Sensible: L2 Translation Assistance by Emulating the Manual Post-Editing Process

Liling Tan, Anne-Kathrin Schumann, Jose M.M. Martinez and Francis Bond[1]
Universität des Saarland / Campus, Saarbrücken, Germany
Nanyang Technological University[1] / 14 Nanyang Drive, Singapore
alvations@gmail.com, anne.schumann@mx.uni-saarland.de,
j.martinez@mx.uni-saarland.de, bond@ieee.org

Abstract

This paper describes the Post-Editor Z system submitted to the L2 writing assistant task in SemEval-2014. The aim of task is to build a translation assistance system to translate untranslated sentence fragments. This is not unlike the task of post-editing where human translators improve machine-generated translations. Post-Editor Z emulates the manual process of post-editing by (i) crawling and extracting parallel sentences that contain the untranslated fragments from a Web-based translation memory, (ii) extracting the possible translations of the fragments indexed by the translation memory and (iii) applying simple cosine-based sentence similarity to rank possible translations for the untranslated fragment.

1 Introduction

In this paper, we present a collaborative submission between *Saarland University* and *Nanyang Technological University* to the L2 Translation Assistant task in SemEval-2014. Our team name is *Sensible* and the participating system is Post-Editor Z (PEZ).

The L2 Translation Assistant task concerns the translation of an untranslated fragment from a partially translated sentence. For instance, given a sentence, "*Ich konnte Bärbel noch on the border in einen letzten S-Bahn-Zug nach Westberlin setzen.*", the aim is to provide an appropriate translation for the underline phrase, i.e. *an der Grenze*.

The aim of the task is not unlike the task of post-editing where human translators correct errors provided by machine-generated translations. The main difference is that in the context of post-editing the source text is provided. A translation workflow that incorporates post-editing begins with a source sentence, e.g. "*I could still sit on the border in the very last tram to West Berlin.*" and the human translator is provided with a machine-generated translation with untranslated fragments such as the previous example and sometimes "fixing" the translation would simply require substituting the appropriate translation for the untranslated fragment.

2 Related Tasks and Previous Approaches

The L2 writing assistant task lies between the lines of machine translation and crosslingual word sense disambiguation (CLWSD) or crosslingual lexical substitution (CLS) (Lefever and Hoste, 2013; Mihalcea et al. 2010).

While CLWSD systems resolve the correct semantics of the translation by providing the correct lemma in the target language, CLS attempts to provide also the correct form of the translation with the right morphology. Machine translation tasks focus on producing translations of whole sentences/documents while crosslingual word sense disambiguation targets a single lexical item.

Previously, CLWSD systems have tried distributional semantics and string matching methods (Tan and Bond, 2013), unsupervised clustering of word alignment vectors (Apidianaki, 2013) and supervised classification-based approaches trained on local context features for a window of three words containing the focus word (van Gompel, 2010; van Gompel and van den Bosch, 2013; Rudnick et al., 2013). Interestingly, Carpuat (2013) approached the CLWSD task with a Statistical MT system .

Short of concatenating outputs of CLWSD / CLS outputs and dealing with a reordering issue

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 541–545,
Dublin, Ireland, August 23-24, 2014.

and responding to the task organizers' call to avoid implementing a full machine translation system to tackle the task, we designed PEZ as an Automatic Post-Editor (APE) that attempts to resolve untranslated fragments.

3 Automatic Post-Editors

APEs target various types of MT errors from determiner selection (Knight and Chander, 1994) to grammatical agreement (Mareček et al., 2011). Untranslated fragments from machine translations are the result of out-of-vocabulary (OOV) words.

Previous approaches to the handling of untranslated fragments include using a pivot language to translate the OOV word(s) into a third language and then back into to the source language, thereby extracting paraphrases to OOV (Callison-burch and Osborne, 2006), combining sub-lexical/constituent translations of the OOV word(s) to generate the translation (Huang et al., 2011) or finding paraphrases of the OOV words that have available translations (Marton et al., 2009; Razmara et al., 2013). [1]

However the simplest approach to handle untranslated fragments is to increase the size of parallel data. The web is vast and infinite, a human translator would consult the web when encountering a word that he/she cannot translate easily. The most human-like approach to post-editing a foreign untranslated fragment is to do a search on the web or a translation memory and choose the most appropriate translation of the fragment from the search result given the context of the machine translated sentence.

4 Motivation

When post-editing an untranslated fragment, a human translator would (i) first query a translation memory or parallel corpus for the untranslated fragment in the source language, (ii) then attempt to understand the various context that the fragment can occur in and (iii) finally he/she would surmise appropriate translations for the untranslated fragment based on semantic and grammatical constraints of the chosen translations.

The PEZ system was designed to emulate the manual post-editing process by (i) first crawling a web-based translation memory, (ii) then extracting parallel sentences that contain the untranslated fragments and the corresponding translations of the fragments indexed by the translation memory and (iii) finally ranking them based on cosine similarity of the context words.

5 System Description

The PEZ system consists of three components, viz (i) a Web Translation Memory (WebTM) crawler, (ii) the XLING reranker and (iii) a longest ngram/string match module.

5.1 WebTM Crawler

Given the query fragment and the context sentence, "*Die Frau kehrte alone nach Lima zurück*", the crawler queries www.bab.la and returns sentences containing the untranslated fragment with various possible tranlsations, e.g:

- ***isoliert** : Darum sollten wir den Kaffee nicht **isoliert** betrachten.*
- ***alleine** : Die Kommission kann nun aber für ihr Verhalten nicht **alleine** die Folgen tragen.*
- ***Allein** : **Allein** in der Europischen Union sind.*

The retrieval mechanism is based on the fact that the target translations of the queried word/phrase are bolded on a web-based TM and thus they can be easily extracted by manipulating the text between <bold>...</bold> tags. Although the indexed translations were easy to extract, there were few instances where the translations were embedded betweeen the bold tags on the web-based TM.

5.2 XLING Reranker

XLING is a light-weight cosine-based sentence similarity script used in the previous CLWSD shared task in SemEval-2013 (Tan and Bond, 2013). Given the sentences from the WebTM crawler, the reranker first removes all stopwords from the sentences and then ranks the sentences based on the number of overlapping stems.

In situations where there are no overlapping content words from the sentences, XLING falls back on the most common translation of the untranslated fragment.

[1] in MT, evaluation is normally performed using automatic metrics based on automatic evaluation metrics that compares scores based on string/word similarity between the machine-generated translation and a reference output, simply removing OOV would have improved the metric "scores" of the system (Habash, 2008; Tan and Pal, 2014).

	en-de			en-es			fr-en			nl-en		
	acc	*wac*	*rec*	*acc*	*wac*	*rec*	*acc*	*wac*	*rec*	*acc*	*wac*	*rec*
WebTM	0.160	0.184	0.647	0.145	0.175	0.470	0.055	0.067	0.210	0.092	0.099	0.214
XLING	0.152	0.178	0.647	0.141	0.171	0.470	0.055	0.067	0.210	0.088	0.095	0.214
PEZ	**0.162**	**0.233**	**0.878**	**0.239**	**0.351**	**0.819**	**0.081**	**0.116**	**0.321**	**0.115**	**0.152**	**0.335**

Table 1: Results for *Best* Evaluation of the System Runs.

5.3 Longest Ngram/String Matches

Due to the low coverage of the indexed translations on the web TM, it is necessary to extract more candidate translations. Assuming little knowledge about the target language, human translator would find parallel sentences containing the untranslated fragment and resort to finding repeating phrases that occurs among the target language sentences.

For instance, when we query the phrase *history book* from the context "*Von ihr habe ich mehr gelernt als aus manchem **history book**.*", the longest ngram/string matches module retrieves several target language sentences without any indexed translation:

- *Ich weise darauf hin oder nehme an, dass dies in den Geschichtsbüchern auch so erwähnt wird.*
- *Wenn die Geschichtsbücher geschrieben werden wird unser Zeitalter, denke ich, wegen drei Dingen erinnert werden.*
- *Ich bin sicher, Präsident Mugabe hat sich nun einen Platz in den Geschichtsbüchern gesichert, wenn auch aus den falschen Gründen.*
- *In den Geschichtsbüchern wird für jeden einzelnen Tag der letzten mehr als 227 Jahre an Gewalttaten oder Tragdien auf dem europäischen Kontinent erinnert.*

By simply spotting the repeating word/string from the target language sentences it is possible to guess that the possible candidates for "*history book*" are *Geschichtsbücher* or *Geschichtsbüchern*. Computationally, this can be achieved by looking for the longest matching ngrams or the longest matching string across the target language sentences fetched by the WebTM crawler.

5.4 System Runs

We submitted three system runs to the L2 writing assistant task in Semeval-2014.

1. **WebTM**: a baseline configuration which outputs the most frequent indexed translation of the untranslated fragment from the Web TM.
2. **XLING**: reranks the WebTM outputs based on cosine similarity.
3. **PEZ**: similar to the XLING but when the WebTM fetches no output, the system looks for longest common substring and reranks the outputs based on cosine similarity.

6 Evaluation

The evaluation of the task is based on three metrics, viz. absolute accuracy (*acc*), word-based accuracy (*wac*) and recall (*rec*).

Absolute accuracy measures the number of fragments that match the gold translation of the untranslated fragments. Word-based accuracy assigns a score according to the longest consecutive matching substring between output fragment and reference fragment; it is computed as such:

$$wac = \frac{|longestmatch(output, reference)|}{max(|output|, |reference|)}$$

Recall accounts for the number of fragments for which output was given (regardless of whether it was correct).

7 Results

Table 1 presents the results for the *best* evaluation scores of the PEZ system runs for the English to German (en-de), English to Spanish (en-es), French to English (fr-en) and Dutch to English (nl-en) evaluations. Figure 1 presents the word accuracy of the system runs for both best and out-of-five (oof) evaluation[2].

The results show that using the longest ngram/string improves the recall and subsequently the accuracy and word accuracy of the system. However, this is not true when guessing untranslated fragments from L1 English to L2. This is due to the low recall of the system when searching for the untranslated fragment in French and

[2] Please refer to http://goo.gl/y9f5Na for results of other competing systems

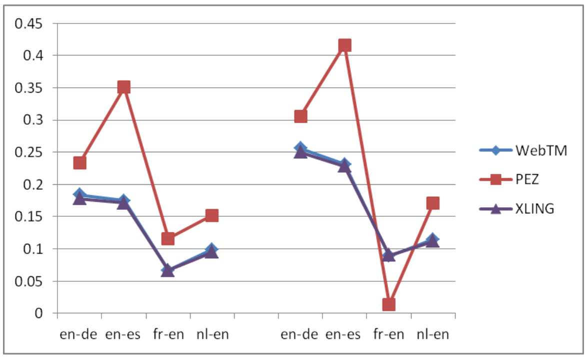

Figure 1: Word Accuracy of System Runs (*best* on the left, *oof* on the right).

Dutch, where the English words/phases indexed in the TM is much larger than other languages.

8 Error Analysis

We manually inspected the English-German outputs from the PEZ system and identified several particularities of the outputs that account for the low performance of the system for this language pair.

8.1 Weird Expressions in the TM

When attempting to translate *Nevertheless* in the context of "*Nevertheless hat sich die neue Bundesrepublik Deutschland unter amerikanischem Druck an der militrischen Einmischung auf dem Balkan beteiligt.*" where the gold translation is *Trotzdem* or *Nichtsdestotrotz*. The PEZ system retrieves the following sentence pairs that contains a rarely used expression *nichtsdestoweniger* from a literally translated sentence pair in the TM:

- **EN**: But *nevertheless* it is a fact that nobody can really recognize their views in the report.

- **DE**: Aber *nichtsdestoweniger* kann sich niemand so recht in dem Bericht wiederfinden.

Another example of weird expression is when translating "*husband*" in the context of "*In der Silvesternacht sind mein husband und ich auf die Bahnhofstraße gegangen.*". PEZ provided a lesser use yet valid translation *Gemahl* instead of the gold translation *Mann*. In this case, it is also a matter of register where in a more formal register one will use *Gemahl* instead of *Mann*.

8.2 Missing / Additional Words from Matches

When extracting candidate translations from the TM index or longest ngram/string, there are several matches where the PEZ system outputs a partial phrase or phrases with additional tokens that cause the disparity between the absolute accuracy and word accuracy. An instance of missing words is as follows:

- **Input**: Eine genetische Veranlagung *plays a decisive role*.

- **PEZ**: *Eine genetische Veranlagung eine entscheidende rolle*.

- **Gold**: *Eine genetische Veranlagung spielt (dabei) eine entscheidende rolle.*

For the addition of superfluous words is as follows:

- **Input:** *Geräte wie Handys sind not permitted wenn sie nicht unterrichtlichen Belangen dienen.*
- **PEZ**: *Geräte wie Handys sind es verboten, wenn sie nicht unterrichtlichen Belangen dienen.*
- **Gold**: *Geräte wie Handys sind verboten wenn sie nicht unterrichtlichen Belangen dienen.*

8.3 Case Sensitivity

For the English-German evaluation , there are several instances where the PEZ system produces the correct translation of the phrase but in lower cases and this resulted in poorer accuracy. This is unique to German target language and possibly contributing to the lower scores as compared to the English-Spanish evaluation.

9 Conclusion

In this paper, we presented the PEZ automatic post-editor system in the L2 writing assistant task in SemEval-2014. The PEZ post-editing system is a resource lean approach to provide translation for untranslated fragments based on no prior training data and simple string manipulations from a web-based translation memory.

The PEZ system attempts to emulate the process of a human translator post-editing out-of-vocabulary words from a machine-generated translation. The best configuration of the PEZ system involves a simple string search for the longest common ngram/string from the target language sentences without having word/phrasal alignment and also avoiding the need to handle word reordering for multi-token untranslated fragments.

Acknowledgements

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n ° 317471.

References

Marianna Apidianaki. 2013. Limsi : Cross-lingual word sense disambiguation using translation sense clustering. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 178–182, Atlanta, Georgia, USA, June.

Chris Callison-burch and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *In Proceedings of HLT/NAACL-2006*, pages 17–24.

Marine Carpuat. 2013. Nrc: A machine translation approach to cross-lingual word sense disambiguation (semeval-2013 task 10). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 188–192, Atlanta, Georgia, USA, June.

Nizar Habash. 2008. Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation. In *ACL*, pages 57–60.

Chung-Chi Huang, Ho-Ching Yen, Ping-Che Yang, Shih-Ting Huang, and Jason S. Chang. 2011. Using sublexical translations to handle the oov problem in machine translation. *ACM Trans. Asian Lang. Inf. Process.*, 10(3):16.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *AAAI*, pages 779–784.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. 2011. Two-step translation with grammatical post-processing. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 426–432, Stroudsburg, PA, USA.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *EMNLP*, pages 381–390.

Majid Razmara, Maryam Siahbani, Reza Haffari, and Anoop Sarkar. 2013. Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115, Sofia, Bulgaria, August.

Alex Rudnick, Can Liu, and Michael Gasser. 2013. Hltdi: Cl-wsd using markov random fields for semeval-2013 task 10. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 171–177, Atlanta, Georgia, USA, June.

Liling Tan and Francis Bond. 2013. Xling: Matching query sentences to a parallel corpus using topic models for wsd. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 167–170, Atlanta, Georgia, USA, June.

Liling Tan and Santanu Pal. 2014. Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, August.

Maarten van Gompel and Antal van den Bosch. 2013. Wsd2: Parameter optimisation for memory-based cross-lingual word-sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 183–187, Atlanta, Georgia, USA, June.

Maarten van Gompel. 2010. Uvt-wsd1: A cross-lingual word sense disambiguation system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 238–241, Stroudsburg, PA, USA.