**DIKU NLP PROJECT**

**Ellia Morse, Federico Fiorio,**

**Gonza Villalba**

**Abstract**

The aim of this project is to create a multilingual question answering system. The dataset used is the publicly available TyDi QA dataset, a set of question-document-answer items. This project covered English, Finnish, and Japanese. The project built up through the duration of the course following in-class lessons.

1. **Introduction and Work Distribution**

This project was completed by Ellia Morse, Federico Fiorio, and Gonza Villalba. While everyone collaborated on each section, Federico was responsible for weeks 1 and 2, Gonza was responsible for week 3, and Ellia was responsible for the report.

1. **Introduction to NLP (5-11 Sept)**

**2.1 Preprocessing and Data Analysis**

The goal of this section was to create a preprocessing pipeline to tokenize instances of English, Finnish, and Japanese at the word level, and familiarise ourselves with the data by understanding what tokens usually begin and end questions. We utilised the Nltk library for English and Finnish, and Janome for Japanese. Due to the difference in characters, a specialised tool would be better fitting for more accurate tokenization.

While reviewing the data, we found that there were words most common for starting and ending a sentence in each language.

English: ['?', 'the', 'When', 'was', 'What', 'is', 'of', 'How', 'in', 'did']

Finnish: ['?', 'on', 'Milloin', 'Mikä', 'Missä', 'Kuka', 'oli', 'Mitä', 'syntyi', 'kuoli']

Japanese: ['は', 'の', '？', 'た', 'い', 'つ', '何', 'し', 'どこ', 'が']

**2.2 Binary Question Classification**

When designing a classifier, we originally attempted bag-of-words. Our goal was for the model to output “1” if a question was answerable or “0” if it was not. However, when predicting whether a question was answerable, the model consistently predicted exactly 50% answerability. Instead, we used tfidf, or term frequency–inverse document frequency. This model creates a statistic to reflect the importance of a word to the overall collection of data.

This gave us better accuracy. Results below are rounded to the hundredth decimal.

English: training 89.2%, validation 72.42%

Finnish: 91.69%, validation 71.23%

Japanese: 98.33%, validation 57.92%

The results for our Japanese model predicted that a question was answerable significantly more often. However, the accuracy was about four times worse.

Perhaps the problem with the Japanese model was the tokenization process with tfidf, as we had issues implementing the Janome library.

1. **Representation Learning (12-18 Sept)**

The goal of this week was to explore different methods of representation. We implemented an extension to our binary classifier. Rather than full sentences, this extension featured continuous vector representations of words. We used a continuous bag of words method that trained the model on individual words, and then averaged them to obtain a sentence representation. The input layer contained context words, output layer contained the current word, and the hidden layer had the number of dimensions.

Compared to our previous model, our results for English tokenisation were training set accuracy of 86.32% and validation set accuracy of 72%. These numbers were a slightly better match than previously.

When using just a continuous bag of words method in English, the training set accuracy was 69.58% accuracy and the validation set accuracy was 65.46% accurate. While these numbers were an even closer match, the overall accuracy was significantly lower.

1. **Language Modelling (18-25 Sept)**

During this week, the objective was to implement different language models and compare their outcomes. Specifically, we aimed to observe the different predictions they offered and attempted to identify the reasons if and why different models structured predicted sentences inconsistently. Additionally, we attempted to identify how the different hyperparameters affected each model’s predictions.

Initially, we implemented pretrained models for each of the three required languages. We focused on the models *roberta-base-squad2*, *deberta-base-japanese-aozora-ud-head*, and *bert-base-finnish-cased-squad2-fi* for Enlgish, Japanese, and Finnish, respectively. These three were the most popular language models for each individual language on Hugging Face, a platform recommended by the DIKU NLP instructors. However, we contrasted the aforementioned models with additional ones not listed on Hugging Face, particularly for the Finnish language as the options were limited in said platform.

Currently, we have run into problems with implementing the stated models with our dataset using the *transformers* package. Thus, our next step is to attempt the implementation using the *haystack* package.

1. **Error Analysis and Interpretability (26 Sept - 2 Oct)**
2. **Sequence Labelling (3-9 Oct)**
3. **Multilingual QA (10 Oct +)**

**Footnotes**

Footnotes are inserted using Insert / Footnote… URLs should be added as Hyperlinks and formatted in 10pt Courier New font w

ithout underlining.

* 1. **Figures and tables**

**Creating:** To create a new Figure or Table, insert a Text Box where you want it to appear (generally, centred at the top of a column close to where it is referred to) and then fill it in with the Figure (or Table). Highlight and right click to add Caption, with the ACL Caption style (or ACL Caption Long style for multi-line captions), which places 10 pt below and above the caption.

**Numbering:** To update numbering, highlight all the relevant text (*e.g.*, **Ctrl-A + F9**). This will update all the numbering applicable to tables, figures, equations, and headings.

**Cross-referencing:** To add a cross reference to a figure or table:

* Place the mouse pointer at the lo
* cation where you wish to add the cross-reference.
* Click on the **Insert** menu, (then click **Reference**), and then **Cross-reference** in the **Links** panel.
* In the **Cross-reference** dialog box, click the caption to which you are building the text reference.
* For a figure, under **Reference Type**, click **Figure**.
* Under Insert Reference To, click Only Label and Number, then click OK.
* As much as possible, fonts in figures should conform to the document fonts (this is not the case in the example figure).

|  |
| --- |
| Figure 1: A figure with a caption that runs for more than one line**.** |

This is an example reference to Figure 1.

* 1. **Hyperlinks**

Within-document and external hyperlinks are indicated with Dark Blue text, Color Hex #000099.

* 1. **References**

To create hyperlinks between citations and references, as you insert each full reference in the References section, highlight it and then select Insert, Bookmark. Link back to the reference from its citations in the text by highlight the citation, right clicking, and selecting Insert, Cross-Reference, then selecting the Bookmark you’ve saved. Highlight the citation again to give make it dark blue (included in this theme), if it is not automatically applied. If there are problems saving the hyperlinks when you convert the document to PDF, use an online converter such as <http://go4convert.com>.

* 1. **Citations**

Citations can be created by creating in-document hyperlinks to bookmarks you’ve created. Go to Insert / Hyperlink / This Document / Bookmarks, and select your bookmark.

* 1. **Equations**

An example equation is shown below:

(1)

To add new equations, authors are encouraged to copy this existing equation line, and then replace with the new equation. The numbering and alignment of equation line elements is automatic. To update equation numbering, press **Ctrl-A + F9**. Note: this will only update the number to the right of the equation; to update numbering within the text you must create a cross-reference.

**Cross-referencing:** To create a cross-reference for an equation:

* Create a bookmark for it.
* Select the number to the right of the equation. Go to **Insert**, **Bookmark** (in the **Links** panel),andthen create a name for your equation. Press **Add** to create the bookmark.
* To refer back, place the mouse pointer at the location where you wish to add the cross reference.
* Go to **Insert, Cross-reference** (in the **Links** panel).In the dialogue box, select **Bookmark** and **Bookmark Text** from each dropdown list. Uncheck **Insert as Hyperlink**, then click **OK**.
* This will make it such that whenever a new equation is added, the references to the equation will be updated when **Ctrl-A + F9** is pressed.
* This an example cross-reference to Equation 1.

* 1. **Appendices**

Appendices, if any, directly follow the text and the

references. Letter them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

1. **MS Word** **STREAM Tools**

This Microsoft Word file was updated in 2016 with STREAM Tools, designed for creating well-formatted reports and papers with Microsoft Word (Mamishev, 2010; Mamishev, 2013).

**Acknowledgments**

An example acknowledgment.

**References**

Alfred. V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling, volume 1*. Prentice-Hall, Englewood Cliffs, NJ.

1. **Appendices**

Appendices are added after the References section by restarting the header numbering using style “A, B, C”.

1. **Supplementary Material**

Supplementary material also be included with the Appendices.