A Method for the Assisted Translation of QA Datasets Using Multilingual Sentence Embeddings

Thomas Vakili
KTH alumnus
thomas@vakili.net

1 Introduction

In recent years, a number of large English datasets for various NLP tasks have been constructed and made publicly available. Similar datasets are often not available for lesser-resourced languages, but are crucial if state-of-the-art NLP results are to be realized in more languages.

Building a large dataset from the ground up is often infeasible due to resource constraints. Thus, translating already existing datasets to new languages could be a more feasible approach. However, automatic translations are often unreliable, especially for lesser-resourced languages. In this paper, we instead present a method for efficiently assisting the manual translation of a QA dataset.

SQuAD (Rajpurkar et al., 2016) is a QA dataset consisting of a large number of Wikipedia articles and a set of questions related to each article. More specifically, every datapoint in SQuAD consists of a question, its answer, and a paragraph from an article containing the answer. The task is to locate the answer to a question within a paragraph.

The method introduced in this paper was first put forward in Vakili (2020) and significantly reduces the time needed to translate SQuAD into another language. The scope of our evaluation is English-to-Swedish translation, but the method is general and could be applied to any Wikipedia language.

2 The Method

The method reduces the cost of translating SQuAD datapoints by searching for likely answers to English questions in Swedish Wikipedia articles.

Thus, we cannot translate a datapoint if its English Wikipedia article lacks a Swedish counterpart. Further, the question will only be answerable in the Swedish article if it overlaps with the English article to a sufficient degree.

When this is the case, multilingual sentence embeddings can be used to find good translation candidates. These are found by embedding the English sentence which answers a SQuAD question and then finding the most similar sentences in the relevant Swedish Wikipedia article.

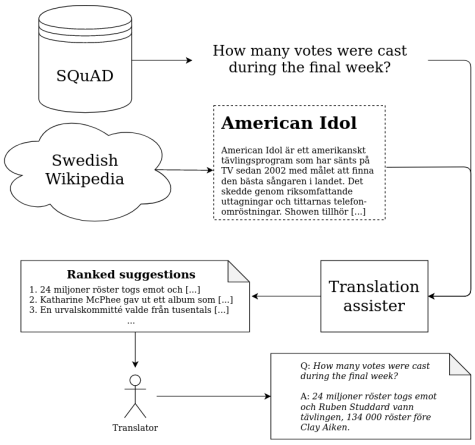

Figure 1: An overview of the proposed method. Here, a translator uses the method to pair an English question regarding the TV show *American Idol* with a Swedish answer from the show's Swedish Wikipedia article.

When the required content overlap is not present, a good Swedish answer will not exist. We therefore want to minimize the time wasted by looking for a good translation when one is unlikely to be found. This is accomplished by only suggesting the translation candidates that are the most similar to the English sentence.

3 Searching for Translation Candidates

LASER (Artetxe & Schwenk, 2019) is a sentence embedding toolkit which produces embeddings that are language-agnostic. This property was leveraged to find sentences in the Swedish Wikipedia articles that were semantically similar to English answers from SQuAD.

First, we extracted each English sentence containing the answer to a SQuAD question. Then,

Minimum relevance	Translations, minimum	Translations, maximum
perfect	2,830	6,777
similar	6,072	11,190
edge-case	8,342	14,028

Table 1: The estimated number of SQuAD questions for which a Swedish answer exists. The upper and lower bounds ($p > 0.95$) are different depending on the minimum relevance considered.

Max-rank	Labour reduction	Translations, minimum	Translations, maximum
5	94.92%	2,187	8,599
10	90.13%	2,358	10,170

Table 2: Results for max-ranks 5 and 10 are shown alongside the expected number of discovered translations based on the ground-truth. The minimum is based on the lower bound for perfect matches, and the maximum on the upper bound for edge-cases.

every Swedish Wikipedia article with an English counterpart in SQuAD was split into its constituent sentences. All of these English and Swedish sentences were embedded using LASER.

The English embeddings were then used to search for similar sentences among the Swedish embeddings. The Swedish sentences were ranked according to their cosine distances to each English embedding, and the best ranked Swedish sentences for each question were selected as the suggestions.

4 A Manually Annotated Ground-Truth

A ground-truth dataset was manually constructed from a randomly selected sample of 300 SQuAD questions from the 64,530 questions concerning articles available in both English and Swedish. The sample size was chosen to balance the time constraints imposed by having only one annotator (the author), and achieving results of a meaningful level of significance.

The annotator manually determined whether or not the answer to each English question could be found within its related Swedish article. When an answer could be found, the annotator recorded a relevance score reflecting how similar the Swedish answer was to its English counterpart. Out of the 300 randomly selected questions, 21 had perfectly matching answers, 18 semantically similar answers, and 12 had answers marked as edge-cases requiring a slight rephrasing of the question.

Using this sample, we could estimate the number of SQuAD questions that are answerable in the Swedish edition of Wikipedia using the Clopper-Pearson (1934) method (shown in Table 1).

5 Labour Reductions

The ground-truth dataset was used to estimate the amount of labour saved when translating SQuAD using the method.

Table 2 shows the impact of limiting the suggestions shown to a translator to the best-ranked results with the smallest distance to the English embedding. When only the top five or ten results are shown, the number of Swedish sentences that need to be reviewed is reduced by 94.92% and 90.13%, respectively. Meanwhile, we retain 73-83% of the existing translations when showing ten results and 61-77% when showing five results, with *edge-case* translations being the least-retained category.

Time was also saved by optimizing the processing order of the questions. The best translations will be found faster if questions that are likely to have good suggestions are processed early on. This is useful when, for example, an annotator is content with finding a fixed number of translations.

This optimization was achieved by sorting the processing order of the questions based on the cosine distances of their best-ranked suggestions. The impact of sorting was analyzed by simulating a translation of the dataset, keeping track of how many suggested translations had to be examined to reach a certain number of discovered translations. When the maximum rank was 10, a majority of all existing translations were found after examining only 2.4% of all sentences. Similar results were found for other configurations as well.

6 Conclusions

The results show that the method does significantly reduce the amount of labour required to translate SQuAD into Swedish. This reduction entails a certain amount of missed translations, and the parameters governing this trade-off have been explored.

Further research into reducing the amount of labour remains. Translating the SQuAD question into English is not tackled by this paper, nor is finding the short-form answer within the Swedish sentence containing it. These issues need to be solved to retain the same structure as SQuAD, either through manual processing or further research.

References

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv:1812.10464 [cs]*, arxiv 1812.10464. http://arxiv.org/abs/1812.10464

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*(4), 404–413. https://doi.org/10.2307/2331986

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. *arXiv:1606.05250 [cs]*, arxiv 1606.05250. http://arxiv.org/abs/1606.05250

Vakili, T. (2020). *A method for the assisted translation of QA datasets using multilingual sentence embeddings* (Master's thesis 2020:581). KTH, School of Electrical Engineering and Computer Science (EECS). http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-281826