CMPUT 650: A1 Bitext

Jai Riley
Department of Computing Science
University of Alberta
Edmonton, CA
jrbuhr@ualberta.ca

Mohammad Tavakoli
Department of Computing Science
University of Alberta
Edmonton, CA
tavakol5@ualberta.ca

Abstract

The goal of this assignment is to learn three essential semantic tasks: language translation, alignment and word sense disambiguation (WSD). Given a preprocessed English dataset tagged with senses, we translated sentences from English to Farsi, aligned the translated tokens with the original English tokens, and used the alignments to project senses to the new translations. Code is available from: https://github.com/jai-riley/CMPUT650_Project/tree/main/Assignment%201.

1 Machine Translation

The first step for this assignment is to extract the sentences from the given preprocessed dataset of English sentences and use machine translation to translate them into Farsi. Once the tokens from the given dataset were concatenated into their full sentences we performed the translation with the Google Translate API through the translators (version=5.8.9) python library with the following function call:

translators.translate_text(translator='google', query_text=s, from_language='en', to_language='fa'),

where s, is the sentence to be translated from English to Farsi. Once the sentences were translated into Farsi, a .txt file was created where each line consisted both English and Farsi versions of a sentence in the format:

English sentence ||| Farsi sentence.

2 Alignment

After translating the text from English to Farsi, the next step is to align the tokens given by the translation with their corresponding tokens from the original sentence. To accomplish this, an attempt was originally made to use AwesomeAlign. Unfortunately, when trying to compile the source code, we received an error that could not be solved and had to abandon this approach. Instead, we opted to use FastAlign (Dyer et al., 2013), an unsupervised aligner. To increase the performance of FastAlign, we retrieved supplementary parallel data (English to Farsi aligned datasets) from OPUS to compile with the given dataset. We combined the 'TED2020 v1' dataset (Reimers and Gurevych, 2020) to the original set of sentences and compiled FastAlign with this new dataset.

3 Sense Projection

Following the alignment of English sentences with their Farsi counterparts, we assigned senses to the words in the translated document. This was accomplished by extending the sense annotations onto the alignment links obtained in the preceding step using FastAlign. Our methodology involved projecting the BabelNet tag from the English word to its corresponding Farsi counterpart when the two tokens were aligned across translations.

4 Additional Datasets

As stated in the alignment section, we used the 'TED2020 v1' dataset to add parallel data to the main data. This dataset comprises 304,888 English sentences along with their translations into Farsi.

5 Example Errors Found

This section will describe examples of errors that were caught by visual inspection during each of the three steps involved in this process.

5.1 Translation Errors

Upon visual inspection of translations it was noticed that sometimes the translator struggled with choosing the right sense for the word to be translated. For example, there is an error in the translation of the third sentence:

> If you need more information about your medical condition or your treatment, read the Package Leaflet.

The translator translated "Package Leaflet" incorrectly as it detected the sense for the word Leaflet wrongly. The word-level translation for Leaflet is correct but within the context, it is not a good sentence-level translation. The translators also made some errors when translating verbs. As an illustration, in the sentence:

> It is actually not necessary to know MathML to use Kalgebra,

the translator mistakenly translates the verb ’to know’ into the verb ’to use’ in Farsi. Furthermore, there are instances where the translator fails to accurately translate verb tenses. For instance, in the sentence:

> For these purposes, Cerenia can be given for up to five days,

the translator incorrectly translates the tense of the verb ’to be’ from present tense into past tense ’has been’.

5.2 Alignment Errors

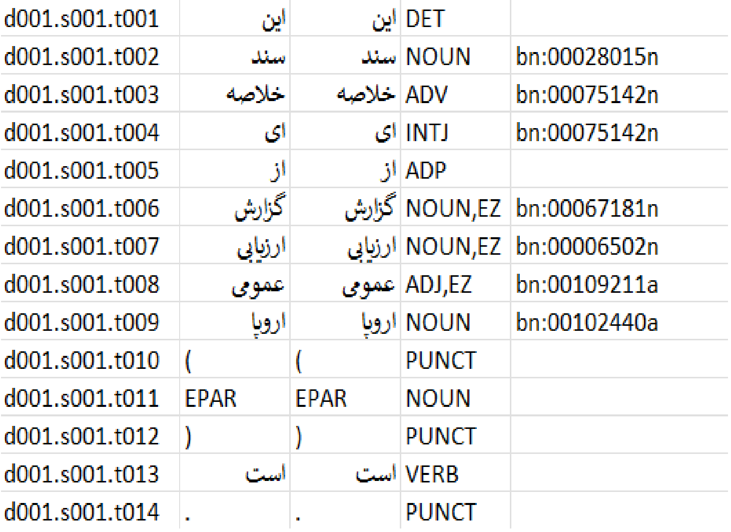
Even though we added the additional data to FastAlign, the output is not very accurate, creating many errors that carry forward into sense projection. Primarily we have notice the aligner struggles to correctly align verbs. This problem is mainly seen for verbs in which the meaning of the English verb requires multiple words in Farsi. In addition, sometimes the aligner aligns English verbs with Farsi prepositions.

5.3 Sense Projection Errors

In this section, errors are associated with alignment. If the alignment is not correct, it follows that the sense projection will also be incorrect. Therefore, the errors mentioned in the alignment section also apply here.

6 Results

We have saved the results in a CSV-formatted file. For example, the output for the first sentence is shown in Figure 1.



d001.s001.t001	این	این	DET	
d001.s001.t002	سند	سند	NOUN	bn:00028015n
d001.s001.t003	خلاصه	خلاصه	ADV	bn:00075142n
d001.s001.t004	ای	ای	INTJ	bn:00075142n
d001.s001.t005	از	از	ADP	
d001.s001.t006	گزارش	گزارش	NOUN,EZ	bn:00067181n
d001.s001.t007	ارزیابی	ارزیابی	NOUN,EZ	bn:00006502n
d001.s001.t008	عمومی	عمومی	ADJ,EZ	bn:00109211a
d001.s001.t009	اروپا	اروپا	NOUN	bn:00102440a
d001.s001.t010	(	(	PUNCT	
d001.s001.t011	EPAR	EPAR	NOUN	
d001.s001.t012	)	)	PUNCT	
d001.s001.t013	است	است	VERB	
d001.s001.t014	.	.	PUNCT	

Figure 1: The output for the first sentence

References

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proc. of NAACL*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.