Team Z: Wiktionary as a L2 Writing Assistant

Anubhav Gupta
UFR SLHS
Université de Franche-Comté
anubhav.gupta@edu.univ-fcomte.fr

Abstract

This paper presents a word-for-word translation approach using Wiktionary for SemEval-2014 Task 5. The language pairs attempted for this task were English-Spanish and English-German. Since this approach did not take context into account, it performed poorly.

1 Introduction

The objective of SemEval-2014 Task 5 is to translate a few words or a phrase from one language (L1) into another (L2). More specifically, a sentence containing primarily L2 and a few L1 words is provided, and the task is to translate the L1 words into the L2. This task is similar to the previous cross-linguistic SemEval tasks involving lexical substitution (Mihalcea et al., 2010) and word-sense disambiguation (Lefever and Hoste, 2013).

For example, consider the following sentence, written entirely in German except for one English word: *Aber auf diesem Schiff wollen wir auch Ruderer sein, wir sitzen im selben Boot und wollen mit Ihnen **row***. Here, the word **row** is polysemous and can be translated as the verb **rudern** or as the noun **Reihe** depending on context. The words to be translated can also form an idiomatic expression, such as **in exchange** in *die 1967 eroberten arabischen Gebiete **in exchange** gegen Frieden*. These examples reveal that this is not a straightforward task, as word-for-word translation may give inaccurate results.

Wiktionary is a multilingual dictionary containing word-sense, examples, sample quotations, collocations, usage notes, proverbs and translations (Torsten et al., 2008; Meyer and Gurevych, 2012). Since Wiktionary data have previously been used for translations (Orlandi and Passant, 2010), it was chosen for looking up the translation of source language (L1) words. However, the translation approach was word-for-word and ignored the target language (L2) context, i.e., the context in which the text fragment to be translated is found. The Wiktionary-based solution is for English-to-Spanish and English-to-German language translation though four language pairs were provided in this shared task.

2 Wiktionary

For a given word, the English version of Wiktionary gives not only its definition but also possible translations. The translations are divided based on part of speech (PoS) and word sense and at times also encode gender and number information. For example, the German and Spanish translations for the English word **book** are stored in Wiktionary as follows:

```
====Noun====
{{en-noun}}

=====Translations=====
{{trans-top|collection of sheets
of paper bound together
containing printed or written
material}}
* German: {{t+|de|Buch|n}}
* Spanish: {{t+|es|libro|m}}

{{trans-top|record of betting}}
* German: {{t|de|Wettliste|f}}
{{trans-top|convenient collection
of small paper items, such as
stamps}}
* German: {{t+|de|Album|n}}
* Spanish: {{t+|es|álbum|m}}

{{trans-top|major division of
a published work, larger than
```

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 624–627,
Dublin, Ireland, August 23-24, 2014.

```
a chapter}}

{{trans-top|script of a musical}}
* Spanish: {{t+|es|libreto|m}}

{{trans-top|usually in plural:
records of the accounts of
a business}}
* German: {{t+|de|Bücher|n-p}}

{{trans-top|ebook}}
* German: {{t+|de|E-Book|n}}

====Verb====
{{en-verb}}

=====Translations=====
{{trans-top|to reserve}}
* German: {{t+|de|buchen}},
{{t+|de|reservieren}}
* Spanish: {{t|es|reservar}}

{{trans-top|to write down,
register, record}}
* German: {{t+|de|notieren}},
{{t+|de|schreiben}}
* Spanish: {{t+|es|anotar}}

{{trans-top|to record the details
of}}
* {{ttbc|de}}: {{t+|de|bestrafen}}

{{trans-top|sports: to issue with
a caution}}

{{trans-top|slang: to travel
very fast}}
* German: {{t+|de|rasen}}
* {{ttbc|es}}: {{t|es|multar}}
```

The Wiktionary dump[1] is an XML file containing the word in the <title> tag and its description under the <text> tag. The translation of the word is indicated by {{t| or {{t+| followed by two letters to denote the target language (*es* for Spanish and *de* for German). This is followed by the translation and gender information in the case of nouns.

The information in Wiktionary was converted into a multidimensional hash table consisting of English words as key and PoS and translations in Spanish and German as the values. This table was used to look up the translations for the task.

Wiktionary also contains lists of the 10000 most frequent words in Spanish and of the 2000 most frequent words in German. This information was used to sort the target language words in the hash table in decreasing order of frequency. The translations absent from these frequency lists were kept in the hash table in the order that they were extracted from Wiktionary.

3 Translation

TreeTagger PoS	Wiktionary PoS
DT	Determiner, Article
NC, NN, NNS	Noun
IN, TO	Preposition
VB, VBG,VBZ, MD	Verb
RB, RBR, RP, WRB	Adverb
CD	Numeral
CC	Conjunction
PP, PRP, WP	Pronoun
JJ, JJS	Adjective

Table 1: PoS Mapping

The TreeTagger (Schmid, 1994) was used to parse the English (L1) phrases to obtain the PoS of each word along with the lemma. The PoS tags returned by the TreeTagger were mapped to the PoS used in Wiktionary as shown in Table 1. The word and its PoS were searched for in the hash table. If the translation was not found, then the lemma and its PoS were looked up. If the lemma lookup also failed then the phrase was not translated.

Once the L2 words were obtained for all the L1 words in the phrase, the L2 words were matched based on the gender and number information provided. For example, for the phrase **this question**, Wiktionary offered **este | m** and **esta | f** as Spanish translations of **this**, and **interrogante | m pregunta | f duda | f cuestión | f incógnita | f** for **question**. The translations were paired based on gender agreement rules (e.g. *este interrogante*, where both are masculine, and *esta pregunta*, where both are feminine) and provided as solutions.

3.1 Rules for English-to-Spanish Translation

Wiktionary only provides translations for the citation form of a word (even though other forms exist in WIktionary as valid entries), which is prob-

[1] For this task the 17 Dec 2013 version was used.

Language Pair	Dataset	Approach	Evaluation	Accuracy	Word Accuracy	Recall
en-es	Trial	Word-by-Word	Best	0.278	0.372	0.876
			Oof	0.382	0.471	0.876
		Word-by-Word + Rules	Best	0.340	0.434	0.844
			Oof	0.444	0.535	0.844
	Test	Word-by-Word	Best	0.200	0.308	0.785
			Oof	0.246	0.356	0.785
		Word-by-Word + Rules	Best	0.223	0.333	0.751
			Oof	0.277	0.386	0.751
en-de	Trial	Word-by-Word	Best	0.210	0.306	0.900
			Oof	0.316	0.422	0.900
	Test	**Word-by-Word**	Best	0.218	0.293	0.851
			Oof	0.307	0.385	0.851

Table 2: Performance of the System.

lematic when translating plural nouns or conjugated (finite) verbs. Lack of this inflectional information degraded the overall performance of both English-to-Spanish and English-to-German translations. Two rules were included in an attempt to improve the English-to-Spanish translations: (1) plural nouns and adjectives were formed by adding **-s** or **-es**, and (2) where a noun was preceded by an adjective in a L1 phrase, after the translation, the positions of the noun and the adjective were switched to respect the noun-adjective word order that is more commonly found in Spanish.

4 Results and Conclusions

Table 2 shows the performance of the system for the English-to-Spanish and English-to-German translations. The approach in bold was submitted for evaluation. The accuracy refers to the percentage of the fragments that were predicted accurately, whereas word accuracy measures the partially correct solutions. For each fragment, up to 5 translations could be submitted with one considered as the best answer and the rest regarded as alternatives. The *best* evaluation considered only the best answers. On the other hand, *oof* (out-of-five) evaluation considered the alternative answers to calculate the scores if the best answer was incorrect.

A context-independent, word-for-word translation approach to L2 Writing Assistant was proposed. The mediocre performance was due to the fact the approach was very basic. The system can be significantly improved by using the Spanish and German versions of Wiktionary to make up for the translations missing from the English version and by considering the L2 context provided. One such example in the German Wiktionary is the {{`Charakteristische Wortkombinationen`}} tag, which refers to the possible collocations. For example, one of the translations of the English word **exchange** in German is **Austausch**, which is most often collocated with **im** or **als**. Also, using a tool like JWKTL[2] would improve the quality of information extracted from Wiktionary.

References

Els Lefever and Véronique Hoste. 2013. SemEval-2013 Task 10: Cross-lingual Word Sense Disambiguation. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA.

Christian M. Meyer and Iryna Gurevych. 2012. Wiktionary: A New Rival for Expert-Built Lexicons? Exploring the Possibilities of Collaborative Lexicography. In *Electronic Lexicography*, edited by Sylviane Granger and Magali Paquot, 259–91. Oxford: Oxford University Press.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. Semeval 2010 Task 2: Cross-lingual Lexical Substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*. Uppsala, Sweden.

Fabrizio Orlandi and Alexandre Passant. 2010. Semantic Search on Heterogeneous Wiki Systems. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, 4:1–4:10. WikiSym ’10. New York, NY, USA: ACM.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of*

[2] https://code.google.com/p/jwktl/

International Conference on New Methods in Language Processing. Manchester, UK.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 20008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco.