A Lightweight Terminology Verification Service for External Machine Translation Engines

Alessio Bosca[†], **Vassilina Nikoulina**[‡], **Marc Dymetman**[‡]
[†]CELI, Turin, Italy
[‡]Xerox Research Centre Europe, Grenoble, France
[†]alessio.bosca@celi.it, [‡]{*first.last*}@xrce.xerox.com

Abstract

We propose a demonstration of a domain-specific terminology checking service which works on top of any generic black-box MT, and only requires access to a bilingual terminology resource in the domain. In cases where an incorrect translation of a source term was proposed by the generic MT service, our service locates the wrong translation of the term in the target and suggests a terminologically correct translation for this term.

1 Introduction

Today there exist generic MT services for a large number of language pairs, which allow relatively easily to make your domain-specific portal multilingual, and allow access to its documents for a broad international public. However, applying a generic MT service to domain-specific texts often leads to wrong results, especially relative to the translation of domain-specific terminology. Table 1 illustrates an example of a terminology inconsistent translation provided by a generic MT system.

English Source: Farmers tend to implement a broad non-focused **weed-control** strategy, on the basis of broad spectrum products and mixtures of different products.
Bing[1]: Los agricultores tienden a aplicar una estrategia amplia para **control de malezas** no centrado, sobre la base de productos de amplio espectro y las mezclas de diferentes productos.

Table 1: Example of the translation produced by a generic MT model for a domain-specific document. Source term : **weed-control**, official Spanish term translation: **control de malas hierbas**.

The importance of domain-specific terminology for Machine Translation has been mentioned in several previous works (eg. (Carl and Langlais, 2002; Skadins et al., 2013)). However, most of these works handle the case where the terminology is tightly integrated into the translation process. This requires both a good expertise in SMT and a large amount of both in-domain and generic parallel texts, which is often difficult, especially for low-resourced languages like Turkish or Estonian. Here, we are targeting the situation where the content provider is not willing to train a dedicated translation system, for some reason such as lack of technical skills or lack of necessary resources (parallel data or computational resources), but has at his disposal a multilingual in-domain terminology which could be helpful for improving the generic translation provided by an external translation service. We propose a demonstration of a multilingual terminology verification/correction service, which detects the wrongly translated terms and suggests a better translation of these terms. This service can be seen as an aid for machine translation post-editing focused on in-domain terminology and as a tool for supporting the workflow of practicing translators.

2 Related Work

There has recently been a growing interest for terminology integration into MT models. Direct integration of terminology into the SMT model has been considered, either by extending SMT training data (Carl and Langlais, 2002), or via adding an additional term indicator feature (Pinnis and Skadins, 2012; Skadins et al., 2013) into the translation model. However none of the above is possible when we deal with an external black-box MT service.

(Itagaki and Aikawa, 2008) propose a post-processing step for an MT engine, where a wrongly translated term is replaced with a user-provided term translation. The authors claim that translating the term directly often gives a different

Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 49–52,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

translation from the one obtained when translating the term in context: for English-Japanese the out-of-context term translation matches exactly the in context term translation in 62% of cases only. In order to address this problem the authors propose 15 simple context templates that induce the same term translation as the one obtained in the initial sentence context. Such templates include ”This is TERM” or ”TERM is a ...”. The main problem with this approach is that these templates are both language-pair and MT engine/model specific. Thus a certain human expertise is required to develop such templates when moving to a new language pair or underlying MT engine.

Our approach is close to the (Itagaki and Aikawa, 2008) approach, but instead of developing specific templates we propose a generic method for wrong terminology translation detection. We do not aim at producing the final translation by directly replacing the wrongly translated term — which can be tricky—, but rather perform the term correction in an interactive manner, where the user is proposed a better term translation and may choose to use it if the suggestion is correct.

3 Terminology-checking service

We assume that the provider of the terminology-checking service has a bilingual domain-specific terminology D at his disposal, which he wishes to use to improve the translation produced by a generic MT service MT. Our method verifies whether the terminology was translated correctly by the MT service (terminology verification), and if not, locates the wrong translation of the term and suggests a better translation for it.

3.1 Terminology checking

The basic terminology verification procedure applied to the source sentence s and to its translation $MT(s)$ by the generic service is done through the following steps:

1. For each term $T = (T_s, T_t)$ in D check whether its source part T_s is present in the source sentence s.

2. If s contains T_s, check whether the target part of the term T_t is present in the translation $MT(s)$. If yes, and the number of occurrences of T_s in s is equal to that of T_t in $MT(s)$: the term translation is consistent with terminological base. Otherwise, we attempt to locate the wrong term translation and suggest a better translation to the user.

Both steps require a sub-string matching algorithm which is able to deal with term detection problems such as morphological variants or different term variants. We describe the approach we take for efficient sub-string matching in more detail in section 3.3.

3.2 Terminology correction

Once we have detected that there is a source term T_s which has been incorrectly translated we would like to suggest a better translation for this term. This requires not only knowing a correct translation T_t of the source term T_s, but also its position in the target sentence. To do that, we need to identify what was the incorrect translation proposed by the MT engine for the term and to locate it in the translation $MT(s)$.

This can be seen as a sub-problem of the word-alignment problem, which is usually solved using bilingual dictionaries or by learning statistical alignment models out of bilingual corpora. However, in practice, these resources are not easily available, especially for low-resourced language pairs. In order to be able to locate the wrong term translation in the target sentence without resorting to such resources, our approach is to rely instead on the *same* external MT engine that was used for translating the whole source sentence in the first place, an approach also taken in (Itagaki and Aikawa, 2008).

To overcome the problem mentioned by (Itagaki and Aikawa, 2008) of non-matching out-of-context terms translations we propose to combine out-of context term translation ($MT(T_s)$) and context-extended term translation, as follows:

- Translate the term T_s extended with its left and/or right n-gram context: $s_{i-n}s_{i-n+1}...T_s...s_{j+n-1}s_{j+n}$, where $T_s = s_i...s_j$;

- Find a fuzzy match in $MT(s)$ for the translation of the context-extended term $MT(s_{i-n}...T_s...s_{j+n})$ using the same sub-string matching algorithm as in the terminology verification step.

Various combinations of out-of-context term translation ($MT(T_s)$) and n-extended term translation ($MT(s_{i-n}...T_s...s_{j+n})$) are possible.

The term location is performed in a sequential way: if the wrong term translation was not located after the first step (out-of-context translation), attempt the following step, extending size of the context (n) until the term is located.

3.3 Implementation

The implementation of the terminology-checking service that we demonstrate exploits Bing Translator[2] as SMT service, refers to the Agriculture domain and supports two terminology resources: the multilingual ontology from the Organic.Edunet portal[3] and Agrovoc, a multilingual theasurus from FAO[4]. The presented prototype enables terminology checking for all the language pairs involving English, French, German, Italian, Portoguese, Spanish and Turkish.

The component for matching the textual input (i.e. either the source or the translation from the SMT service) with elements from domain terminologies is based on the open source search engine Lucene[5] and exploits its built-in textual search capabilities and indexing facilities. We created a search index for each of the supported languages, containing the textual representations of the terminology elements in that language along with their URI (unique for each terminology element). The terms expressions are indexed in their original form as well as in their lemmatized and POS tagged ones; for Turkish, resources for morphological analysis were not available therefore stemming has been used instead of lemmatization.

In order to find the terminological entries within a textual input in a given language a two-steps procedure is applied:

- In a first step, the text is used as a query over the search index (in that language) in order to find a list of all the terminology elements containing a textual fragment present in the query.
- In a second step, in order to retain only the domain terms with a complete match (no partial matches) and locate them in the text, a new search index is built in memory, containing a single document, namely the original textual input (lemmatized or stemmed according to the resources available for that specific language). Then the candidate terminology elements found in the first step are used as queries over the in-memory index and the “highlighter” component of the search engine is exploited to locate them in the text (when found). A longest match criterion is used when the terminology elements found refer to overlapping spans of text.

Following this procedure a list with terminology elements (along with their URIs and the position within the text) is generated for both the source text and its translation. A matching strategy based on the URI allows to pair domain terms from the two collections. For domain terms in the source text without a corresponding terminology element in the translated text, the ”wrong” translation is located in the text according to the approach described in 3.2. The domain term is retranslated with the same SMT (with context extension, if needed) in order to obtain the ”wrong” translation and the translated string is located within the translation text with the same approach used in the second step of the procedure used for locating terminological entries (with an in-memory search index over the full text and the fragment used as query).

The service outputs two lists: one containing the pairs of terminology elements found both in the source and in the translation and another one with the terminology elements without a ”correct” translation (according to the domain terminology used) and for each of those an alternative translation from the domain terminology is proposed. In our demonstration a web interface allows users to access and test the service.

4 Proof of concept evaluation

In order to evaluate the quality of locating the wrong term translation, we applied the terminology verification service to an SMT model trained with Moses (Hoang et al., 2007) on the Europarl (Koehn, 2005) corpus. This SMT model was used for translating a test set in the Agricultural domain from Spanish into English. In these settings we have access to the internal sub-phrase alignment provided by Moses, thus we know the exact location of the wrong term translation, which allows us to evaluate how good our locating technique is.

The test set consists of 100 abstracts in Spanish from a bibliographical database of scientific publications in the Agriculture domain. These abstracts were translated into English with our translation

[2] http://www.bing.com/translator

[3] http://organic-edunet.eu/

[4] http://aims.fao.org/standards/agrovoc/about

[5] https://lucene.apache.org

model, and we then applied terminology verification and terminology correction procedures to these translations.

When applying terminology verification we detected in total 171 terms in Spanish, 71 of them being correctly translated into English (consistent with terminology), and 100 being wrongly translated (not consistent with terminology).

We then attempted to locate these wrongly translated terms in the system translation $MT(s)$.

Matching the out-of-context term translation with initial translation allowed to find a match for 82 wrongly translated terms (out of 100); Matching 1 left/right word extended term translation ($MT(w_{i-1}T_s w_{j+1})$) allowed to find a match for 16 more terms (out of 18 left).

Using the internal word alignments provided by Moses, we also evaluated how precisely the borders of the wrongly translated term were recovered by our term location procedure. This precision is measured as follows:

- The target tokens identified by our procedure (as described in 3) are: $g_T = t_1, \ldots, t_j$;
- We then identify the reference target tokens corresponding to the translation of the term T_s using the Moses word alignment : $r_T = \{r_{t_1}, \ldots, r_{t_k}\}$.

We define term location precision p as $p = \frac{|t_j \in r_T \cap g_T|}{|g_T|}$. The precision of term location with out-of-context term translation is of 0.92; the precision of term location with context-extended term translation is 0.91.

Overall, our approach allows to match 98% of the wrongly translated terms, with an overall location precision of 0.91. Although these numbers may vary for other language pairs and other MT systems, this performance is encouraging.

5 Conclusion

We propose a demonstration of a terminology verification system that can be used as an aid for post-editing machine translations explicitly focused on bilingual terminology consistency. This system relies on an external black-box generic MT engine extended with available domain-specific terminology. The location of the wrong term translation is located via re-translation of the original term with the same MT engine. We show that we partially overcome the situation where the out-of-context translation of the term differs from the original translation of this term (in the full sentence) by extending the term context with surrounding n-grams. The terminology verification method is both MT engine and language independent, does not require any access to the internals of the MT engine used, and is easily portable.

Acknowledgments

This work has been partially supported by the Organic.Lingua project (http://www.organic-lingua.eu/), funded by the European Commission under the ICT Policy Support Programme.

References

Michael Carl and Philippe Langlais. 2002. An intelligent terminology database as a pre-processor for statistical machine translation. In *COLING-02: Second International Workshop on Computational Terminology*, pages 1–7.

Hieu Hoang, Alexandra Birch, Chris Callison-burch, Richard Zens, Rwth Aachen, Alexandra Constantin, Marcello Federico, Nicola Bertoldi, Chris Dyer, Brooke Cowan, Wade Shen, Christine Moran, and Ondej Bojar. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007 Demo and Poster Sessions*, pages 177–180.

Masaki Itagaki and Takako Aikawa. 2008. Post-mt term swapper: Supplementing a statistical machine translation system with a user dictionary. In *LREC*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit X*, pages 79–86, Phuket Thailand.

Marcis Pinnis and Raivis Skadins. 2012. Mt adaptation for under-resourced domains - what works and what not. In *Baltic HLT*, volume 247, pages 176–184.

Raivis Skadins, Marcis Pinnis, Tatiana Gornostay, and Andrejs Vasiljevs. 2013. Application of online terminology services in statistical machine translation. In *MT Summit XIV*, pages 281–286.