Master Thesis Proposal: An end-to-end Acronym Expander System for Dutch

Jesher Appels (12865362, jesher.appels@student.uva.nl)

1 INTRODUCTION

Every project or research that extensively utilizes data requires some form of expository data analysis (EDA) to get a good understanding of the data that is being used, as well as, to verify the data quality. This document describes the EDA and data gathering processes for a Dutch end-to-end acronym expander system called AcX (Acronym eXpansion). There are two ways to extract the meaning of an acronym: in-expansion and out-expansion. With in-expansion, the meaning of a particular acronym is present in the document, whereas with out-expansion, the meaning has to be derived based in the context. Out-expansion in the AcX system utilizes sentiment analysis techniques to learn the context in which the acronym is used.

The main aim of this research is to test and tweak the AcX system for the Dutch language, see if language specific transformers like BERTje, BERT-NL, and ULMFiT generate any additional performance and if additional pre-processing steps yield even better results. The research question is formulated as follows:

To what extend can an end-to-end system for acronym expansion in the Dutch language be built upon AcX, and how does its performance for Dutch improve with pre-trained transformers and additional pre-processing steps in the out-expansion part of the system?

The document will not include the results of the EDA, simply because the data must be label first, and this will be done in the upcoming weeks. However, it will describe the data gathering and verification process. As well as an analysis on a subset of the documents that will be used and a comparable alternative.

2 METHODOLOGY

Data requirements

There is one general requirement for the data that will be used in the project, and that is that it has to be as diverse as possible. Literature shows that much of the research from recent years predominantly focuses on the extraction and identification of acronyms for a specific field, such as Medical [1, 2, 3] or Computer science. The creators of the AcX system aim to make their system for a broad range of topics and do not want to have a particular bias towards a specific field, meaning generalizability is essential. For this reason, the documents will not be constrained to specific topics. Furthermore, the data will be gathered from different sources, including Wikipedia articles, news articles, instruction documents, and blog posts. The expectation is that this research will need approximately 200 to 250 documents of variable lengths and a variable number of acronyms. As for now, we anticipate having around three to four acronyms per document for the Dutch language.

Annotation system

The first step in the data gathering process is creating a text annotation platform suitable for both in- and out-expansion. Creating a task-specific framework is necessary since no single open-source system easily allows different uses to annotate out-expansion acronyms. In the past, open forms like Google-Forms were used as a labeling framework, but these forms require extensive management (dividing the document over different users) and pre-processing (the form's output was a single string of acronyms). The newly created system automates the management task and simplifies the pre-processing tasks. The front-end of the system includes a website built using Flask, a framework in Python that enables the creation of web applications. There is a SQL database running in the back-end, and it will contain all the documents and annotations. As for the user of the system, the system is targeted at other students who are willing to help. Everyone who participates stands a chance to earn a reward for their contribution. The user interacts with the system by first signing up with their email address and selecting the language of interest, followed by two pages of instructions. These instructions should give the user a clear sense of the way in which the documents have to be annotated. For example, one important requirement is that the acronym-expansion pairs are documented in exactly the same way as they are written in the document. Annotating the documents starts after the users have read through the instructions. The page contains the document to be annotated and two sections to fill in the acronym-expansion pairs. The first session is reserved for in-expansions and the second one for the out-expansions.

The back-end of the system contains some rule based approaches to ensure that the data is of proper quality. These rules include:

- Every document is annotated by at least two different people to notice and prevent any mistakes.
- The annotations of the two users for a specific document are compared and if the difference is too great, they are manually checked for a third time to determine which annotations are correct.
- The documents are selected randomly, therefore, a user will annotate documents with different topics (for example, sports and politics)

The output of the systems is a .json file that contains the acronym-expansion pairs for both in-expansions and out-expansions. These can then be merged with the documents and together, they are used as training data for the Dutch version of the AcX system.

3 EXPLORATORY DATA ANALYSIS

Currently, there is no labeled data set, making it impossible to perform a full EDA. However, the first set of documents (60 text files) have already been scrapped and will be used in this preliminary EDA.

One of the data requirements is that it must be as diverse as possible to guarantee good generalisability. Measuring the data variability will be done based on the following three metrics: document length, document sentiment, and topics.

The first step is to clean the data in order for it to be effectively measured against the metrics above. Since this is text data, the following five steps have been performed:

- Removing punctuation
- Removing excess white spaces and newlines
- Convert all letters to lowercase
- Remove stop-words
- Tokenize the documents

Figure 1 shows the word count distribution per document of both the cleaned and dirty cleaned data. The histogram shows that the distribution is right-skewed, and this is caused by an outlier with around 9000 characters. The goal for the final corpus is to have a more homogeneous (and possibly normally distributed) distribution.

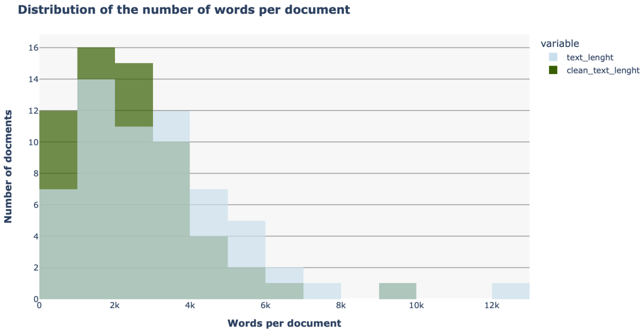

Figure 1: A histogram showing the number of characters per document in corpus for both the cleaned and dirty cleaned data

Figure 2 builds on the histogram and shows a box-plot of the document lengths. The figure shows a median document length of 2000 words and the first and third quartile at 1500 and 3200 words. Again, the final corpus must contain a broader distribution of documents.

Most of the previous works on acronym expansion has been done on research papes, and these have a different sentiment than, for example, blog post or news articles. The goal behind the AcX system is to identify acronyms in all kinds of various documents, which is why a diversity of sentiment is essential. The box-plot plot in figure 3 shows the polarity score of the different documents, and this indicates whether a document has a positive or negative tone. The polarity is determined with the Dutch version of TextBlob. A score of +1 is assigned to documents with a highly positive sentiment

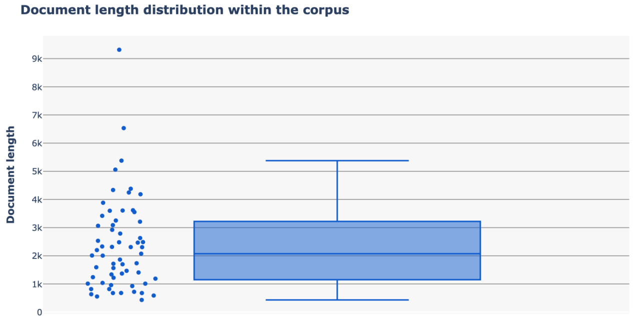

Figure 2: A box-plot that shows the distribution of the document lengths

and -1 to documents with a highly negative sentiment. Figure 3 shows the distribution of this score, and as can be seen, most documents have a polarity score of around 0, with a few outliers of 0.6 and 0.8.

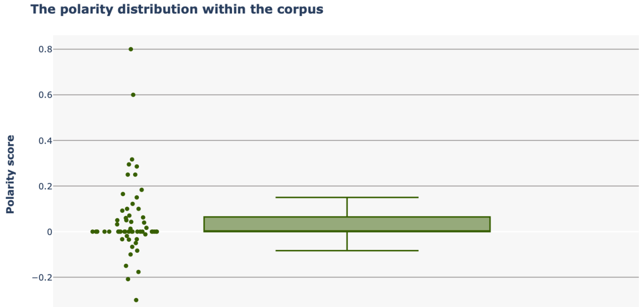

Figure 3: The polarity distribution of the corpus, calculated with the Dutch version of the TextBlob package

Topics are a key performance indicator (KPI) for measuring the diversity of documents. The word cloud in figure 4 was initially created to get a proper sense of the different topics. The image shows that the words "tussen, eerste, werden and twee" are the most common in the corpus. In addition, the letters ’a’ and ’s’ are also frequently used. Regardless, to get a more accurate picture of the topics in the corpus, a Latent Dirichlet Allocation (LDA) model from the Gensim package was used. This model is not optimized for Dutch but for now, gives a reasonable picture of the topics. The results show that the topics ’eerste, tussen, and KLM have the highest weights. This doesn’t say much because the words eerste and tussen are not really topics in the Dutch language (they are not standalone words). On the other hand, the word KLM is a recognizable topic. The word cloud and the LDA results indicate that more data cleaning processes are needed.

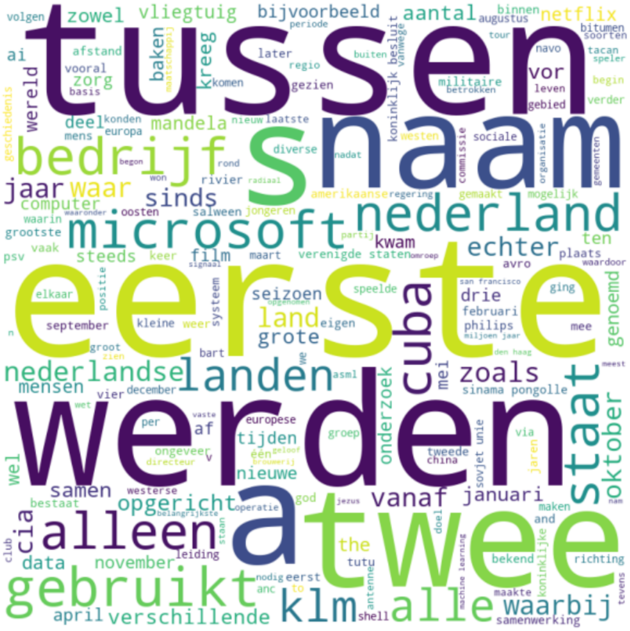

Figure 4: A word cloud with the most common words

EDA on the in-expansion data-set from Veyseh et al.

There are currently no Dutch data sets that contain annotated acronyms. However, performing an exploratory data analysis (EDA) on an English data set should provide a clear sense of the data requirements.

The data used in this part of the analysis comes from Veyseh [1] and is used to test in-expansion methods. The data contains 3980 sentences with acronyms and has three columns; one for the raw text (just one sentence), one for the location of the acronym in the sentence, and one column for the location of the expansion. The locations are based on the index of the starting letter in the sting. For example, ’The master information studies (IS) is awesome.’ contains the acronym IS with its expansion "information studies." These locations would be denoted as [31,32] for the acronym and [11,29] for the expansion. Furthermore, The data contains 7689 acronyms from which there are 7074 unique acronym-expansion pairs.

[1] https://github.com/amirveyseh/AAAI-22-SDU-shared-task-1-AE

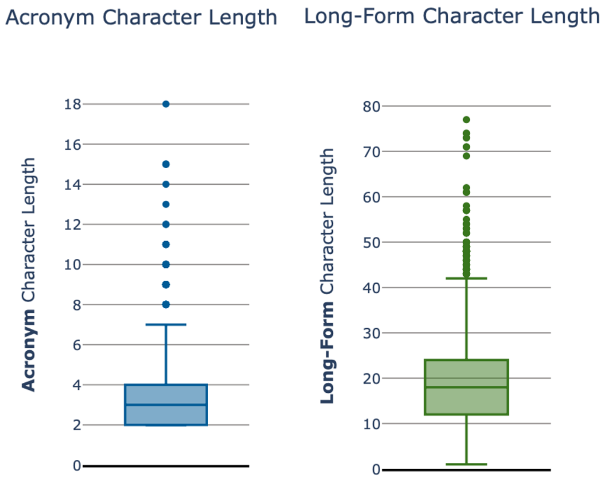

Figure 5: Box plot with the character length distribution for both the acronyms and long-forms

Finally, Figure 5 shows the character length distribution for both the acronyms and expansion. The data shows that the average length of an acronym is 3.3 characters (median=3) and Figure 5 shows that 75 % of the acronyms have a length equal to or less than 4. The general distribution is right-skewed, with the maximum being 18 characters. For the expansion, the average length is 18.5 characters length (including the spaces and punctuation marks). Here, the distribution is also right-skewed.

The Dutch data sets will be different from the one create by Veyseh et al. in the sense that it will not contain the location of the acronyms and corresponding definitions. Also, it will include the out-expansion acronyms, making it a suitable for the AcX system.

REFERENCES

[1] Jean Charbonnier and Christian Wartena. *Using Word Embeddings for Unsupervised Acronym Disambiguation.* Tech. rep., pp. 2610–2619. URL: http://noa.wp-hs-hannover.de.

[2] Aadarsh Singh and Priyanshu Kumar. *SciDr at SDU-2020 : IDEAS-Identifying and Disambiguating Everyday Acronyms for Scientific Domain.* Tech. rep. 2021, pp. 451–462.

[3] Sungrim Moon, Bridget McInnes, and Genevieve B. Melton. “Challenges and practical approaches with word sense disambiguation of acronyms and abbreviations in the clinical domain”. In: *Healthcare Informatics Research* 21.1 (Jan. 2015), pp. 35–42. ISSN: 2093369X. DOI: 10.4258/hir.2015.21.1.35.