Abu-MaTran

Automatic Building of Machine Translation

PIAP- GA-2012-324414

D4.1d MT systems for the fourth development cycle

Dissemination level	Public
Delivery date	2016/12/31
Status and version	Final, v1.0
Authors and affiliation	Víctor Sánchez-Cartagena (Prompsit), Miquel Esplà-Gomis (UA), Jorge Ferrández-Tordera (Prompsit), Filip Klubička (UZ), 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 D4.1d describes work done in the area of machine translation (MT) development and deployment (work package 4) during the period of the fourth milestone of the project (from month 37 to month 48). As part of this fourth and last development cycle, we have worked on scalable language modelling techniques in order to be able to use the vast amounts of monolingual data crawled in the project. We have also built upon work done in the previous milestone concerning the use of linguistic approaches to MT. In this regard, we have investigated the use of factored models in phrase-based and neural MT and we have also investigated the pros and cons of the rule-based and unsupervised approaches to morphological segmentation. We have also worked on collaborative approaches to rapid development of MT for closely-related languages, focusing on the South-Slavic family, i.e. that of the target languages of the project. Another task has concerned the investigation of approaches for rapid adaptation to new domains, which has concerned our main target language (Croatian) but also Greek, as an exercise to demonstrate that the approaches developed are, to a large extent, language-independent. Finally, the deliverable also reports on two related activities carried out within this work package, namely our participation in the translation and quality estimation shared task at WMT16.

1 Introduction

Due to the fact that we have crawled large amounts of data in the project (specially monolingual), we explore how to use such data in an efficient manner. To this end we have developed a cloud-based architecture to build language models.

In the previous development cycle of the Abu-MaTran project there was a focus on augmenting our statistical machine translation (SMT) system with linguistic information. In this cycle we build upon that foundation; we study the best way to use factored models and we study the effects of the morphological segmentation approaches used in the previous milestone.

Due to the emergence of a new paradigm to automatic translation, neural MT (NMT), we build systems making use of this approach and compare them to our previous systems (e.g. as part of our participation in the 2016 edition of the WMT translation shared task). We also conduct novel research to add linguistic information to the NMT architecture.

Concerning rapid development of MT, we look at efficient ways to build MT engines between closely-related languages and also to bring MT to new domains.

2 Language Modelling

Language models (LMs) are an essential element in statistical approaches to natural language processing for tasks such as the one that is the focus of the project: machine translation (MT). The advent of big data leads to the availability of massive amounts of monolingual data that can be easily crawled (cf. Section 2.1 in Deliverable D3.1c (Esplà-Gomis et al., 2015)) and then used to build LMs; in fact, for the most prominent languages, it is not feasible using current techniques and hardware to train LMs with all the data available nowadays. At the same time, it has been shown that the more data is used for a LM the better the performance, e.g. for MT, without any indication yet of reaching a plateau.

Because of the reasons above, we developed CloudLM (Ferrández-Tordera et al., 2016),[1] an open-source cloud-based LM intended for MT, which allows to query distributed LMs. CloudLM relies on the Apache Solr[2] search plat-

[1]`https://github.com/jferrandez/mosesdecoder/tree/cache-cloudlm`
[2]`http://lucene.apache.org/solr/`

form and provides the functionality of state-of-the-art language modelling (it builds upon KenLM[3]), while allowing to query massive LMs (as the use of local memory is drastically reduced), at the expense of slower decoding speed. CloudLM was developed by project member Jorge Ferrández-Tordera from partner Prompsit during secondments at partner DCU.

3 Neural MT

Due to the recent emergence of the new neural approach to MT, during the last year of the project we have built a neural system for the main language direction that we have considered throughout the project, i.e. English-to-Croatian. This system is based on our experience with neural MT at the WMT news translation shared task (cf. Section 7.1).

The system is trained using the same parallel data as in our best statistical MT system built in the previous milestone. This is made up of the top 25% sentence pairs (cf. Section 2.6. of Deliverable D5.1c Pirinen et al. (2015a)) of all the parallel corpora available (cf. Section 2.1 of Deliverable D4.1c Pirinen et al. (2015b)).

We have used the sequence-to-sequence architecture to neural MT with attention and we have applied byte pair encoding (Sennrich et al., 2015) jointly on the source and target languages. This consists of initially segmenting each word in characters, and iteratively joining the most frequent pair of segments in the training corpus. We performed 85 000 join operations. Training is performed for 10 days and a model is saved every 4.5 hours. We decode the test set using an ensemble of 4 models. These are the 4 models with highest BLEU scores on the development set.

4 Linguistically-Augmented Statistical MT

4.1 Factored models and morphological expansion in Phrase-based MT

One of the approaches followed in the English–Croatian SMT systems built in Deliverable 4.1c (Pirinen et al., 2015b) for improving the grammatical quality

[3] `https://kheafield.com/code/kenlm/`

of the output was factored translation models. However, the improvement obtained over a system without factors was really small.

During the last year, we studied more deeply the reasons behind this underperformance (Sánchez-Cartagena et al., 2016). We found out that both the order of the language model that operates on morphosyntactic tags and the way in which the part-of-speech tagging algorithm used to annotate the data uses the information from the lexicon have a huge impact on the translation quality of the final systems. Optimising these parameters and combining factored models with data selection, we built a system that reached the translation quality of the commercial system Google Translate.[4]

Additionally, we experimented with morphological expansion (Sánchez-Cartagena et al., 2016, Sec. 3): we tried to generate synthetic data in order to fill the vocabulary gaps caused by the high degree of inflection of Croatian, but we did not obtain statistically significant improvements.

4.2 Factored models in Neural MT

While in SMT the factored model architecture permits annotating each SL or TL word with additional linguistic information that can be leveraged by the statistical models to achieve better translation quality, very little effort has been put in the integration of linguistic information into neural machine translation (NMT). The most relevant approach is that by Sennrich and Haddow (2016), who enabled the use of linguistic factors in the SL: the real-valued vector that represents each word is built by combining its different factors and a single encoder recurrent neural network is used. They report statistically significant gains for 3 language pairs.

We explored the use of a different approach for integrating SL linguistic information. We followed a similar approach to that followed by Zoph and Knight (2016) for multi-source neural machine translation and use of a different encoder for each SL factor. The potential advantage of this approach over the strategy devised by Sennrich and Haddow (2016) is that the attention mechanism operates independently for each SL factor. This allows the system to deal in a more effective way with phenomena such as agreement since, when producing a TL word, its morphological inflection information may need to be obtained from a SL word different from the SL word from which its lemma comes from.

[4] `http://translate.google.com`

Experiments with the German–English language pair and the same data used by Sennrich and Haddow (2016) showed no statistically significant difference between their approach and ours. The code of our implementation is available at `https://github.com/vitaka/nematus/tree/sl-factors-1`. Further work will be carried out in order to assess the role of word segmentation (Sennrich and Haddow (2016) segmented the words in smaller units and so did we in our experiments) and size of the encoders in the results.

4.3 Morph segmentation

Morphological segmentation is recognised as a potential solution in statistical machine translation (SMT) to deal with data sparsity posed by morphologically complex languages. Two approaches have been used in the literature, rule-based and statistical, but always in isolation. In addition, previous work has failed to bring significant improvement and conclusive analyses of the effects of segmentation.

We carried out an investigation (Pirinen et al., 2016) where we used both rule-based and unsupervised approaches to segmentation jointly and aimed to find out where they excel and where they fail. Our case study is an in-depth analysis on our English-to-Finnish systems submitted to the WMT 2015 shared task (Rubino et al., 2015). To that end, we evaluated these SMT systems built with different segmenters from different perspectives: intrinsic evaluation, MT automatic metrics, MT human evaluation and MT linguistic evaluation. In terms of automatic metrics, we found out the best system to be the one that combines both rule-based and unsupervised segmentations, outperforming an unsegmented system by a wide margin (1.08 BLEU and 3.64 TER points). Human evaluation showed that the outputs produced by an SMT system with rule-based segmentations are preferred over those of the system that uses unsupervised segmentations.

5 MT systems between closely-related languages

This section describes work carried out during year 4 on the project aimed at developing MT systems between languages of the South-Slavic family.

Translation between closely related languages is one of the scenarios in which rule-based machine translation can still be a competitive approach, as structural changes between the languages are easy to encode by transfer rules and the powerful lexical selection methods of corpus-based approaches are not required. Words with multiple meanings in the source language often have the same meanings in the target language and so it is not necessary to infer the disambiguated meaning of a word in order to correctly translate it. Moreover, if the languages are highly inflected (as happens with South-Slavic languages), the advantage of rule-based machine translation over corpus-based approaches grows because morphological dictionaries and inflection paradigms allow the developers of the rule-based system to easily encode all the inflected forms of each word.

Be that as it may, building rule-based machine translation systems usually requires linguists trained on the inner workings of the system to invest a considerable amount of time in creating the linguistic resources of the system. In order to mitigate this and make the process less time-consuming, the open-source Apertium platform (Forcada et al., 2010) aims at allowing the collaborative building of rule-based machine translation systems for language pairs with scarce resources thanks to its intuitive shallow-transfer formalism.

However, even with Apertium at our disposal, there is still the difficulty of employing linguists that are trained in both languages that could feed the necessary linguistic data into our system. So we took the initial line of thinking a step further, and we combat this problem by designing an approach focused on collaboration between non-linguists and linguists. This allowed us to combine automated data collection with non-expert annotation in order to build a bidirectional Croatian-Serbian rule-based system in considerably little time.

We organized two workshops in Zagreb which were attended by a mixed audience of linguists and non-linguists, all of whom were native Croatian speakers with circumstantial knowledge of Serbian. The first workshop was focused on building a bilingual dictionary by presenting the attendees with automatically produced bilingual lexical candidates and asked them to verify whether these words were indeed translations. The second workshop was focused on transfer rules, where we automatically extracted translation rules and had the participants validate them based examples of rules in use - all they had to do was answer the question “Is this a valid translation?” Once we obtained all the data, we double-checked it and added it into our system, which currently has 88521 bilingual entries as well as 99 transfer rules in the Croatian–Serbian direction, and 86 transfer rules in the Serbian–Croatian direction. The system was developed in a total of approximately 6 person months, which is very little time considering how time-consuming building rule-based systems is. An additional achievement is the fact that, when compared to Google Translate, the only other available system at the time, our system yielded higher TER scores. More details on the construction process and evaluation can be found in the paper written by Klubička et al. (2016).

Even though machine translation between closely related languages is less challenging and exhibits a smaller number of translation errors than translation between distant languages, there are still obstacles which should be addressed in order to improve such systems. This is why an additional evaluation of the resulting system has been carried out. We wished to analyze these obstacles on machine translation systems between closely related South Slavic languages, namely Croatian, Serbian and Slovenian, in order to identify problems and try to solve them so we could further improve our systems. For details on the comparative evaluation, refer to Popović et al. (2016); general conclusions drawn after much comparison and evaluation, show that for all language pairs and for both translation systems that were investigated, the main obstacles are the differences between syntactic properties. This suggests that more work should be done in regards to defining transfer rules, perhaps employing a more comprehensive approach or devising a more complex transfer framework.

6 MT systems for specific domains

This section covers work carried out during year 4 concerning the adaptation of MT to specific text domains.

During this year we have published a journal paper (Toral et al., 2016) that presents a widely applicable methodology to bring MT to new domains of under-resourced languages in a cost-effective and rapid manner. It relies on web crawling to automatically acquire parallel data to train statistical MT systems if any such data can be found for the language pair and domain of interest. If that is not the case, we resort to (1) crowdsourcing to translate small amounts of text (hundreds of sentences), which are then used to tune statistical MT models, and (2) web crawling of vast amounts of monolingual data (millions of sentences), which are then used to build language models for MT.

We have applied these methods to two respective use cases for the main target language of the project: Croatian. The first use case regards tourism, given the importance of this sector to Croatia's economy, and builds on the work done for this domain during the second year of the project (cf. Section 2.2. in Deliverable 2.2b), while the second use case has to do with tweets, due to the growing importance of social media. For tourism, we crawled parallel data from 20 web domains using two state-of-the-art crawlers developed in the project (Bitextor and ILSP Focused Crawler) and explored how to combine the crawled data with bigger amounts of general-domain data. In the social media use case, we dealt with tweets from the 2014 edition of the soccer World Cup. We built domain-adapted systems by (1) translating small amounts of tweets to be used for tuning by means of crowdsourcing and (2) crawling vast amounts of monolingual tweets. In both use cases our domain adapted systems outperformed strong baselines.

The methodology developed in the project to build domain-specific MT systems, in spite of having been so far only applied to the target languages of the project, is, to a large extent, language-independent. During this year we have applied it to a new language pair and domain: tourism in English–Greek. The development of this MT system involved a transfer of knowledge to the partner ILSP by means of a secondment of Prompsit's researcher Víctor M. Sánchez-Cartagena.

In order to build MT systems focused on the tourism domain for English-to-Greek and Greek-to-English, firstly in-domain parallel and monolingual data were obtained by crawling content from museum, archaeological and tourism-related websites with ILSP Focused Crawler (Papavassiliou et al., 2013). We also collected all the publicly available English–Greek parallel data (to be used our out-of-domain parallel corpus) as well as all the publicly available Greek monolingual corpora. As English monolingual corpora, we used those published for the WMT 2016 shared translation task.[5]

We evaluated some of the most relevant methods for combining in-domain and out-of-domain data for SMT that can be found in the literature and chose the best performing method for each translation direction. Some of the combination methods, such as linear interpolation of two phrase tables (Sennrich, 2012) and parallel data selection (Rubino et al., 2014) were already tested during the project, but other ones, like monolingual data selection (Ruiz et al., 2012), fill-up (Bisazza et al., 2011), or linear interpolation of multiple phrase tables (Sennrich, 2012), were not.

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

As a result, we obtained useful insights about the scenarios that best fit each domain adaptation method. Namely, we found out that, when Greek is the target language (i.e., when the target language has a higher inflection degree than the source language) only domain adaptation methods that operate on the target language model are useful and translation quality does not grow when increasing the amount of out-of-domain parallel data. In the opposite direction, however, the combination of methods that operate on the phrase table and on the language model is the best option and the addition of more out-of-domain parallel data yields higher translation quality. When Greek is the source language, the out-of-domain parallel data helps to reduce the out-of-vocabulary words and thus phrase table adaptation is necessary. When Greek is the target language, on the contrary, the language model is a very important part of the system: the number of translations of each English phrase is higher and a powerful language model helps to correctly combine the different hypotheses.

7 Participation in Shared Tasks

As in the previous year, we have taken part in shared tasks related to machine translation (cf. Section 7.1) and quality estimation (cf. Section 7.2). Our submissions build upon the work we carried out for these shared tasks last year (cf. Section in 5.1 and 5.2 in Deliverable D4.1c (Pirinen et al., 2015b)).

7.1 WMT16 Translation Task

As in previous years we participated in the WMT translation shared task (Sánchez-Cartagena and Toral, 2016). The language direction in which we focused is, as last year, English–Finnish. This year we applied morphological segmentation and deep learning in order to address (i) the data scarcity problem caused by the lack of in-domain parallel data in the constrained task and (ii) the complex morphology of Finnish. We submitted a neural machine translation system, a phrase-based machine translation system re-ranked with a neural language model and the combination of their outputs tuned on character sequences. Our combination and our neural system were ranked first and second respectively according to automatic evaluation metrics and tied for the first place in the human evaluation.

7.2 WMT16 Quality Estimation

Partner UA took part in the WMT16 shared task for MT quality estimation (QE), namely, in two of the sub-tasks: word-level QE and phrase-level QE (Esplà-Gomis et al., 2016). The participation continued with the method presented by Esplà-Gomis et al. (2015) in the previous edition of WMT'15, and adapted it so it could be used for phrases. The approaches submitted use a multilayer perceptron to exploit any source of bilingual information available in an agnostic way for detecting possible edits, both at the level of words or phrases. The authors define sources of bilingual information as any resource that allow to translate sub-segments either from the source language into the target language or vice versa. The methods by Esplà-Gomis et al. (2016) use two MT systems as sources of bilingual information, namely Lucy and Google Translate, as well as the on-line bilingual concordancer Reverso Context, which is available online[6] and has been developed the partner Prompsit Language Engineering. Two systems were submitted to each of the subtasks at WMT16 shared task, one using the features described by Esplà-Gomis et al. (2016) and the other combining them with the baseline features provided by the organisation. While the best-performing submission for word-level QE clearly outperformed the baseline, the results were not that good for phrase-level QE.

8 Conclusions

This deliverable has covered the work done in the area of MT development and deployment (work package 4) during the period of the forth milestone of the project (M37–M48). During this period we have conducted a series of activities that build upon the developments done in this area of the project and also in other areas in previous years.

We have developed a cloud-based architecture to build language models, which allows us to use the vast amounts of data crawled in the project in an efficient manner. We have continued the work devoted to augmenting our statistical machine translation (SMT) system with linguistic information. In particular, we have studied the best way to use factored models and also the effects of the morphological segmentation approaches used in the previous milestone.

[6] `http://context.reverso.net/translation`

Due to the emergence of a new paradigm to automatic translation, neural MT (NMT), we have built systems making use of this approach. For example, in our participation in the 2016 edition of the WMT translation shared task, which obtained the top rank. On top of that we have conducted novel research to add linguistic information to the NMT architecture.

Finally, concerning rapid development of MT, we have built MT engines between South-Slavic languages using a collaborative approach and we have demonstrated the large degree of language independence of the approaches we have developed for domain adaptation, by applying them to a new language pair: English–Greek.

References

Bisazza, A., Ruiz, N., and Federico, M. (2011). Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, USA.

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. (2016). UAlacant word-level and phrase-level machine translation quality estimation systems at WMT 2016. In *Proceedings of the First Conference on Machine Translation*, pages 782–786, Berlin, Germany. Association for Computational Linguistics.

Esplà-Gomis, M., Forcada, M. L., Ljubešić, N., Papavassiliou, V., Prokopidis, P., Ortiz-Rojas, S., Rubino, R., Sánchez-Cartagena, V., and Toral, A. (2015). Abu-matran deliverable D3.1c: Acquisition for the third development cycle. Technical report.

Ferrández-Tordera, J., Ortiz-Rojas, S., and Toral, A. (2016). CloudLM: a Cloud-based Language Model for Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 105:51–61.

Forcada, M. L., Rosell, M. G., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Tyers, F. M. (2010). Apertium: a free/open-source platform for rule-based machine translation platform. *Machine Translation*.

Klubička, F., Ramírez-Sánchez, G., and Ljubešić, N. (2016). Collaborative development of a rule-based machine translator between croatian and serbian. *Baltic Journal of Modern Computing*, 4(2):361–367.

Papavassiliou, V., Prokopidis, P., and Thurmair, G. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51, Sofia, Bulgaria. Association for Computational Linguistics.

Pirinen, T., Rubino, R., Sánchez-Cartagena, V., Klubička, F., and Toral, A. (2015a). Abu-matran deliverable D5.1c: Evaluation for the third development cycle. Technical report.

Pirinen, T., Rubino, R., Sánchez-Cartagena, V. M., and Toral, A. (2015b). Abu-matran deliverable D4.1c: Mt systems for the third development cycle. Technical report.

Pirinen, T. A., Toral, A., and Rubino, R. (2016). Rule-based and statistical morph segments in english-to-finnish smt. In *Proceedings of the 2nd International Workshop on Computational Linguistics for Uralic Languages*, Szeged, Hungary.

Popović, M., Arcan, M., and Klubička, F. (2016). Language related issues for machine translation between closely related south slavic languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 43–52, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Rubino, R., Toral, A., Sánchez-Cartagena, V. M., Ferrández-Tordera, J., Ortiz Rojas, S., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Way, A. (2014). Abu-MaTran at WMT 2014 translation task: Two-step data selection and rbmt-style synthetic rules. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 171–177, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ruiz, N., Bisazza, A., Cattoni, R., and Federico, M. (2012). FBK's machine translation systems for IWSLT 2012's TED lectures. In *Proceedings of the 9th International Workshop on Spoken Language Translation*, pages 61–68, Hong Kong.

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. and Toral, A. (2016). Abu-MaTran at WMT 2016 translation task: Deep learning, morphological segmentation and tuning on character sequences. In *Proceedings of the First Conference on Machine Translation*, pages 362–370, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549, Avignon, France.

Sennrich, R. and Haddow, B. (2016). Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Toral, A., Esplá-Gomis, M., Klubička, F., Ljubešić, N., Papavassiliou, V., Prokopidis, P., Rubino, R., and Way, A. (2016). Crawl and crowd to bring machine translation to under-resourced languages. *Language Resources and Evaluation*, pages 1–33.

Zoph, B. and Knight, K. (2016). Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.