ONTS: "Optima" News Translation System

Marco Turchi*, Martin Atkinson*, Alastair Wilcox+, Brett Crawley,
Stefano Bucci+, Ralf Steinberger* and Erik Van der Goot*
European Commission - Joint Research Centre (JRC), IPSC - GlobeSec
Via Fermi 2749, 21020 Ispra (VA) - Italy
*[name].[surname]@jrc.ec.europa.eu
+[name].[surname]@ext.jrc.ec.europa.eu
brettcrawley@gmail.com

Abstract

We propose a real-time machine translation system that allows users to select a news category and to translate the related live news articles from Arabic, Czech, Danish, Farsi, French, German, Italian, Polish, Portuguese, Spanish and Turkish into English. The Moses-based system was optimised for the news domain and differs from other available systems in four ways: (1) News items are automatically categorised on the source side, before translation; (2) Named entity translation is optimised by recognising and extracting them on the source side and by re-inserting their translation in the target language, making use of a separate entity repository; (3) News titles are translated with a separate translation system which is optimised for the specific style of news titles; (4) The system was optimised for speed in order to cope with the large volume of daily news articles.

1 Introduction

Being able to read news from other countries and written in other languages allows readers to be better informed. It allows them to detect national news bias and thus improves transparency and democracy. Existing online translation systems such as *Google Translate* and *Bing Translator*[1] are thus a great service, but the number of documents that can be submitted is restricted (Google will even entirely stop their service in 2012) and submitting documents means disclosing the users' interests and their (possibly sensitive) data to the service-providing company.

For these reasons, we have developed our in-house machine translation system ONTS. Its translation results will be publicly accessible as part of the Europe Media Monitor family of applications, (Steinberger et al., 2009), which gather and process about 100,000 news articles per day in about fifty languages. ONTS is based on the open source phrase-based statistical machine translation toolkit Moses (Koehn et al., 2007), trained mostly on freely available parallel corpora and optimised for the news domain, as stated above. The main objective of developing our in-house system is thus not to improve translation quality over the existing services (this would be beyond our possibilities), but to offer our users a rough translation (a "gist") that allows them to get an idea of the main contents of the article and to determine whether the news item at hand is relevant for their field of interest or not.

A similar news-focused translation service is "Found in Translation" (Turchi et al., 2009), which gathers articles in 23 languages and translates them into English. "Found in Translation" is also based on Moses, but it categorises the news after translation and the translation process is not optimised for the news domain.

2 Europe Media Monitor

Europe Media Monitor (EMM)[2] gathers a daily average of 100,000 news articles in approximately 50 languages, from about 3,400 hand-selected web news sources, from a couple of hundred specialist and government websites, as well as from about twenty commercial news providers. It visits the news web sites up to every five minutes to

[1] http://translate.google.com/ and http://www.microsofttranslator.com/

[2] http://emm.newsbrief.eu/overview.html

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 25–30,
Avignon, France, April 23 - 27 2012.

search for the latest articles. When news sites offer RSS feeds, it makes use of these, otherwise it extracts the news text from the often complex HTML pages. All news items are converted to Unicode. They are processed in a pipeline structure, where each module adds additional information. Independently of how files are written, the system uses UTF-8-encoded RSS format.

Inside the pipeline, different algorithms are implemented to produce monolingual and multilingual clusters and to extract various types of information such as named entities, quotations, categories and more. ONTS uses two modules of EMM: the named entity recognition and the categorization parts.

2.1 Named Entity Recognition and Variant Matching.

Named Entity Recognition (NER) is performed using manually constructed language-independent rules that make use of language-specific lists of trigger words such as titles (president), professions or occupations (tennis player, playboy), references to countries, regions, ethnic or religious groups (French, Bavarian, Berber, Muslim), age expressions (57-year-old), verbal phrases (deceased), modifiers (former) and more. These patterns can also occur in combination and patterns can be nested to capture more complex titles, (Steinberger and Pouliquen, 2007). In order to be able to cover many different languages, no other dictionaries and no parsers or part-of-speech taggers are used.

To identify which of the names newly found every day are new entities and which ones are merely variant spellings of entities already contained in the database, we apply a language-independent name similarity measure to decide which name variants should be automatically merged, for details see (Pouliquen and Steinberger, 2009). This allows us to maintain a database containing over 1,15 million named entities and 200,000 variants. The major part of this resource can be downloaded from `http://langtech.jrc.it/JRC-Names.html`

2.2 Category Classification across Languages.

All news items are categorized into hundreds of categories. Category definitions are multilingual, created by humans and they include geographic regions such as each country of the world, organizations, themes such as natural disasters or security, and more specific classes such as earthquake, terrorism or tuberculosis,

Articles fall into a given category if they satisfy the category definition, which consists of Boolean operators with optional vicinity operators and wild cards. Alternatively, cumulative positive or negative weights and a threshold can be used. Uppercase letters in the category definition only match uppercase words, while lowercase words in the definition match both uppercase and lowercase words. Many categories are defined with input from the users themselves. This method to categorize the articles is rather simple and user-friendly, and it lends itself to dealing with many languages, (Steinberger et al., 2009).

3 News Translation System

In this section, we describe our statistical machine translation (SMT) service based on the open-source toolkit Moses (Koehn et al., 2007) and its adaptation to translation of news items.

Which is the most suitable SMT system for our requirements? The main goal of our system is to help the user understand the content of an article. This means that a translated article is evaluated positively even if it is not perfect in the target language. Dealing with such a large number of source languages and articles per day, our system should take into account the translation speed, and try to avoid using language-dependent tools such as part-of-speech taggers.

Inside the Moses toolkit, three different statistical approaches have been implemented: *phrase based statistical machine translation* (PB-SMT) (Koehn et al., 2003), *hierarchical phrase based statistical machine translation* (Chiang, 2007) and *syntax-based statistical machine translation* (Marcu et al., 2006). To identify the most suitable system for our requirements, we run a set of experiments training the three models with Europarl V4 German-English (Koehn, 2005) and optimizing and testing on the News corpus (Callison-Burch et al., 2009). For all of them, we use their default configurations and they are run under the same condition on the same machine to better evaluate translation time. For the syntax model we use linguistic information only on the target side. According to our experiments, in terms of performance the hierarchical model

performs better than PBSMT and syntax (18.31, 18.09, 17.62 Bleu points), but in terms of translation speed PBSMT is better than hierarchical and syntax (1.02, 4.5, 49 second per sentence). Although, the hierarchical model has the best Bleu score, we prefer to use the PBSMT system in our translation service, because it is four times faster.

Which training data can we use? It is known in statistical machine translation that more training data implies better translation. Although, the number of parallel corpora has been is growing in the last years, the amounts of training data vary from language pair to language pair. To train our models we use the freely available corpora (when possible): Europarl (Koehn, 2005), JRC-Acquis (Steinberger et al., 2006), DGT-TM[3], Opus (Tiedemann, 2009), SE-Times (Tyers and Alperen, 2010), Tehran English-Persian Parallel Corpus (Pilevar et al., 2011), News Corpus (Callison-Burch et al., 2009), UN Corpus (Rafalovitch and Dale, 2009), CzEng0.9 (Bojar and Žabokrtský, 2009), English-Persian parallel corpus distributed by ELRA[4] and two Arabic-English datasets distributed by LDC[5]. This results in some language pairs with a large coverage, (more than 4 million sentences), and other with a very small coverage, (less than 1 million). The language models are trained using 12 model sentences for the content model and 4.7 million for the title model. Both sets are extracted from English news.

For less resourced languages such as Farsi and Turkish, we tried to extend the available corpora. For Farsi, we applied the methodology proposed by (Lambert et al., 2011), where we used a large language model and an English-Farsi SMT model to produce new sentence pairs. For Turkish we added the Movie Subtitles corpus (Tiedemann, 2009), which allowed the SMT system to increase its translation capability, but included several slang words and spoken phrases.

How to deal with Named Entities in translation? News articles are related to the most important events. These names need to be efficiently translated to correctly understand the content of an article. From an SMT point of view, two main issues are related to Named Entity translation: (1) such a name is not in the training data or (2) part of the name is a common word in the target language and it is wrongly translated, e.g. the French name "Bruno Le Maire" which risks to be translated into English as "Bruno Mayor". To mitigate both the effects we use our multilingual named entity database. In the source language, each news item is analysed to identify possible entities; if an entity is recognised, its correct translation into English is retrieved from the database, and suggested to the SMT system enriching the source sentence using the xml markup option [6] in Moses. This approach allows us to complement the training data increasing the translation capability of our system.

How to deal with different language styles in the news? News title writing style contains more gerund verbs, no or few linking verbs, prepositions and adverbs than normal sentences, while content sentences include more preposition, adverbs and different verbal tenses. Starting from this assumption, we investigated if this phenomenon can affect the translation performance of our system.

We trained two SMT systems, $SMT_{content}$ and SMT_{title}, using the Europarl V4 German-English data as training corpus, and two different development sets: one made of content sentences, News Commentaries (Callison-Burch et al., 2009), and the other made of news titles in the source language which were translated into English using a commercial translation system. With the same strategy we generated also a Title test set. The SMT_{title} used a language model created using only English news titles. The News and Title test sets were translated by both the systems. Although the performance obtained translating the News and Title corpora are not comparable, we were interested in analysing how the same test set is translated by the two systems. We noticed that translating a test set with a system that was optimized with the same type of data resulted in almost 2 Blue score improvements: Title-TestSet: 0.3706 (SMT_{title}), 0.3511 ($SMT_{content}$); News-TestSet: 0.1768 (SMT_{title}), 0.1945 ($SMT_{content}$). This behaviour was present also in different language pairs. According to these results we decided to use two different translation systems for each language pair, one optimized using title data

[3] http://langtech.jrc.it/DGT-TM.html

[4] http://catalog.elra.info/

[5] http://www.ldc.upenn.edu/

[6] http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc4

and the other using normal content sentences. Even though this implementation choice requires more computational power to run in memory two Moses servers, it allows us to mitigate the workload of each single instance reducing translation time of each single article and to improve translation quality.

3.1 Translation Quality

To evaluate the translation performance of ONTS, we run a set of experiments where we translate a test set for each language pair using our system and Google Translate. Lack of human translated parallel titles obliges us to test only the content based model. For German, Spanish and Czech we use the news test sets proposed in (Callison-Burch et al., 2010), for French and Italian the news test sets presented in (Callison-Burch et al., 2008), for Arabic, Farsi and Turkish, sets of 2,000 news sentences extracted from the Arabic-English and English-Persian datasets and the SE-Times corpus. For the other languages we use 2,000 sentences which are not news but a mixture of JRC-Acquis, Europarl and DGT-TM data. It is not guarantee that our test sets are not part of the training data of Google Translate.

Each test set is translated by Google Translate - Translator Toolkit, and by our system. Bleu score is used to evaluate the performance of both systems. Results, see Table 1, show that Google Translate produces better translation for those languages for which large amounts of data are available such as French, German, Italian and Spanish. Surprisingly, for Danish, Portuguese and Polish, ONTS has better performance, this depends on the choice of the test sets which are not made of news data but of data that is fairly homogeneous in terms of style and genre with the training sets.

The impact of the named entity module is evident for Arabic and Farsi, where each English suggested entity results in a larger coverage of the source language and better translations. For highly inflected and agglutinative languages such as Turkish, the output proposed by ONTS is poor. We are working on gathering more training data coming from the news domain and on the possibility of applying a linguistic pre-processing of the documents.

Source L.	ONTS	Google T.
Arabic	0.318	0.255
Czech	0.218	0.226
Danish	0.324	0.296
Farsi	0.245	0.197
French	0.26	0.286
German	0.205	0.25
Italian	0.234	0.31
Polish	0.568	0.511
Portuguese	0.579	0.424
Spanish	0.283	0.334
Turkish	0.238	0.395

Table 1: Automatic evaluation.

4 Technical Implementation

The translation service is made of two components: the connection module and the Moses server. The connection module is a servlet implemented in Java. It receives the RSS files, isolates each single news article, identifies each source language and pre-processes it. Each news item is split into sentences, each sentence is tokenized, lowercased, passed through a statistical compound word splitter, (Koehn and Knight, 2003), and the named entity annotator module. For language modelling we use the KenLM implementation, (Heafield, 2011).

According to the language, the correct Moses servers, title and content, are fed in a multi-thread manner. We use the multi-thread version of Moses (Haddow, 2010). When all the sentences of each article are translated, the inverse process is run: they are detokenized, recased, and untranslated/unknown words are listed. The translated title and content of each article are uploaded into the RSS file and it is passed to the next modules.

The full system including the translation modules is running in a 2xQuad-Core with Intel Hyper-threading Technology processors with 48GB of memory. It is our intention to locate the Moses servers on different machines. This is possible thanks to the high modularity and customization of the connection module. At the moment, the translation models are available for the following source languages: Arabic, Czech, Danish, Farsi, French, German, Italian, Polish, Portuguese, Spanish and Turkish.

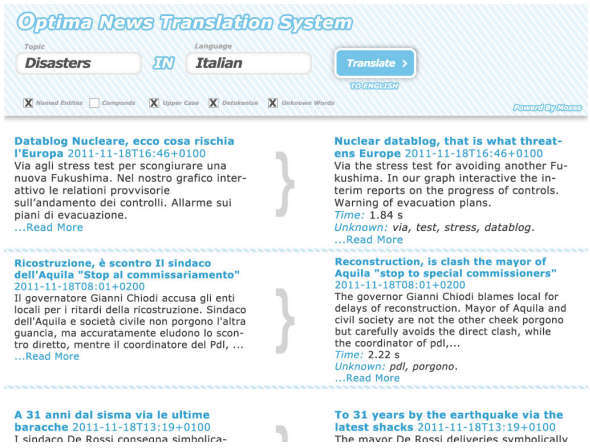

Figure 1: Demo Web site.

4.1 Demo

Our translation service is currently presented on a demo web site, see Figure 1, which is available at `http://optima.jrc.it/Translate/`. News articles can be retrieved selecting one of the topics and the language. All the topics are assigned to each article using the methodology described in 2.2. These articles are shown in the left column of the interface. When the button "Translate" is pressed, the translation process starts and the translated articles appear in the right column of the page.

The translation system can be customized from the interface enabling or disabling the named entity, compound, recaser, detokenizer and unknown word modules. Each translated article is enriched showing the translation time in milliseconds per character and, if enabled, the list of unknown words. The interface is linked to the connection module and data is transferred using RSS structure.

5 Discussion

In this paper we present the Optima News Translation System and how it is connected to Europe Media Monitor application. Different strategies are applied to increase the translation performance taking advantage of the document structure and other resources available in our research group. We believe that the experiments described in this work can result very useful for the development of other similar systems. Translations produced by our system will soon be available as part of the main EMM applications.

The performance of our system is encouraging, but not as good as the performance of web services such as Google Translate, mostly because we use less training data and we have reduced computational power. On the other hand, our in-house system can be fed with a large number of articles per day and sensitive data without including third parties in the translation process. Performance and translation time vary according to the number and complexity of sentences and language pairs.

The domain of news articles dynamically changes according to the main events in the world, while existing parallel data is static and usually associated to governmental domains. It is our intention to investigate how to adapt our translation system updating the language model with the English articles of the day.

Acknowledgments

The authors thank the JRC's OPTIMA team for its support during the development of ONTS.

References

O. Bojar and Z. Žabokrtský. 2009. *CzEng0.9: Large Parallel Treebank with Rich Annotation*. Prague Bulletin of Mathematical Linguistics, 92.

C. Callison-Burch and C. Fordyce and P. Koehn and C. Monz and J. Schroeder. 2008. *Further Meta-Evaluation of Machine Translation*. Proceedings of the Third Workshop on Statistical Machine Translation, pages 70–106. Columbus, US.

C. Callison-Burch, and P. Koehn and C. Monz and J. Schroeder. 2009. *Findings of the 2009 Workshop on Statistical Machine Translation*. Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 1–28. Athens, Greece.

C. Callison-Burch, and P. Koehn and C. Monz and K. Peterson and M. Przybocki and O. Zaidan. 2009. *Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation*. Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 17–53. Uppsala, Sweden.

D. Chiang. 2005. *Hierarchical phrase-based translation*. Computational Linguistics, 33(2): pages 201–228. MIT Press.

B. Haddow. 2010. *Adding multi-threaded decoding to moses*. The Prague Bulletin of Mathematical Linguistics, 93(1): pages 57–66. Versita.

K. Heafield. 2011. *KenLM: Faster and smaller language model queries*. Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, UK.

P. Koehn. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation*. Proceedings of the Machine Translation Summit X, pages 79-86. Phuket, Thailand.

P. Koehn and F. J. Och and D. Marcu. 2003. *Statistical phrase-based translation*. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 48–54. Edmonton, Canada.

P. Koehn and K. Knight. 2003. *Empirical methods for compound splitting*. Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, pages 187–193. Budapest, Hungary.

P. Koehn and H. Hoang and A. Birch and C. Callison-Burch and M. Federico and N. Bertoldi and B. Cowan and W. Shen and C. Moran and R. Zens and C. Dyer and O. Bojar and A. Constantin and E. Herbst 2007. *Moses: Open source toolkit for statistical machine translation*. Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstration session, pages 177–180. Columbus, Oh, USA.

P. Lambert and H. Schwenk and C. Servan and S. Abdul-Rauf. 2011. *SPMT: Investigations on Translation Model Adaptation Using Monolingual Data*. Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 284–293. Edinburgh, Scotland.

D. Marcu and W. Wang and A. Echihabi and K. Knight. 2006. *SPMT: Statistical machine translation with syntactified target language phrases*. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 48–54. Edmonton, Canada.

M. Pilevar and H. Faili and A. Pilevar. 2011. *TEP: Tehran English-Persian Parallel Corpus*. Computational Linguistics and Intelligent Text Processing, pages 68–79. Springer.

B. Pouliquen and R. Steinberger. 2009. *Automatic construction of multilingual name dictionaries*. Learning Machine Translation, pages 59–78. MIT Press - Advances in Neural Information Processing Systems Series (NIPS).

A. Rafalovitch and R. Dale. 2009. *United nations general assembly resolutions: A six-language parallel corpus*. Proceedings of the MT Summit XIII, pages 292–299. Ottawa, Canada.

R. Steinberger and B. Pouliquen. 2007. *Cross-lingual named entity recognition*. Lingvisticæ Investigationes, 30(1) pages 135–162. John Benjamins Publishing Company.

R. Steinberger and B. Pouliquen and A. Widiger and C. Ignat and T. Erjavec and D. Tufiş and D. Varga. 2006. *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*. Proceedings of the 5th International Conference on Language Resources and Evaluation, pages 2142–2147. Genova, Italy.

R. Steinberger and B. Pouliquen and E. van der Goot. 2009. *An Introduction to the Europe Media Monitor Family of Applications*. Proceedings of the Information Access in a Multilingual World-Proceedings of the SIGIR 2009 Workshop, pages 1–8. Boston, USA.

J. Tiedemann. 2009. *News from OPUS-A Collection of Multilingual Parallel Corpora with Tools and Interfaces*. Recent advances in natural language processing V: selected papers from RANLP 2007, pages 309:237.

M. Turchi and I. Flaounas and O. Ali and T. DeBie and T. Snowsill and N. Cristianini. 2009. *Found in translation*. Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, pages 746–749. Bled, Slovenia.

F. Tyers and M.S. Alperen. 2010. *South-East European Times: A parallel corpus of Balkan languages*. Proceedings of the LREC workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages, Valletta, Malta.