Application of Online Terminology Services in Statistical Machine Translation

Raivis Skadiņš, Mārcis Pinnis, Tatiana Gornostay, Andrejs Vasiļjevs
Tilde, Vienibas gatve 75a, Riga, Latvia
{raivis.skadins|marcis.pinnis|tatiana.gornostay|andrejs}@tilde.lv

Abstract

In this system demonstration paper we present a cloud-based platform providing online terminology services for human and machine users. We focus on the use case for the application of online terminology services in statistical machine translation and describe the applied methods for monolingual and bilingual terminology integration into statistical machine translation during training and translation phases.

Keywords: online services, terminology service, statistical machine translation, terminology extraction, terminology translation

1 Introduction

Accurate use of terminology is critical within content life cycle from its creation to distribution, including content translation and localisation, to ensure efficient and precise professional communication.

Traditionally, terminology resources are collected and stored in terminology databases, mostly used by human users through Web-based interfaces or database integration into authoring and/or computer-assisted translation (CAT) tools. Such databases are usually populated manually with new terms by terminologists or domain experts. Furthermore, most terminology databases fail to provide extensive up-to-date multilingual terminology, since in the dynamic pace of technological and societal development new terms are coined every day by industry, translation agencies, collective and individual authors. Moreover, terms in under-resourced languages and/or specific domains are particularly poorly represented in online terminology databases. Most of the online terminology databases offer not much more than the typical database features of storing and querying terminology entries.

The evolution of the Internet and cloud-computing opens the opportunity to advance the automation of terminology and translation work by creating cloud-based terminology services for the key terminology tasks. Such work is being carried out in the FP7 project Terminology as a Service[1] (TaaS). TaaS platform provides a variety of online terminology services, to serve the needs for automated acquisition, processing, and application of terminological data by human users (i.e., language workers), for example:

- **Automatic extraction of monolingual term candidates**, using state-of-the-art terminology extraction techniques, from documents uploaded by users;
- **Automatic lookup of translation equivalent term candidates** in user-defined target language(s) from different terminology databases (for automatically extracted monolingual term candidates);
- **Automatic extraction of translation equivalent term candidates** from parallel and/or comparable Web data, using state-of-the-art terminology extraction and bilingual terminology alignment techniques (for automatically extracted monolingual term candidates);
- **Facilities for cleaning up** automatically acquired raw terminological data;
- **Facilities for exporting** terminological data in different formats, e.g., TSV, CSV, TBX, and others.

Terminology services can be also exploited by machine users (i.e., language processing applications), such as CAT tools, machine translation systems, search engines, and others. Thus, termi-

[1] More information about the *TaaS project* can be found on the project's official Web page: http://www.taas-project.eu.

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 281–286.

nology services have the potential to significantly enhance the quality of language tools and natural language processing in general.

In this paper we focus on a particular use case of the application of online terminology services in statistical machine translation (SMT), i.e., terminology services exploited for the adaptation of SMT with domain and task specific terminology (these are monolingual term candidate extraction and automatic extraction of translation equivalent term candidates), with a special focus on under-resourced languages and the languages with a high degree of inflection (i.e., rich morphology).

In the next section we overview the related work in the field and consider existing methods for terminology translation in statistical MT. In the third section we describe our solution of handling terminology in SMT via online terminology services being developed within the TaaS project.

The conceptual design for the integration of terminology services into SMT is also outlined within this paper. Finally, we make conclusions and outline future work in the proposed direction.

2 Related work: terminology handling in statistical machine translation

There are several research works reporting improvements of translation quality in terms of automatic machine translation evaluation metrics after integration of multiword expressions in a parallel corpus. Bouamor et al. (2012) observed a gain of +0.3 BLEU points for French-English SMT. Nikoulina et al. (2012) proposed a framework for integrating Named Entities (NE) within SMT. It was shown that the introduced model can lead to +2-3 BLEU points improvement over a baseline system for two different test sets.

Current SMT phrase-based models, including Moses (Koehn et al., 2007), do not handle terminology translation. Although domain adaptation can be done using additional in-domain training data (Koehn and Schroeder, 2007), such an approach is very resource intensive and requires SMT model training for each specific domain. In cases when language resources are very limited or a user requires translation of a document that is written in a different domain, not covered by available SMT models, domain adaptation is not applicable. This means that terminology diversity within domains is not well-managed with current approaches. For example, a term „tablet" is ambiguous – it can refer to a popular consumer electronics product (a tablet computer), a number of sheets of paper fastened together along one edge (WordNet 3.1), a pill used in medicine, and others. An SMT system would translate this term in every single case according to its statistical translation and language models. In other words, a term would be translated using the most probable phrase alignment, which in most cases may not be in the domain specified by a user.

Another common terminology translation issue is the absence of terms in phrase-based SMT translation models. The lack of language (terminology) resources causes the "so-called" missing terminology to be ignored and not translated (i.e., the output is the same as the input). This issue can be solved if SMT systems provide a runtime integration with existing terminology databases or terminology collections provided by users. Such research has already been proposed, for instance, the popular Moses SMT platform allows the pre-processing of the translatable content during translation by providing possible translation equivalents for phrases. Carl and Langlais (2002) in their research showed that using terminology dictionaries in such a way could increase the translation performance for the English-French language pair. Babych and Hartley (2003) showed that for NE (namely, organisation names) special "*do-not-translate*" lists allowed increasing translation quality for the English-Russian language pair using a similar pre-processing technique that restricts translation of identified phrases. However, such approaches have been investigated either for languages with simple morphology or categories of phrases that are rarely translated or even left untranslated (e.g., many company and organisation names). A recent study in the FP7 project TTC (2013) has shown that for English-Latvian the pre-processing does not yield positive results for term translation. Hálek et al. (2011) also showed that the translation performance with on-line pre-processing drops according to BLEU for English-Czech named entity translation. This proves that the method is not stable when translating into morphologically rich languages, or the languages with the high level of inflection (e.g., the Baltic and Slavic languages). For such languages the task of terminology translation would also require a morphological synthesiser to be integrated into an SMT system in order to synthesise the correct inflected word form (or word forms for multiword terms) in case a morphologically rich language is used as the target language.

There has been research done in terminology translation and in usage of user-provided terminology, in particular. For instance, Itagaki and Aikawa (2008) proposed a module called "Term Swapper" that operated as a wrapper around an SMT system. Okuma et al. (2008) proposed a method for term substitution with high frequency terms from the training data and translation by analogy. The Moses SMT system also provides support for additional phrase table usage along with a general domain phrase table, as well as explicit user-specified translation of known phrases. The above mentioned methods show their potential. However, a certain adaptation of these methods is needed for morphologically rich languages.

Another way how to include terminology in phrase-based SMT is through a specific feature which indicates terms in a translation table (Pinnis and Skadiņš, 2012). Using additional phrase tables and explicit user-specified translations of known phrases is a general practice in SMT for different purposes (e.g., Chen and Eisele (2010) use it to create hybrid SMT systems). However, it is not explicitly used for integrating terminology in SMT systems.

If we focus on building a domain specific SMT engine, pooling together all available data (especially a significant portion of data that is out of the desired domain) can lead to negative changes in quality, since the out-of-domain training data will overwhelm the in-domain data (Koehn and Schroeder, 2007). Unfortunately, this drawback of domain specific SMT, when only in-domain data is used, is its failure to capture generalisations relevant to the target language. This can lead to poor translation quality (Thurmair, 2004).

A domain specific SMT engine needs to capture the generalisations of an engine trained on a large and sufficient supply of parallel data, yet not lose the crucial domain orientation. It was shown that to achieve this, an SMT engine can be trained on all available parallel data including out-of-domain data, and language model training data must be split into in-domain and out-of-domain sets, generating separate language models (LM) for each of the sets (Koehn and Schroeder, 2007; Lewis et al., 2010).

Although SMT domain adaptation has been an active field in the machine translation research community, the majority of practical SMT applications relay solely on collecting big amounts of domain specific corpora. Moreover, there are not so many even more advanced solutions which would focus on a special handling of terminology. It is assumed that training data will contain translations with terminology and SMT will learn accurate terminology from training data. However, it is not usually the case as training data, even if it is in the same domain, can contain contradicting terminology – industry or corporate specific synonyms in product- or vendor-biased terminology.

3 Proposed solution: Terminology services for SMT

3.1 Term extraction workflows for SMT

One of the prerequisites for accurate handling of terminology for an SMT system is its ability to identify terms in the translatable content. In this paper we propose to identify terms in SMT system training data (i.e., parallel and monolingual corpora used for the creation of models) and in the translatable content prior to translation (i.e., by pre-processing the text with existing terminology resources). These are two different steps, at which terminology integration in SMT systems and the availability of surrounding context (i.e., how much data is available – a phrase, a sentence, a full document etc.) have different requirements with respect to term identification and data processing speed. For instance, term identification in a large parallel corpus has to be fast and efficient and it has to be able to bilingually identify terms in the source language and in the target language content. Whereas term identification during translation requires just monolingual analysis in order to identify term candidates. Depending on the length of the available context, term identification can be context dependent or context independent. In order to satisfy the requirements, we propose two different term tagging workflows:

- Document level term tagging prior to translation is performed with statistically and linguistically motivated term extraction methods following Pinnis et al. (2012) in three steps. At first, term candidates are acquired using part-of-speech pattern filtering. Then, terms are weighed using different statistical association measures; the weights are normalised with the help of the TF*IDF (Spärck Jones, 1972) measure using reference corpora statistics (i.e., an inverse document frequency list calculated on a broad domain corpus). Finally, terms are tagged in the translatable content. In the proposed workflow we treat multiword term phrases as non-breakable phrases (i.e., phrases

that have to be translated by an SMT engine so that the reordering process would not break the phrases in multiple fragments).

- The second workflow performs sentence and phrasal level term tagging for SMT training as well as speed-critical sentence-by-sentence translation scenarios (e.g., commercial translation or high volume translation). Term tagging techniques slightly differ for parallel and monolingual data. For parallel data domain-specific bilingual term collections, which are acquired from the online terminology services (specified by users when training SMT systems), are transformed into transducers that identify bilingual terms in parallel sentences (or even phrases of an SMT system's translation model). For monolingual data (e.g., during translation or for monolingual corpora used during training), the transducers tag the terms identified in the text span. In the translation process, the transducers also provide translation equivalent candidates from bilingual term collections, thus ensuring that terminology is translated consistently (as required by a user). Although this method is able to identify terms present only in bilingual term collections, it is fast and it can be applied for text spans as short as one word, whereas the first method is applicable only when there is large enough context available from which to draw statistics.

3.2 Terminology integration into SMT

Terminology can be integrated in SMT systems in two levels – the training phase and the translation phase. In our proposed scenario (see Figure 1) online terminology services are used to acquire monolingual and bilingual term collections in order to adapt an SMT to specified domains in both levels. Terminology integration in SMT depends on the availability of data and other specific issues and (as elaborated below) may vary significantly.

The easiest method for bilingual terminology integration in SMT training is by adding the bilingual term collection to the parallel corpus that is used for training an SMT system. Although the size of the term collection usually is relatively small in comparison to the whole parallel corpus, namely the presence of a term collection in training data helps the SMT training engine to build better word and phrase alignments, and it also fills gaps in the vocabulary by allowing translation of previously unknown terms. In addition to this simple approach, we also propose to use online terminology services to tag terms in both parallel and monolingual corpora used in SMT training.

Following earlier work by Pinnis and Skadiņš (2012) we introduce an additional feature indicating phrases containing in-domain term translations in an SMT system's translation model. In order to do that, we use the phrasal level term tagging method as described in section 3.1. Using online terminology services we acquire a bilingual term collection from the corpus/corpora specified by a user and identify bilingual terms in SMT phrase tables. Our experiments building an English-Latvian SMT system in the mechanical engineering domain show that such an approach achieves a relative SMT quality improvement of up to 6% according to BLEU (Pinnis and Skadiņš, 2012). This method is also used to tag terms in a parallel corpus prior to building a phrase table. Theoretically, information about terms and their alignment might improve the phrase extraction process. However, we have not further investigated this path and left it for future work.

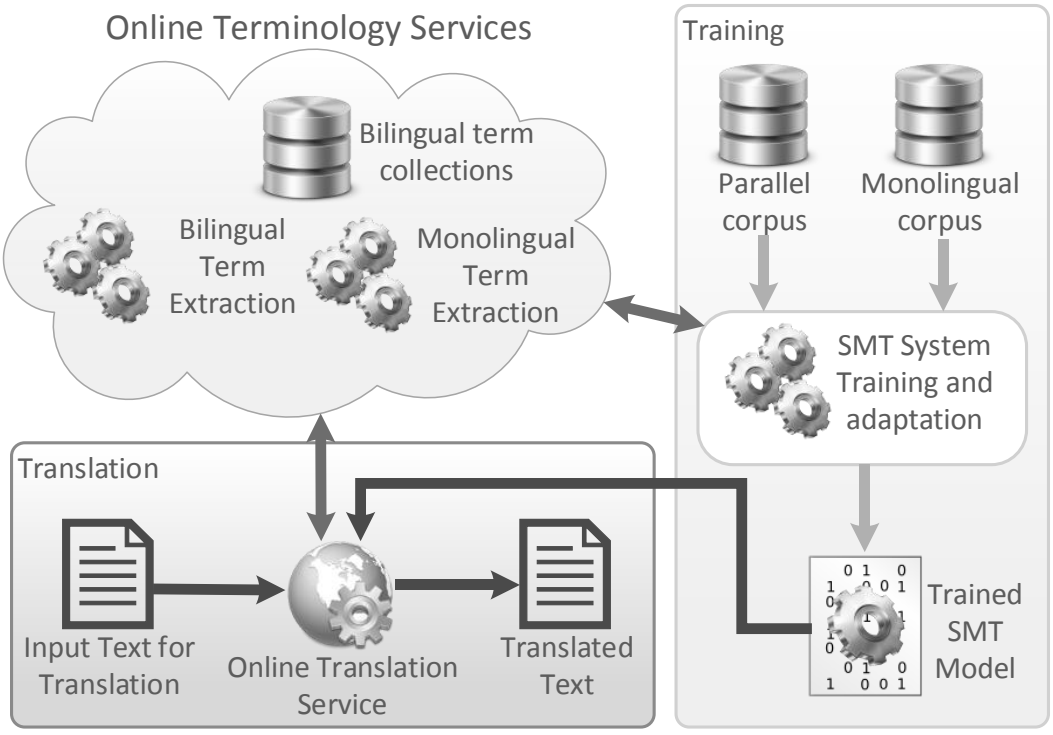

Figure 1. The conceptual design of the terminology service integration into SMT

As it was shown before (Koehn and Schroeder, 2007; Lewis et al., 2010) SMT system domain adaptation can also be achieved by using two language models – general and in-domain, and it is possible to select in-domain sentences automatically from a big general domain corpus (Moore and Lewis, 2010). This method works well if we have a reasonably big initial in-domain corpus. However, we show that, in cases when such a corpus is not available, it is possible to use bilingual term collections acquired from online terminology services. Such collections are then be used to select sentences containing domain specific terms from general domain corpora with the sentence level term tagging method described in section 3.1. Our experiments show that such an approach improves SMT quality by relative 35.6% over a baseline system according to BLEU (Pinnis and Skadiņš, 2012).

Besides the integration of online terminology services in the SMT system training phase, terminology services are also beneficial when used in the translation process. The translatable content is then pre-processed using term tagging methods described in section 3. In case if the input text is large enough (e.g., in the case of full documents, complete news articles etc.), a linguistically motivated term tagging method is applied in order to identify multi-word term phrases. Furthermore, the sentence level term tagging method is applied in order to find possible translation candidates for terms identified in the input text. Translation candidates are selected from bilingual term collections acquired from online terminology services. During translation the terminology annotation is used in the SMT decoder to limit reordering, so that multi-word terms are not split in multiple parts and reordered.

4 Conclusions

In this paper we have introduced the cloud-based terminology platform providing online terminology services for human and machine users to speed up and increase efficiency of terminology work. We have presented the use case for the application of online terminology services for SMT. We have described methods for terminology integration in both the SMT system training phase and the translation phase and outlined future work. Our experiments show that such an approach improves SMT quality over a baseline system according to the BLEU score.

The fully functional prototype of the platform is available for demonstration and testing. The demonstration workflow includes automatic ex-

traction of monolingual term candidates from documents uploaded by users, automatic lookup and extraction of translation equivalent term candidates from online term banks and in statistically aligned parallel and comparable data, and application of the created terminology collection in the Moses-based SMT.

Acknowledgement

The research within the project TaaS leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), Grant Agreement no 296312.

References

Babych, B., & Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition. *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*.

Bouamor, D., Semmar, N., & Zweigenbaum, P. (2012). Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pp. 674-679.

Carl, M., & Langlais, P. (2002). An intelligent Terminology Database as a pre-processor for Statistical Machine Translation. *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology*, Vol. 14, pp. 1-7.

Chen, Y., & Eisele, A. (2010). Integrating a Rule-based with a Hierarchical Translation System. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta, pp. 1746-1752.

Hálek, O., Rosa, R., Tamchyna, A., & Bojar, O. (2011). Named entities from Wikipedia for machine translation. *Proceedings of the Conference on Theory and Practice of Information Technologies (ITAT 2011)*, pp. 23-30.

Itagaki, M., & Aikawa, T. (2008). Post-MT Term Swapper: Supplementing a Statistical Machine Translation System with a User Dictionary. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, Morocco, pp. 1584-1588.

Koehn, P., Federico, M., Cowan, B., Zens, R., Duer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation, *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, pp. 177-180.

Koehn, P., & Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, pp. 224-227.

Lewis, W. D., Wendt, C., & Bullock, D. (2010). Achieving Domain Specificity in SMT without Overt Siloing. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pp. 2878-2883.

Moore, R. C., & Lewis, W. (2010). Intelligent selection of language model training data. *Proceedings of the ACL 2010 Conference Short Papers*, Uppsala, Sweden, pp. 220-224.

Nikoulina, V., Sandor, A., & Dymetman, M. (2012). Hybrid Adaptation of Named Entity Recognition for Statistical Machine Translation. *Proceedings of the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT (ML4HMT-12)*, Mumbai, India, pp. 1-16.

Okuma, H., Yamamoto, H., & Sumita, E. (2008). Introducing a translation dictionary into phrase-based smt. *IEICE transactions on information and systems*, 91(7), 2051–2057.

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., & Gornostay, T. (2012). Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012)*, Madrid, pp. 193–208.

Pinnis, M., & Skadiņš, R. (2012). MT Adaptation for Under-Resourced Domains – What Works and What Not. In A. Tavast, K. Muischnek, & M. Koit (Eds.), *Human Language Technologies – The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012* (Vol. 247, pp. 176–184). Tartu: IOS Press. Doi: 10.3233/978-1-61499-133-5-176.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, Vol. 28, pp. 11-21.

Thurmair, G. (2004). Comparing rule-based and statistical MT output. *Proceedings of the Workshop on the Amazing Utility of Parallel and Comparable Corpora*, Lisbon, Portugal, pp. 5-9.

TTC Project. (2013). Public deliverable D7.3: Evaluation of the impact of TTC on Statistical MT. *TTC Project: Terminology Extraction, Translation Tools and Comparable Corpora*. Retrieved from http://ttc-project.eu/images/stories/TTC_D7.3.pdf, p. 38.