FRI Natural language processing course 2022
Project report

University of Ljubljana
Faculty of Computer and Information Science

Cross-lingual question answering

Nace Gorenc, Jernej Vrhunc in Žan Pečovnik

Abstract

The abstract goes here. The abstract goes here.

Keywords

Keyword1, Keyword2, Keyword3 ...

Advisors: Slavko Žitnik

Introduction

In this project we will take a closer look at natural language processing, more precisely the area of question answering and machine translation. We begin with the overview of related work and short presentation of potentially viable solutions.

Lewis et al. [1] tried to evaluate different question answering strategies, they came to the idea that it would be interesting when searching the web for answers, that different target languages could increase the number of possible answers. Eventhough it is good that more potential answers is overall better, it is still preferable to have answers in the same language as the question was. They focused on bilingual question answering systems, processing questions in both French and English. They approached the problem from 2 angles. The first one was to translate each question to the target language and then process the question analysis in the target language. The second approach is a term-by-term translation meaning that the question is analysed in the same language and the answer is then translated to the target language. Both approaches have a major flaw which is that a lot of valuable information can be lost during the translation.

Ligozat et al. [2] created a new multi-way aligned extractive question answering evaluation benchamrk dataset. QA models have shown rapid progress due to new high quality benchmark datasets eventhough most of them are in English, since they are very costly to collect. This dataset has over 12000 instances in English language and between 5000 and 6000 instances in Arabic, German, Spanish, Hindi, Vietnamese and Simplified Chinese. The sentences gathered from Wikipedia articles were translated by professionals and annotated in the aligned contexts for target languages. Each instance has an aligned equivalent in multiple other languages (always including English). Comparison was made between multilingual BERT and XLM models, both trained in English on SQuAD as training dataset. It turns out that zero-shot XLM model transfers best, but the transfer results are far worse that training-language performance.

Ojokoh and Adebisi [3] reviewed multiple known frameworks for question answering, passage retrieval and answer extraction. They addressed some important issues associated with QA systems, which include question processing, question classes, cross-lingual and real time question answering. Finally, they classified QA systems based on some identifying criteria like application domain, question type and data source.

McCann et al. [4] introduced the Natural Language Decathlon, which is a challenge that spans ten tasks: question answering, machine translation, summarization, natural language inference, sentiment analysis, semantic role labeling, relation extraction, goal-oriented dialogue, semantic parsing, and commonsense pronoun resolution. They casted all tasks as question answering over a context and proposed a multitask question answering network (MQAN) which jointly learns all tasks in their challenge.

In the paper [5], the authors presented the Natural Question corpus, a question answering data set. It is said to be natural, because the dataset consists of real anonymous questions collected from the Google search engine. They set three

goals in their research:

- provide large scale end-to-end training data,
- provide dataset that drives research in natural language understanding,
- to study human performance in providing QA annotations.

The process of the method is as follows. An annotator is presented with a pair of question and the Wikipedia page. It returns also a pair of short answer and a long answer, which is usually a paragraph or a table from the Wikipedia page in HTML format. Both answers can be returned as null, if the annotator hasn't found any answer for the question that has been asked. They present the metrics that can be used with natural questions, for the purposes of evaluating the performance of question answering systems. The NQ corpus is designed to provide a benchmark with which we can evaluate the performance of the QA system. They represent the upper bound on these metrics and show that existing methods currently do not approach this upper bound.

Short dataset analysis

SQuAD 2.0 is an improvement of the SQuAD 1.1 dataset which had 107.785 question-answer pairs on 536 Wikipedia articles. In the SQuAD 1.1 dataset the correct answer to the question always existed somewhere, the models just had to pick the most probable answer instead of checking that the answer is entailed by the text, so they extended the new dataset with 53775 new unanswerable questions about the same paragraphs as in SQuAD 1.1, which were crafted by people in such way that they are relevant to the paragraph and that this paragraph contains a plausible answer. New dataset is significantly more challenging since the state-of-the-art models had a 66.3% F1 score on this dataset, while the same models had a 85.8% F1 score on the previous SQuAD 1.1 dataset, whereas human F1 score is 89.5% meaning that the new dataset is approximately 20% more challenging. Together there are now 151.054 questions on 505 articles, approximately a third of that is unanswerable.

Proposed methods

We suggest 2 different approaches to this problem. In both cases we would used the pre-trained BERT model for question answering and then evaluate both approaches.

1. Translate the question to English language and perform question analysis in the English language, translate answer back to source language
2. Translate the corpus to source language (e.g. Slovene) and fine-tune BERT on source language

Equations

You can write equations inline, e.g. $\cos\pi = -1$, $E = m \cdot c^2$ and α, or you can include them as separate objects. The Bayes's rule is stated mathematically as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1}$$

where A and B are some events. You can also reference it – the equation 1 describes the Bayes's rule.

Lists

We can insert numbered and bullet lists:

1. First item in the list.
2. Second item in the list.
3. Third item in the list.

- First item in the list.
- Second item in the list.
- Third item in the list.

We can use the description environment to define or describe key terms and phrases.

Word What is a word?.

Concept What is a concept?

Idea What is an idea?

Random text

This text is inserted only to make this template look more like a proper report. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam blandit dictum facilisis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Interdum et malesuada fames ac ante ipsum primis in faucibus. Etiam convallis tellus velit, quis ornare ipsum aliquam id. Maecenas tempus mauris sit amet libero elementum eleifend. Nulla nunc orci, consectetur non consequat ac, consequat non nisl. Aenean vitae dui nec ex fringilla malesuada. Proin elit libero, faucibus eget neque quis, condimentum laoreet urna. Etiam at nunc quis felis pulvinar dignissim. Phasellus turpis turpis, vestibulum eget imperdiet in, molestie eget neque. Curabitur quis ante sed nunc varius dictum non quis nisl. Donec nec lobortis velit. Ut cursus, libero efficitur dictum imperdiet, odio mi fermentum dui, id vulputate metus velit sit amet risus. Nulla vel volutpat elit. Mauris ex erat, pulvinar ac accumsan sit amet, ultrices sit amet turpis.

Phasellus in ligula nunc. Vivamus sem lorem, malesuada sed pretium quis, varius convallis lectus. Quisque in risus nec lectus lobortis gravida non a sem. Quisque et vestibulum sem, vel mollis dolor. Nullam ante ex, scelerisque ac efficitur vel, rhoncus quis lectus. Pellentesque scelerisque efficitur purus in faucibus. Maecenas vestibulum vulputate nisl sed vestibulum. Nullam varius turpis in hendrerit posuere.

Figures

You can insert figures that span over the whole page, or over just a single column. The first one, Figure 1, is an example of a figure that spans only across one of the two columns in the report.

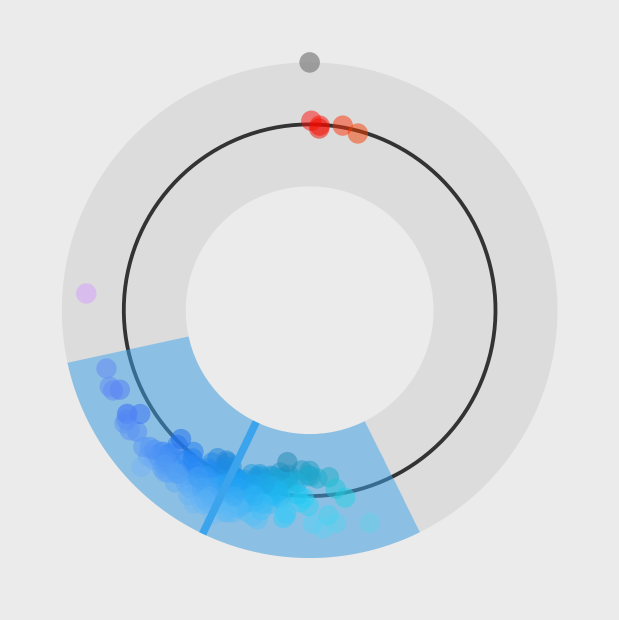

Figure 1. A random visualization. This is an example of a figure that spans only across one of the two columns.

On the other hand, Figure 2 is an example of a figure that spans across the whole page (across both columns) of the report.

Tables

Use the table environment to insert tables.

Table 1. Table of grades.

Name		
First name	Last Name	Grade
John	Doe	7.5
Jane	Doe	10
Mike	Smith	8

Code examples

You can also insert short code examples. You can specify them manually, or insert a whole file with code. Please avoid inserting long code snippets, advisors will have access to your repositories and can take a look at your code there. If necessary, you can use this technique to insert code (or pseudo code) of short algorithms that are crucial for the understanding of the manuscript.

Listing 1. Insert code directly from a file.

```
import os
import time
import random

fruits = ["apple", "banana", "cherry"]
for x in fruits:
  print(x)
```

Listing 2. Write the code you want to insert.

```
import(dplyr)
import(ggplot)

ggplot(diamonds,
           aes(x=carat, y=price, color=cut)) +
  geom_point() +
  geom_smooth()
```

Results

Use the results section to present the final results of your work. Present the results in a objective and scientific fashion. Use visualisations to convey your results in a clear and efficient manner. When comparing results between various techniques use appropriate statistical methodology.

More random text

This text is inserted only to make this template look more like a proper report. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam blandit dictum facilisis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Interdum et malesuada fames ac ante ipsum primis in faucibus. Etiam convallis tellus velit, quis ornare ipsum aliquam id. Maecenas tempus mauris sit amet libero elementum eleifend. Nulla nunc orci, consectetur non consequat ac, consequat non nisl. Aenean vitae dui nec ex fringilla malesuada. Proin elit libero, faucibus eget neque quis, condimentum laoreet urna. Etiam at nunc quis felis pulvinar dignissim. Phasellus turpis turpis, vestibulum eget imperdiet in, molestie eget neque. Curabitur quis ante sed nunc varius dictum non quis nisl. Donec nec lobortis velit. Ut cursus, libero efficitur dictum imperdiet, odio mi fermentum dui, id vulputate metus velit sit amet risus. Nulla vel volutpat elit. Mauris ex erat, pulvinar ac accumsan sit amet, ultrices sit amet turpis.

Phasellus in ligula nunc. Vivamus sem lorem, malesuada sed pretium quis, varius convallis lectus. Quisque in risus nec lectus lobortis gravida non a sem. Quisque et vestibulum sem, vel mollis dolor. Nullam ante ex, scelerisque ac efficitur vel, rhoncus quis lectus. Pellentesque scelerisque efficitur purus in faucibus. Maecenas vestibulum vulputate nisl sed vestibulum. Nullam varius turpis in hendrerit posuere.

Nulla rhoncus tortor eget ipsum commodo lacinia sit amet eu urna. Cras maximus leo mauris, ac congue eros sollicitudin ac. Integer vel erat varius, scelerisque orci eu, tristique purus. Proin id leo quis ante pharetra suscipit et non magna. Morbi in volutpat erat. Vivamus sit amet libero eu lacus pulvinar pharetra sed at felis. Vivamus non nibh a orci viverra rhoncus

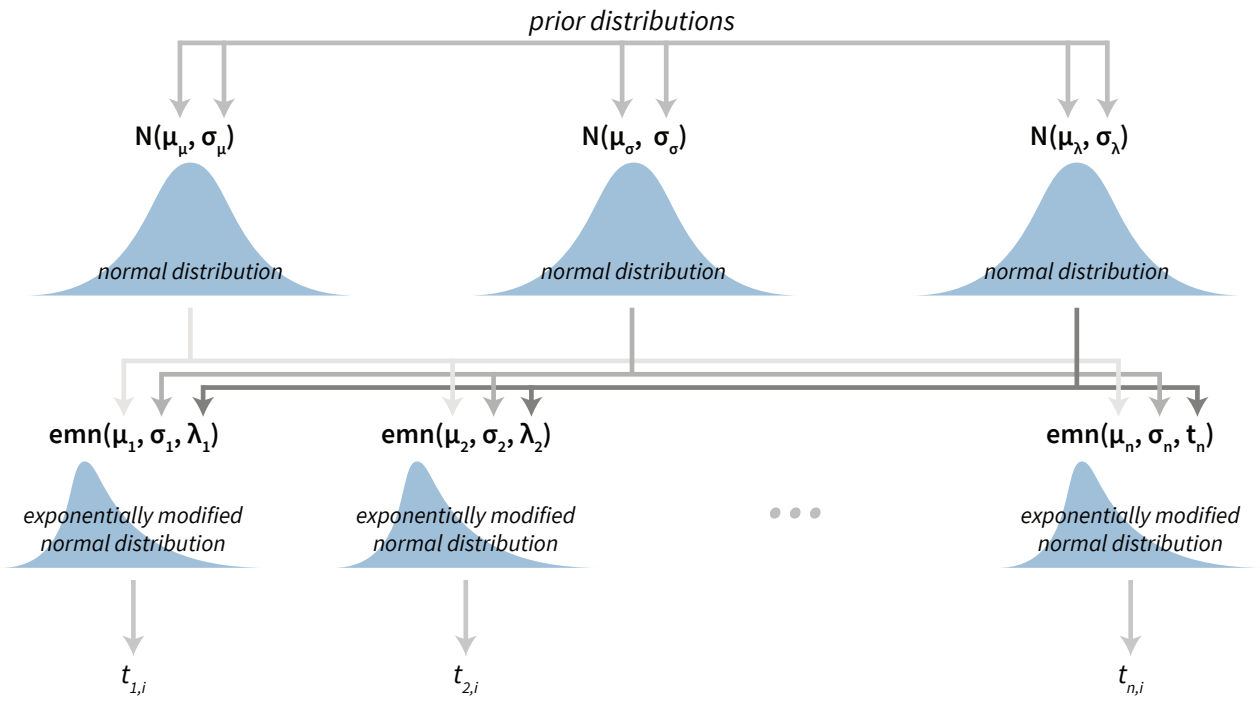

Figure 2. Visualization of a Bayesian hierarchical model. This is an example of a figure that spans the whole width of the report.

sit amet ullamcorper sem. Ut nec tempor dui. Aliquam convallis vitae nisi ac volutpat. Nam accumsan, erat eget faucibus commodo, ligula dui cursus nisi, at laoreet odio augue id eros. Curabitur quis tellus eget nunc ornare auctor.

Discussion

Use the Discussion section to objectively evaluate your work, do not just put praise on everything you did, be critical and exposes flaws and weaknesses of your solution. You can also explain what you would do differently if you would be able to start again and what upgrades could be done on the project in the future.

References

[1] Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*, 2019.

[2] Anne-Laure Ligozat, Brigitte Grau, Isabelle Robba, and Anne Vilnat. Evaluation and improvement of cross-lingual question answering strategies. In *Proceedings of the Workshop on Multilingual Question Answering at EACL Conference*, pages 23–30, Trento, Italy, May 2006.

[3] Bolanle Ojokoh and Emmanuel Adebisi. A review of question answering systems. *Journal of Web Engineering*, 17(8):717–758, 2018.

[4] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018.

[5] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.