

The conversion from Berkeley parser syntax to Moses was done by a script provided in the Moses repository.[15]

For training the models we used the default options in Moses with flags indicating that the system is hierarchical (`-hierarchical`) and either tree-to-string (`-source-syntax`) or string-to-tree (`-target-syntax`).

The syntax-based training with Moses allows additional processing to the syntax model to improve performance called relaxation (Wang et al., 2007).

4 Resources from Related Languages

One of the aims of the current milestone is on extending to related languages and using resources from related languages. In this regard, we take advantage of the Serbian–Croatian rule-based system based on Apertium that has been developed recently during the secondments of Prompsit at UZ,[16] to translate a monolingual corpus of the Serbian TLD (srWaC, cf. section 2 in deliverable D3.1c) from Serbian into Croatian and use it as an additional language model in statistical MT systems for the English-to-Croatian direction.

The procedure is as follows:

1. Translate srWaC into Croatian with Apertium and extract three versions of the MT output:
    - Whole output.
    - Only the subset of sentences without any unknown word.

`mbotmoses.en.html`

[15] `http://github.com/moses-smt/mosesdecoder/scripts/training/conversion/berkeley2moses.perl`

[16] `https://svn.code.sf.net/p/apertium/svn/staging/apertium-hbs_HR-hbs_SR/`

- Only the subset of sentences without any unknown word but discarding unknown words that regard capitalised words. We assume that most of them are proper nouns that would be written the same in both languages.

2. Build 3 LMs on those subsets and add each one at a time as an additional LM (log-linear combination) to a SMT system trained on hrenWaC (TM) and hrWaC (LM).

Taking into account that the hrWaC corpus is around double the size of srWaC and that the Apertium system is bound to make some mistakes, we do not expect the addition of srWaC to bring sizable improvements, but nonetheless it might lead to interesting findings. Because of this, the application of this approach in the reverse direction, i.e. translating hrWaC into Serbian and adding it as a LM for statistical MT for the English–Serbian direction, should bring further improvements. This is left as future work as the Apertium system does not support that translation direction at this time.

corpus	# sentences	# words
hrWaC	67,403,231	1,404,303,868
srWaC	25,636,554	557,727,450
srWaC Apertium (all)		557,038,636
srWaC Apertium (no unk)		186,878,197
srWaC Apertium (no unk lower)		302,829,433

Table 1: Quantitative figures of the srWaC corpus as is, its translations (all sentences, all except sentences with unkwnowns and all except sentences with unknowns that start with lowercase character) and hrWaC.

5 Participation in Shared Tasks

5.1 WMT15 Translation Task

Abu-MaTran took part in the WMT15 shared task for statistical machine translation between Finnish and English. There were two task setups: the first setup was based on using Europarl as parallel corpus and crawled news data as monolingual additional data, the translations for shared task were

from news domain (constrained task). The second task setup was the same, but allowed participants to augment the data e.g., with their own crawled data (unconstrained task). Abu-MaTran participated in three categories out of four in the Finnish–English pair: English-to-Finnish unconstrained and constrained, and Finnish-to-English constrained. The full description of the system is in the publication (Rubino et al., 2015). It uses the approaches and technologies described in this deliverable: various morph segmentation models as well as deliverable D3.1c (Esplà-Gomis et al., 2016): corpus acquisition. Our submission was the top-ranking system in English-to-Finnish both tasks second in English-to-Finnish unconstrained task using the BLEU cased metric. For total scoring including human evaluation, our system was top scoring constrained system, for both directions and in category of English-to-Finnish unconstrained, we came in third cluster or fourth absolute position (Bojar et al., 2015).

5.2 WMT15 Quality Estimation

Partner UA took part in the WMT15 shared task for MT quality estimation (QE), namely, in the task focused on MT QE at the word level. The methods developed build on the ideas developed by Esplà-Gomis et al. (2015a) for word-level QE in computer-aided translation tools based on translation memories. In this shared task, word-level MT QE is described as the task of detecting the words that need to be post-edided in the output of an MT system.

The approaches developed by Esplà-Gomis et al. (2015b) for word-level MT QE use a binary classification approach using a multilayer perceptron, and are able to exploit any source of bilingual information available for detecting possible editions. The authors define sources of bilingual information as any resource that allow to translate sub-segments. The method by Esplà-Gomis et al. (2015b) uses as sources of bilingual information two MT systems, Apertium and Google Translate, as weel as the on-line available bilingual concordancer Reverso Context, developed by partner Prompsit Language Engineering. Two systems were submitted to the WMT15 shared task (Esplà-Gomis et al., 2015), one using the features described by Esplà-Gomis et al. (2015b) and the other combining them with the baseline features provided by the organisation. The two submissions ranked third and first, respectively, among the 16 systems competing in this shared task (Bojar et al., 2015).

5.3 TweetMT 2015

Partner DCU took part in the TweetMT shared task (Alegria et al., 2015), consisting in the use of MT for translating tweets between a set of Iberian language pairs (Spanish from/to Catalan, Basque, Galician and Portuguese).

The systems we submitted to this competition (Toral et al., 2015) made extensive use of techniques developed in the project, namely crawling of tweets, morph segmentation and data selection. Systems were evaluated automatically and ours was the best of all systems submitted for five out of the six language directions in which we participated.

6 Conclusions and Future Work

This deliverable has covered the work done in the area of MT development and deployment (Work Package 4) during the period of the third milestone of the project (M25–M36). We have built new phrase-based statistical MT systems for English–Croatian by building upon the systems we prepared for milestone 2. We have then extended our systems by experimenting with the application of linguistic knowledge, including morph segmentation as well as syntax-based and hierarchical approaches.

Other related activities in this work-package have regarded our participation in WMT15 and TweetMT shared tasks.

Regarding future work for milestone 4, we consider mainly the following two lines of action:

1. Improving upon our current MT systems based on the errors found in a forthcoming diagnostic human evaluation.

2. Building neural MT systems.

References

Agić, Ž. and Merkler, D. (2013). Three syntactic formalisms for data-driven dependency parsing of Croatian. In *Text, Speech, and Dialogue*, pages 560–567. Springer.

Alegria, I., Aranberri, N., España-Bonet, C., Gamallo, P., Oliveira, H. G., Garcia, E. M., Vicente, I. S., Toral, A., and Zubiaga, A. (2015). Overview

of TweetMT: A Shared Task on Machine Translation of Tweets at SEPLN 2015. In *Proceedings of the Tweet Translation Workshop 2015 co-located with 31st Conference of the Spanish Society for Natural Language Processing (SEPLN 2015)*, pages 8–19.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Creutz, M., Lagus, K., Lindén, K., and Virpioja, S. (2005). Morfessor and hutmegs: Unsupervised morpheme segmentation for highly-inflecting and compounding languages.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J. (2014). Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *Proceedings of ACL*, pages 1370–1380.

Durrani, N., Schmid, H., and Fraser, A. (2011). A Joint Sequence Translation Model with Integrated Reordering. In *Proceedings of ACL/HLT*, pages 1045–1054.

Erjavec, T. (2004). Multext-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *LREC*.

Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M. (2015). Ualacant word-level machine translation quality estimation system at wmt 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 309–315, Lisbon, Portugal. Association for Computational Linguistics.

Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M. L. (2015a). Using machine translation to provide target-language edit hints in computer aided translation based on translation memories. *Journal of Artificial Intelligence Research*, 53:169–222.

Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M. L. (2015b). Using on-line available sources of bilingual information for word-level machine

translation quality estimation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 19–26, Antalya, Turkey.

Esplà-Gomis, M., Ljubešić, N., Ortiz-Rojas, S., Papavassiliou, V., Prokopidis, P., Sánchez-Cartagena, V., and Toral, A. (2016). Abu-matran deliverable d3.1c acquisition for the third development cycle. Technical report.

Forcada, M. L., Ortiz-Rojas, S., Pirinen, T., Rubino, R., and Toral, A. (2015). Abu-matran deliverable d4.1b mt systems for the second development cycle. Technical report.

Galley, M. and Manning, C. D. (2008). A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856. Association for Computational Linguistics.

Grönroos, S.-A., Virpioja, S., Smit, P., and Kurimo, M. (2014). Morfessor flatcat: An hmm-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). Hunpos: an open source trigram tagger. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 209–212. Association for Computational Linguistics.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rubino, R., Pirinen, T., Esplà-Gomis, M., Ljubešić, N., Ortiz Rojas, S., Papavassiliou, V., Prokopidis, P., and Toral, A. (2015). Abu-matran at wmt 2015 translation task: Morphological segmentation and web crawling. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 184–191, Lisbon, Portugal. Association for Computational Linguistics.

Toral, A., Wu, X., Pirinen, T., Qiu, Z., Bicici, E., and Du, J. (2015). Dublin City University at the TweetMT 2015 Shared Task. In *Proceedings of the Tweet Translation Workshop 2015 co-located with 31st Conference of the Spanish Society for Natural Language Processing (SEPLN 2015)*, pages 33–39.

Virpioja, S., Smit, P., Grönroos, S.-A., and Kurimo, M. (2013). Morfessor 2.0: Python implementation and extensions for morfessor baseline. Technical report. Published as `http://urn.fi/URN:ISBN:978-952-60-5501-5`.

Wang, W., Knight, K., and Marcu, D. (2007). Binarizing syntax trees to improve syntax-based machine translation accuracy. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 746–754, Prague, Czech Republic. Association for Computational Linguistics.