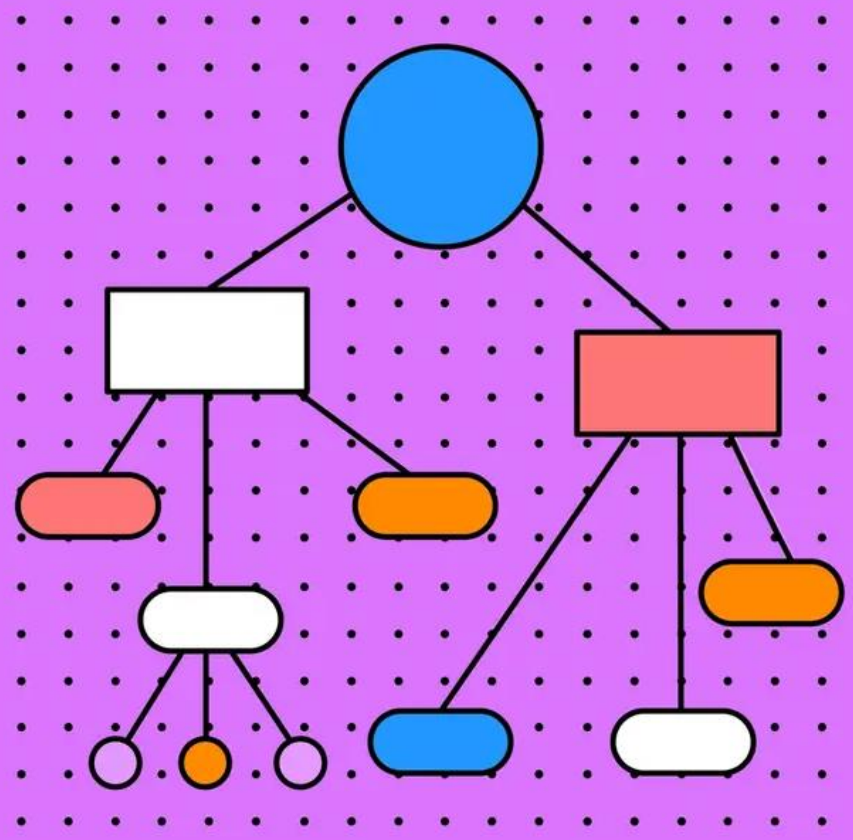
CONCEPTUAL MAP GENERATION FROM TEXT

A preliminary analysis of how Natural Language Processing tools can help facilitating students' lives when studying for an exam

Course: Natural Language Processing – 20597
Professors: Amanda Cercas Curry, Dirk Hovy, Debora Nozza



<u>Group 9:</u>

d'Imporzano Lorenzo (3091665)
Izzo Domitilla (3060888)
Matarazzo Luca Silvestro (3081393)
Rossi Elena (3110844)
Valdo Riccardo (3112743)

1. Introduction

When studying for an exam, students usually summarize texts into conceptual maps so as to facilitate memorization and recall procedures. As stated by the University of North Carolina, conceptual maps are a “powerful study strategy because they help you see the big picture: by starting with higher-level concepts, concept maps help you chunk information based on meaningful connections” (UNC-Chapel Hill Learning Center, 2022)[1]. Nowadays, Artificial Intelligence and computational methods are achieving an ever-increasing boost and growth. These advancements could and actually do bring really helpful innovations into everyday life, with the academic/scholastic worlds being no exception. As a matter of fact, several applications of Natural Language Processing techniques are already of common use into these environments and are helping professors and students to achieve higher performance. The aim of this report is to apply NLP methodologies to transform a text into a conceptual map. In the following paragraphs, all the steps to reach the final result will be explained thoroughly. In particular, after a brief section about literature review in which the main developments in this field are analysed, the data collection procedures are explained. Afterwards, the actual strategy used to transform text into conceptual maps will be presented followed by a section in which results will be analysed and the output maps will be compared both to “dumb” baselines, developed by a non-optimized version of the previously presented model, and to hand-made maps developed personally by us. Additionally, results will also be evaluated according to the ROUGE score, a commonly used evaluation metric for text summarization and machine translation. In particular, all the sentences of a conceptual map will be joined so as to form a single text, which will be then compared in terms of ROUGE to the summary of the respective paragraph generated with BART.
Finally, conclusions will be presented alongside an exploration of the model’s limitations and further analysis which could be conducted starting from the work included in this report.

2. Literature Review

There have been numerous attempts to automatically generate concept maps from text, dealing with either individual document (Valerio & Leake, 2006)[2] or document clusters (Rajaraman & Tan, 2015)[3], but only a few clearly state that a summary is their desired goal. Traditional summarization procedures, such as information and keyword extraction, are still applicable to selecting the most significant relationships for a concept map, and Zubrinic et al. (2015)[4] propose the TF-IDF metric on the level of concept labels combined with a similarity measure to identify and merge similar concepts. Falke and Gurevych (2017)[5] pioneered the concept-map-based document summarizing task, where they employed the Open Information Extraction (Banko et al., 2007)[6] paradigm in conjunction with an SVM classifier for relationship selection. By considering a concept map to be fully characterized by the set of propositions it contains, they computed ROUGE-2 between the two proposition sets of the generated map and a benchmark for evaluation. Most of previous approaches were, however, evaluated on basic and short documents with a limited number of subordinates. Furthermore, they often limit concepts to one to three words, which can make the map difficult to interpret and oversimplified in comparison to the information that the original text seeks to convey. This paper contributes to the literature by proposing, building on earlier work, a novel strategy for conceptual map production that aims to go beyond the constraints of using short paragraphs or simple syntax.

3. Data Collection

To conduct the analysis presented in this report, more than forty thousand paragraphs were collected via web scraping, both from Wikipedia pages, in their greatest part concerning historic events or persons, and textbooks (in PDF format). The final data frame contained more than 1.2 million sentences of different lengths and various arguments. It has to be noticed that not all will be used for the sake of this paper, though they are available and ready to be used for further analysis or model improvements, as will be suggested in the section about limitations and further analysis. To perform such scraping, in particular, the

wikipedia Python library was used for Wikipedia's pages, along with different libraries for the text extraction from the PDFs. As a matter of fact, different libraries seemed to work better than others, probably due to the way in which the specific PDF textbook was formatted. Overall, the great majority of the data still comes from Wikipedia pages.
After having collected the data, it was noticed that some paragraphs were too long to be properly summarized in a conceptual map. Therefore, paragraphs exceeding a word count of 600 were split so that no paragraph in the final data frame exceeds this word length, which approximately amounts to one standard A4 page. Clearly, paragraphs were not split in the middle of sentences but rather when the last sentence before the one that leads to the exceeding of the 600 words ends.

4. Methodology

4.1 Initial Extraction

After having collected the data, the actual code to transform a text into a conceptual map was developed. Clearly, all the following steps must be repeated for every text for which a map needs to be generated.
In particular, the first task was the extraction of the relations present inside the paragraph under analysis. Relation refers to the semantic relationships or associations that are extracted from natural language sentences. Practically, relations are tuples of the form (argument 1, relation, argument 2). For instance, (Jeff Bezos, founded, Amazon). To extract them, OpenIE (Open Information Extraction) was used. This is a Java written tool developed by the Stanford NLP Group which focuses on extracting relations between entities or concepts mentioned in the text, typically represented as subject-predicate-object tuples, as in the example presented above. The subject and object represent the entities or noun phrases, while the predicate represents the relation or verb phrase that connects them. Hence, starting from a text, this tool provides a list of relations with an associated confidence, which, as the name suggests, is an indication of how much the extractor is "confident" the result is correct.
Before actually applying this tool, the texts were passed through NeuralCoref. This Python library, built on top of SpaCy, provides coreference resolution capabilities. In particular, this step was deemed necessary as, many times, relations extracted with OpenIE included pronouns like he, she, or they as argument 1. This kind of arguments are clearly not suitable to be included in a conceptual map, as there is no reference to who "he" or "she" is. NeuralCoref partially solves this problem through the use of neural mention-ranking models and allows the pronoun to be linked to the name to which it refers. At this point, a sample of the relations extracted from different paragraphs were manually and visually analysed to understand whether there were some systematic differences and discernible patterns that were, usually, associated with either irrelevant or not conceptual map-suited extractions. As a matter of fact, after having examined more than five thousand relations, some patterns were identified and, thus, immediately after the extractions, relations were removed if they:

- Do not contain any noun (either proper or not) in the first argument;
- Contain only one or two words which are pronouns in the second argument;
- Contain a relation made up by just two words of which the first one is "to" (and hence the verb is in the infinitive form);
- Have a relation which does not contain a verb;
- Had either the first argument or the second argument ending in preposition, as this means that the sentence has been truncated and hence the meaning is probably lost.

These were indeed the common characteristics shared by most of the erroneous relations identified.

4.2 Pre-processing functions

After this step, several functions were coded and applied to additionally prepare the data before building the final conceptual map.
To begin with, a function (*filter_rows*) to further filter the remaining relations was defined. In particular, it merges the three parts of each tuple into a single sentence and compares it to the other so formed sentences. This function then drops all duplicate instances, and, in the case there are some substrings, the longest one is kept (e.g., if both "Barack Obama was born in Hawaii" and "Barack Obama was born in Hawaii in 1961" are present, only the second one is kept). Then, another function (*merge_concepts*) was also defined. This function takes all the unique instances of the first argument and the title of the paragraph and creates clusters of arguments based on the

Cosine Similarity. More in depth, the threshold, identified via trial and error, to be included in the same cluster is 0.4 in Cosine Similarity (sentence embeddings, necessary to compute the similarity measure, were obtained through TF-IDF as it was seen to perform quite well on the data under analysis). Then, a for loop is also run to merge all those clusters which have common elements so as to avoid overlaps. After all these procedures, for each cluster, the function selects the argument to be used based on the length and on whether the cluster contains the title or not. In particular, if the cluster contains the title, then *merge_concepts* selects it as the representing string; otherwise, the shortest among the arguments is selected, which should also represent the simplest (in any case, the title was then included in every map and linked, by default, to every other argument 1 node since the title is usually included in every conceptual map). This function was defined to avoid the final map to have distinct nodes for the same subjects. For instance, without this intervention, a map could have had two distinct nodes, one containing "Obama" and one containing "President Obama", which made the maps very cumbersome, and hence it was deemed appropriate to have just one node ("Obama" in this example's case) for each subject.
Subsequently, the function *concat_concepts* was defined and applied. This function has a two-fold application: it concatenates both the objects and the subjects (namely the second and first arguments). Starting from the objects, the function groups the data frame by the first argument and the relation. Then, it considers all objects that share the same argument 1 and relation and identifies whether there are any common substrings in them. This is done so that, in the final conceptual map, there are no repeated substrings in the objects. For instance, if two objects are "happened in Moscow" and "happened in London" these will not be two distinct nodes in the final output but rather one node containing "happened in Moscow and London". After this, almost the same procedure is repeated for the subjects by grouping the data frame by relation and argument two. The goal in this case is the same as for the objects.
Afterwards, the function *most_similar_phrases* was defined to remove sentences with very affine meanings. This is achieved by comparing the phrases (obtained by joining, once again, argument one, relation, and argument two) and, if two or more of them display a Cosine Similarity greater than 0.3 then only the longest is kept as it is likely the most detailed and informative one at this point.
After all these functions had been defined and applied, the remaining relations were filtered according to the confidence score that was automatically assigned by OpenIE during the extraction phase. In particular, only the relations exceeding the median confidence were kept. Still, some irrelevant or not conceptual map-suitable relations remained, and hence a classifier was built and trained so as to contribute to the pruning procedure and identify which relationships are worthy of ending up in the final conceptual map and which are not.

4.3 Classifier

A first attempt to build a classifier was done so as to develop one able to predict whether the entire sentence, namely argument one + relation + argument two, is meaningful or not. To this end, more than four thousand instances were manually annotated assigning a 1 to meaningful and informative sentences while a 0 to those that are not. A RandomForestClassifier was then applied by feeding as features the confidence OpenIE has given for the specific relation, the vectorisation of the sentence obtained via the pretrained model all-MiniLM-L6-v2 from HuggingFace, the dummy indicating whether the sentence is in the negative form and the dummy indicating whether it is in the passive voice (both indications provided by OpenIE). On top of this, also the length of the sentence and the count of the POS (Part-Of-Speech) tags was fed as input, as previously done also by Falke and Gurevych (2017)[5]. More in depth, the count of nouns, pronouns, adjectives, adverbs, verbs, determinants, and numerals were provided both in absolute and in relative terms. This was done since during the annotation procedure it was noticed that, for instance, most of the first arguments which do not make sense were formed either by many pronouns or by many adjectives, hence, having the count was thought of relevance for the classifier.
Though, even after a Randomized Search to tune the hyperparameters of the forest, the model proved to achieve a very poor accuracy and was performing almost as a random classifier. This was probably because, as noted during the annotation procedure, there were very similar sentences, both in terms of meaning and in terms of structure, which were in some cases annotated as valuable and in

others as irrelevant. Indeed, sometimes OpenIE outputs sentences like "The Romans invaded Palestine" and "Palestine was by the Romans", clearly the second is non sensical, though it is not an easy task to teach the model the difference between the two. For this precise reason, a second attempt was performed to manually re-annotate six hundred more relations and selecting them carefully so that there was no overlap and that they were as informative as possible, so as to likely help the classifier to understand which patterns have to be considered informative and which do not. Still, even with this new data, when using the full sentence, the problem persisted. Some advancements have been made when the three parts of the extraction were considered separately and three different classifiers were trained. In this case, the annotation on the six hundred selected instances was done to identify three labels, namely whether the first argument, the relation, and the second argument are respectively valuable. To decide whether the singular parts of the extraction were valuable or not, strict rules were applied:

- Each part should be in correct English;
- Each part should be sensical even without context (especially so for argument one and two, but not so much for the relation which is usually just a verb);
- There should not be vague references ("e.g., "Someone" or "Many" were not considered as valid arguments)

After having applied these rules and obtained the annotations, the three classifiers were built and trained, and their hyperparameters have been tuned. Among the three parts of the extraction, the only one displaying balanced classes was the one for argument one. Indeed, the relation and argument two displayed a large majority of class 1 instances, which also makes sense logically as the relation is usually just a verb, and it is very hard that it is not valid, while argument two is usually a long sentence taken from the original paragraph and, hence, it is also hard that it does not follow the rules listed above. Among the three classifiers, once tuned and trained on similar features as the ones for the classifiers for the entire sentence, the one for argument two and the relation performed poorly, just a bit better than a random classifier. This is probably due to the fact that those few instances that were labeled as 0 did not have many differences compared to the ones classified as 1, as happened in the case for the classifiers for the entire sentences. This was the case even after a custom loss function was used in the tuning of the hyperparameters. In particular, several configurations for the loss function were attempted (e.g., giving five times the weight to the error of misclassifying 0 as 1 rather than 1 as 0), though none seemed to significantly improve the situation. Also, other models were attempted (e.g., XGBoost, HistGradientBoosting, SVM...) but none helped. This, once again, shows that the task at hand is not an easy one. An attempt was also done to try and perform both under and oversampling. Though the former led to having too many data points to be considered, whilst the latter led to overfit. Therefore, none was of any help in solving the problem present in these classifiers.
Nevertheless, the classifier for argument one was performing well, achieving an accuracy of more than 81%. This success was likely due to the higher internal similarity between instances of classes 0 and 1 in the argument one's case. As this classifier was performing well, it was added to the pipeline to create the final conceptual map. An additional hint on why the classifier for argument 1 was working well is that there are much more differences, also in the topics, when it comes to the distinction between valuable and irrelevant instances, as can be seen in Appendix 1. Hence, after the function *most_similar_phrases* this classifier (pretrained on the second 600 annotations) was applied to the first argument of each remaining relation as a way to prune them even further.

4.4 Keyword filtering and final map

After having performed all of the above, a filtering based on whether the sentence contained or not any keyword of the paragraph was carried out. In particular, keywords were extracted from the original paragraph both with TextRank and RAKE (Rapid Automatic Keyword Extraction algorithm), which both extract and associate a score to every recognized keyword in a text. The main difference between the two is that RAKE uses a combination of statistical and linguistic heuristics to identify potential keywords, while TextRank represents the text as a graph, where sentences or words are nodes and the relationships between them are edges. RAKE then assigns scores based on the frequency of the words and their co-occurrence patterns, while TextRank instead ranks the nodes based on their importance in the graph, which is determined as the sum of the importance scores of its connected nodes. It was deemed

appropriate to consider both scores, as they each have their strengths and weaknesses.
After having extracted the keywords and the associated scores, all the remaining relations which displayed zero both in the RAKE and in the TextRank scores were removed from the ones to be included in the conceptual map. This amounts to removing all those sentences (once again, argument one + relation + argument 2) which do not contain any keyword, according to both algorithms. Indeed, likely, if a sentence does not include any keyword, then it is not a relevant one for a conceptual map.
Last but not least, before generating the map, a last pruning of the relations was performed. Specifically, all those extractions which still presented either 1 single-word argument two or a relation with more than 6 words were removed as likely they are not suitable for a conceptual map, which should include short relations (as it is for the greatest part of the remaining extractions) and significant objects (second argument).
This last trim concludes the pipeline for the map generation. Therefore, after all these steps, the final map was generated with nodes containing argument one and argument two, while the relation was displayed on the edge linking them. Moreover, the title of the paragraph was also added as a node in every map, as it is clearly of vital interest for students who are summarizing a specific text. An example of the result obtained with this procedure is included in Appendix 2.

5. Model Validation

Given the difficulty of the task, it was also hard to find a proper way to validate and assess the obtained results. As a way to try to do so, a two-fold approach was taken. In the first place, a user study was run. In particular, the final map resulting from the aforementioned steps, a manually made map (computerized so that differences could not be immediately told) and the baseline map were presented to several individuals, and the comments on them were collected. The baseline map consists of the one resulting from the relations extracted via OpenIE cut at the 75th percentile of the confidence distribution (given that no other pruning strategies were applied, it was deemed appropriate to cut at a higher confidence compared to the one that was used in the case of the steps listed in the sections above, which was at the median, so that the baseline map does not have too many nodes). An example of the three maps presented to the individuals taking part in the user study can be found in Appendix 3. More in depth, five different sets of maps corresponding to five different paragraphs were shown to twenty different people to have variability between maps whilst also having a valid sample. As a matter of fact, showing a different set of maps to every participant would have led to one feedback per map and this would have come at a cost for objectivity. With four opinions per map, the effect of personal biases should be in part reduced. Overall, irrespective of the set of maps presented, the results of this study showed that both the model-generated and the human-created maps were considered satisfactory in terms of overall understanding. On the other hand, the baseline was regarded as useless, with nonsensical and difficult-to-interpret connections. The handwritten map, in most cases, was deemed to depict the text more precisely than the other maps, although, according to most respondents, both the handwritten and model maps were complete and did not leave out any significant information contained in the original paragraph. The map produced by the model was considered, however, quite minimal, and occasionally failed to recognize some concepts that should have been merged together. Furthermore, the model did not always accurately record the dates of the events, and the subjects were sometimes assessed as lacking context. For these reasons, the map generated by the model was thought to be useful for saving time and revising concepts after reading the paragraph, but not to be blindly trusted and studied from. On the contrary, the human-created map was thought to be good enough to comprehend the entire paragraph without having to read it. In the context of applications to historical events, some respondents suggested arranging the map by date rather than subject or event to emphasize the chronological order. However, the goal is to develop, starting with this application, a model that can be applied to a variety of subjects and topics rather than just history. Therefore, arranging information by concept is still viewed as more generic and task-appropriate. There were no more suggestions regarding the map structure, which was deemed neat and visually appealing.
On top of the user study, some objective metrics were also used to evaluate the correctness of the generated maps. In particular, given that

conceptual maps should, in a sense, summarise the content of a paragraph, it was deemed sensible to confront them with a summary of the same paragraph; this approach is also suggested in the literature[7]. More in depth, to do so the summary was generated using a pretrained BART model from HuggingFace[1], designed for extractive summarization, which seemed the most suited as, like the concept map, it retains the most relevant sentences from the original text to create a summary. First, a random sample of 253 Wikipedia paragraphs was collected, and it was fed into both the concept map and the BART summarization model (which was adequately parametrized). The sentences collected in the concept maps were concatenated together to form paragraphs, acting as summaries, thus allowing to confront them with summaries. Then, the ROUGE-2 metric was computed, using as reference the BART summaries. The choice to consider ROUGE-2 was made in accordance with the literature on concept map evaluation[7]. Average values of precision, recall, and F1 scores were similar: around 0.2; distribution plots are collected in Appendix 4. These results are in line with other analyses in the literature[8] and signal the ability of the model to partially retrieve relevant elements from the text in the same way as BART does. However, there are two relevant limitations to this analysis: firstly, the different nature of the summarization techniques. Indeed, concept grouping in nodes caused text repetition when concatenating relationships in the maps. At the same time, the concept map model, differently from BART, does not really summarize text, as it is "relation-based", hence whenever it considers a relation important, it includes it in the map, even though it may be long and contain unnecessary details, which BART may instead discard. Secondly, the automatic concept maps display varying lengths, depending on the paragraph's characteristics, while BART can be parametrized to output summaries of a bounded length. This is not necessarily a defect of the proposed model, as certain paragraphs may be denser in information than others, and in such cases, it makes sense to produce larger maps for a fixed length of input. Nonetheless, the difference increased variability in summarization performance, which was quite high, as seen from the distribution plots in Appendix 4.

As a last evaluation, a sample of 18 paragraphs was considered, with four different summaries: the BART ones, the textualized concept maps from the full model, the textualized concept maps from the baseline model (stemming at the 75-th percentile of the confidence distribution) and the textualized handmade concept maps (used for the user study). The latter three concept maps were confronted with the BART summaries, this time only in terms of ROUGE-2 precision. In fact, for ease of representation, precision seems the most suitable metric to represent the quality of an automated concept map summary, as it represents the portion of the latter which is contained in the reference summary. This means precision controls for the size of the automatic concept map (which, as we said, is generally variable and may bias recall), and it can be thought of as a purity measure of the respective summaries. On average, the full model performed better than the baseline and worse than the handmade maps (respectively, 0.15, 0.19, and 0.22). Thus, the ability of our model to encompass the relevant information contained in the paragraphs seems not too far from that of a human. Interestingly, handmade maps displayed more variability in precision than automated maps, which may reflect differences in personal summarization ability across authors of this report. Distribution plots are displayed in Appendix 5.

6. Conclusions

This paper illustrates how, with the help of Natural Language Processing, it is possible to achieve great results when trying to generate automatically conceptual maps starting from a text. In particular, it was showed that, by combining pretrained models, established libraries, and custom-made functions and classifiers, it was possible to filter out some of the irrelevant and meaningless relations from a starting paragraph with the final goal of building a conceptual map containing both meaningful and correct sentences, and not simply single words or bigrams. Overall, the results obtained, once compared with hand-made maps, were satisfactory, as displayed in the section above. Still, some improvements are clearly possible, and a deeper analysis could be conducted, as highlighted in the final section of this report hereafter.

[1] sshleifer/distilbart-cnn-12-6

7. Limitations and further analyses

As highlighted throughout the discussion above, the task of transforming text into conceptual maps is a though one. As a matter of facts, it is very hard to ask a machine to understand whether a sentence is meaningful whilst being correct from a grammatical point of view and also having the potential to be useful in a conceptual map. Clearly, the above explained procedure can be improved and even better results can be achieved.

One possible way of doing so is to manually annotate even more instances so as to better tune the classifier on argument one and also, potentially, solve the problems associated with the classifiers of argument two and of the relation. Indeed, classes' unbalance could be solved by finding new relevant instances of class 0, and, hence, results could be potentially improved as the classifier will have more instances of class 0 from which to learn. Moreover, more fine techniques could also come into hand. For instance, a recurrent neural network with a personalised loss function could prove to work well in this scenario. More in depth, if a RNN is used to classify the arguments and the relation then it would be possible to try and give to the model the vectorization also of the other two parts of the tuple and not only the one of the parts under analysis. This could be beneficial as having the full context in which the specific part of the extraction is related to might help the model to better predict whether that specific part is a valid one or not.

On top of this, to reduce personal bias, having different people annotating the same data and then check their agreement rate could be of great help. As a matter of facts, due to time constraints, for the sake of this paper, annotations have been carried out by different individuals (4 out of 5 members of the group). Even if clear rules were defined and agreed upon, each person has its own personal biases and it is possible that the criteria used by each one of the annotators do not match perfectly resulting in a possible source of variation, and hence confusion, for the model which might have had reduced performances due to this problem.

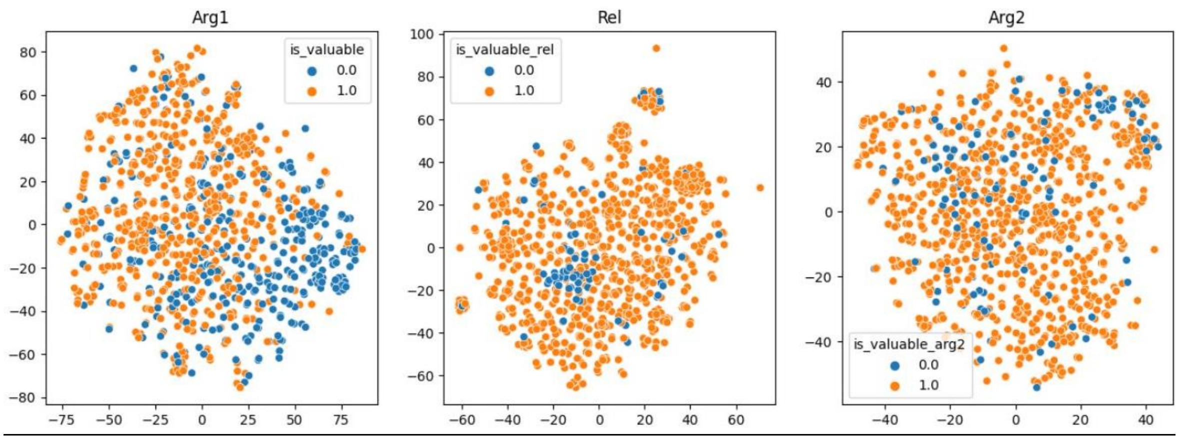
Personal bias could also be a concern in the user study. Indeed, evaluating the set of three maps requires quite a bit of time as individuals needed to read the three maps as well as the paragraph they come from (made up of almost 600 words most of the times). Because of this, it was hard to find people available to perform this operation. Though, if results could be evaluated by a greater amount of people (like 100-150 persons per set of maps) then the feedbacks would have a much higher statistical significance and the stochasticity due to personal biases could be reduced.

Having access to more data and to more computational power to better tune the models could also clearly be of great help. Indeed, LLM which are working well nowadays have huge amount of training data and computational power to elaborate them.

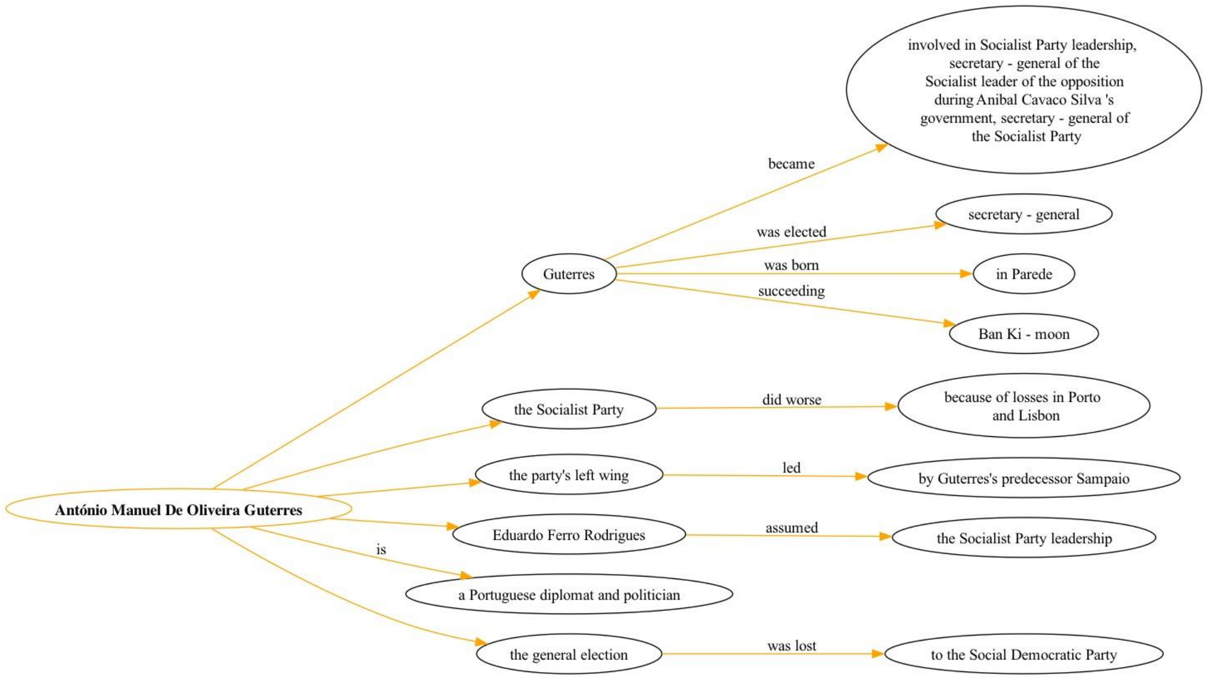
Though, overall, the code developed with the steps explained in this report still performs quite well considering all the valuation techniques used and the constraint faced. Clearly, this has to be intended as a base for further analysis and research and not a proper final point, as often is the case of language models.

Appendix

Appendix 1: Representation of the vectorization of the three parts of the extraction of the newly 600 annotated ones.



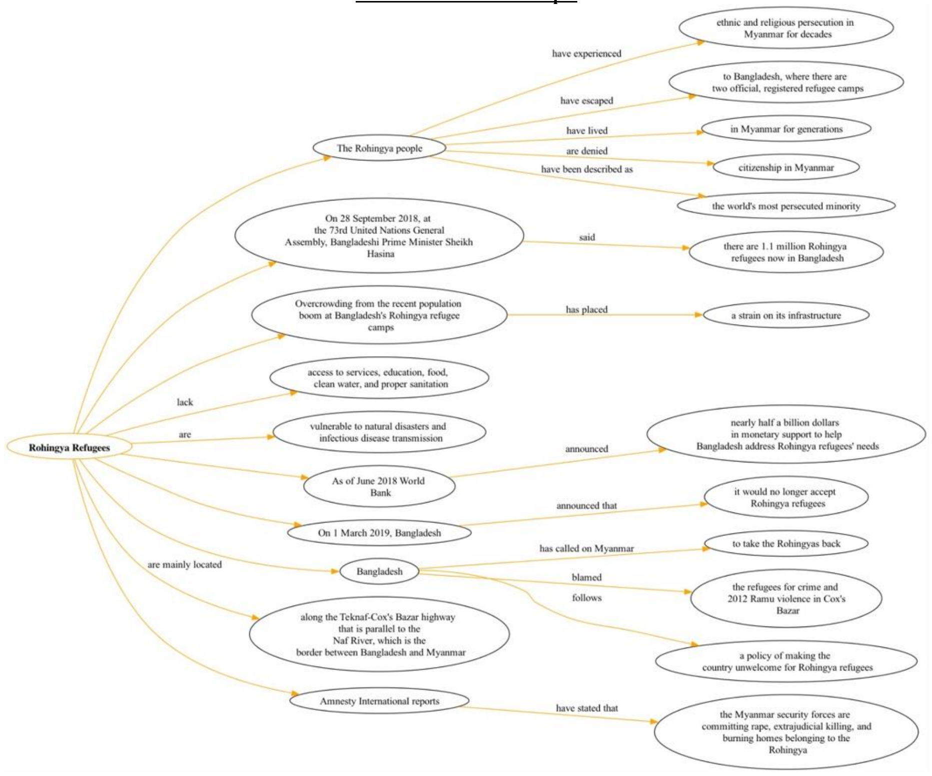
Appendix 2: An example of a conceptual map generated by the pipeline explained in this paper



António Manuel De Oliveira Guterres
Guterres
became
involved in Socialist Party leadership, secretary - general of the Socialist leader of the opposition during Anibal Cavaco Silva 's government, secretary - general of the Socialist Party
was elected
secretary - general
was born
in Parede
succeeding
Ban Ki - moon
the Socialist Party
did worse
because of losses in Porto and Lisbon
the party's left wing
led
by Guterres's predecessor Sampaio
Eduardo Ferro Rodrigues
assumed
the Socialist Party leadership
is
a Portuguese diplomat and politician
the general election
was lost
to the Social Democratic Party

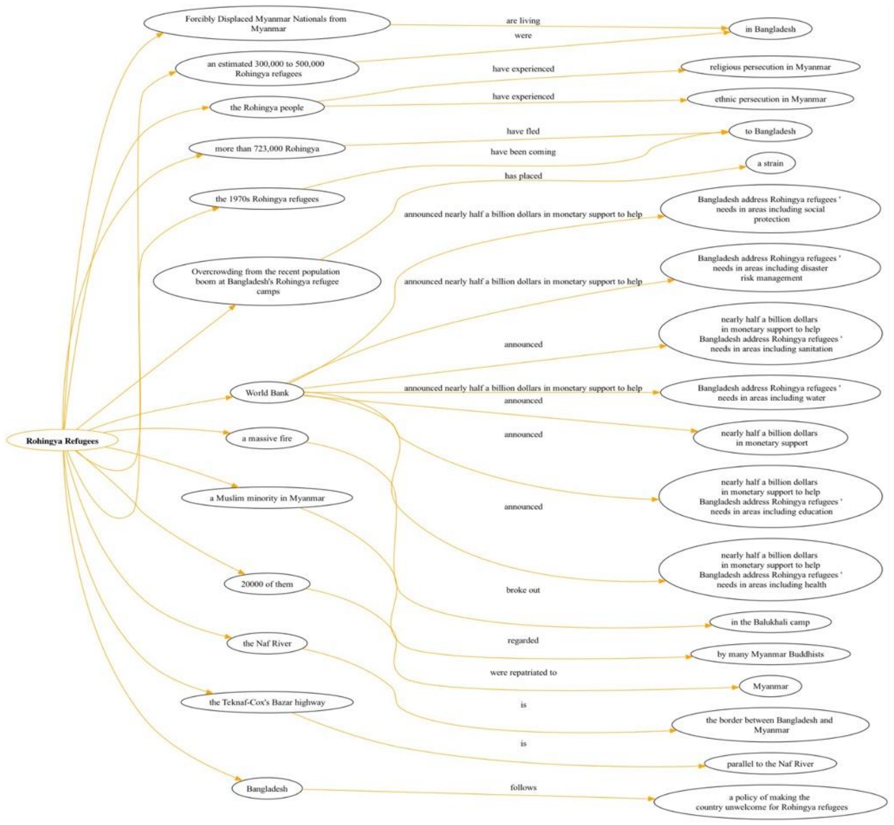
<u>Appendix 3</u>: An example of a set of the three maps presented to individuals in the user study, in order respectively there is the hand-made one, then the baseline and, to conclude, the map generated by the model.

<u>Hand-made map:</u>



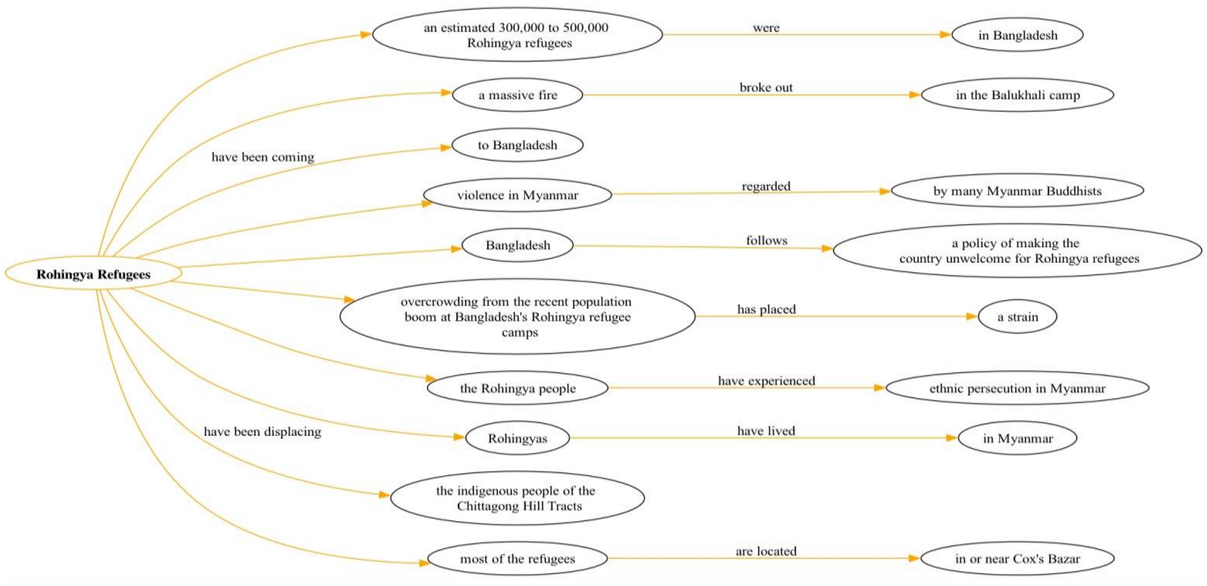
Rohingya Refugees
The Rohingya people
have experienced
ethnic and religious persecution in Myanmar for decades
have escaped
to Bangladesh, where there are two official, registered refugee camps
have lived
in Myanmar for generations
are denied
citizenship in Myanmar
have been described as
the world's most persecuted minority
On 28 September 2018, at the 73rd United Nations General Assembly, Bangladeshi Prime Minister Sheikh Hasina
said
there are 1.1 million Rohingya refugees now in Bangladesh
Overcrowding from the recent population boom at Bangladesh's Rohingya refugee camps
has placed
a strain on its infrastructure
lack
access to services, education, food, clean water, and proper sanitation
are
vulnerable to natural disasters and infectious disease transmission
As of June 2018 World Bank
announced
nearly half a billion dollars in monetary support to help Bangladesh address Rohingya refugees' needs
On 1 March 2019, Bangladesh
announced that
it would no longer accept Rohingya refugees
Bangladesh
has called on Myanmar
to take the Rohingyas back
blamed
the refugees for crime and 2012 Ramu violence in Cox's Bazar
follows
a policy of making the country unwelcome for Rohingya refugees
are mainly located
along the Teknaf-Cox's Bazar highway that is parallel to the Naf River, which is the border between Bangladesh and Myanmar
Amnesty International reports
have stated that
the Myanmar security forces are committing rape, extrajudicial killing, and burning homes belonging to the Rohingya

<u>Baseline Map:</u>

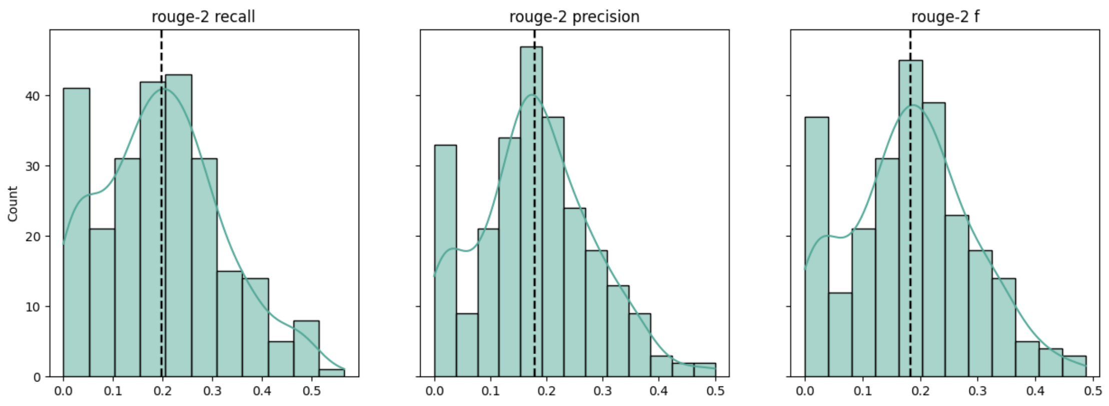


Rohingya Refugees
Forcibly Displaced Myanmar Nationals from Myanmar
are living
in Bangladesh
an estimated 300,000 to 500,000 Rohingya refugees
were
the Rohingya people
have experienced
religious persecution in Myanmar
have experienced
ethnic persecution in Myanmar
more than 723,000 Rohingya
have fled
to Bangladesh
the 1970s Rohingya refugees
have been coming
Overcrowding from the recent population boom at Bangladesh's Rohingya refugee camps
has placed
a strain
World Bank
announced nearly half a billion dollars in monetary support to help
Bangladesh address Rohingya refugees ' needs in areas including social protection
announced nearly half a billion dollars in monetary support to help
Bangladesh address Rohingya refugees ' needs in areas including disaster risk management
announced
nearly half a billion dollars in monetary support to help Bangladesh address Rohingya refugees ' needs in areas including sanitation
announced nearly half a billion dollars in monetary support to help
Bangladesh address Rohingya refugees ' needs in areas including water
announced
announced
nearly half a billion dollars in monetary support
announced
nearly half a billion dollars in monetary support to help Bangladesh address Rohingya refugees ' needs in areas including education
nearly half a billion dollars in monetary support to help Bangladesh address Rohingya refugees ' needs in areas including health
a massive fire
broke out
in the Balukhali camp
a Muslim minority in Myanmar
regarded
by many Myanmar Buddhists
20000 of them
were repatriated to
Myanmar
the Naf River
is
the border between Bangladesh and Myanmar
the Teknaf-Cox's Bazar highway
is
parallel to the Naf River
Bangladesh
follows
a policy of making the country unwelcome for Rohingya refugees

Model's map:



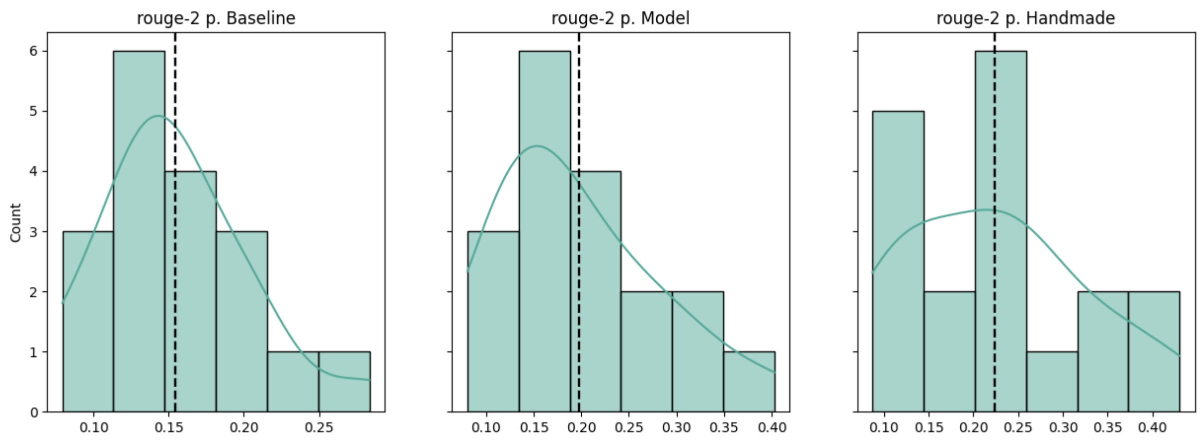
Appendix 4: Distribution of ROUGE-2 recall, precision and F1 scores for a sample of 253 randomly selected Wikipedia paragraphs, with BART summaries are reference and textualized concept maps from the full model as candidates. Averages are shown by dashed vertical lines.

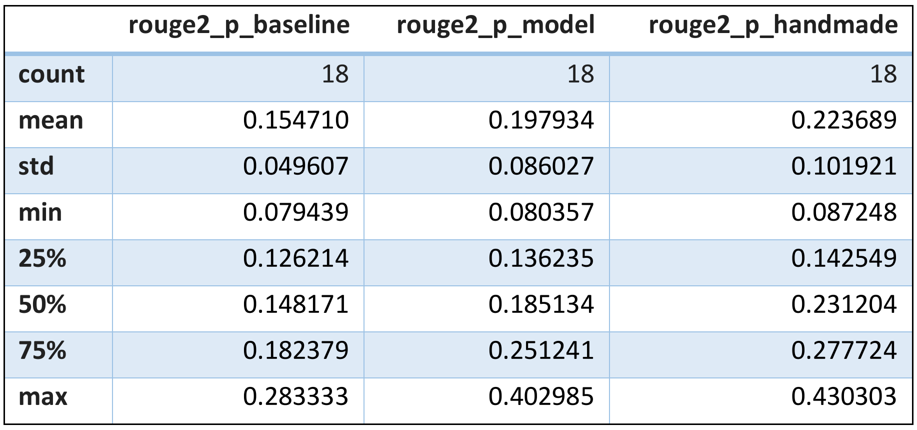


	rouge2_r	rouge2_p	rouge2_f
count	252	252	252
mean	0.197506	0.179135	0.183759
std	0.123498	0.103134	0.107118
min	0.000000	0.000000	0.000000
25%	0.109342	0.114827	0.114405
50%	0.193130	0.175219	0.184220
75%	0.272922	0.245797	0.249195
max	0.565574	0.500000	0.487395

Appendix 5: Distribution of
ROUGE-2 precision for a sample of 18

randomly selected Wikipedia paragraphs, with BART summaries as reference and with candidates being textualized concept maps from the full model, from the baseline model and from handmade maps. Averages are shown by dashed vertical lines.





	rouge2_p_baseline	rouge2_p_model	rouge2_p_handmade
count	18	18	18
mean	0.154710	0.197934	0.223689
std	0.049607	0.086027	0.101921
min	0.079439	0.080357	0.087248
25%	0.126214	0.136235	0.142549
50%	0.148171	0.185134	0.231204
75%	0.182379	0.251241	0.277724
max	0.283333	0.402985	0.430303

[1] UNC-Chapel Hill Learning Center. (2022, July 11). *Concept Maps – Learning Center*. Learning Center. https://learningcenter.unc.edu/tips-and-tools/using-concept-maps/#:~:text=They%20are%20a%20powerful%20study,significant%20and%20easier%20to%20remember.

[2] Alejandro Valerio and David B. Leake. 2006. Jump- Starting Concept Map Construction with Knowledge Extracted from Documents. In *Proceedings of the 2nd International Conference on Concept Mapping*, pages 296–303, San José, Costa Rica.

[3] Kanagasabai Rajaraman and Ah-Hwee Tan. 2002. Knowledge discovery from texts: A Concept Frame Graph Approach. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pages 669–671, McLean, VA, USA.

[4] Krunoslav Zubrinic, Ines Obradovic, and Tomo Sjekavica. 2015. Implementation of method for generating concept map from unstructured text in the Croatian language. In *23rd International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pages 220–223, Split, Croatia.

[5] Tobias Falke and Iryna Gurevych. 2017. Bringing Structure into Summaries: Crowdsourcing a Benchmark Corpus of Concept Maps. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2951–2961, Copenhagen, Denmark. Association for Computational Linguistics.

[6] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web. In Proceedings of the 20th International Joint Conference on Artifical Intelligence, pages 2670–2676, Hyderabad, India.

[7] A. Nugumanova and Y. Baiburin. Evaluation metrics for automatically constructed concept maps. 2021 21st International Conference on Control, Automation and Systems (ICCAS), Jeju, Korea, Republic of, 2021, pp. 365-370, doi: 10.23919/ICCAS52745.2021.9649911.

[8] Tobias Falke, Christian M. Meyer and Iryna Gurevych. Concept-Map-Based Multi-Document Summarization using Concept Coreference Resolution and Global Importance Optimization. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 801-811, Taipei, Taiwan. Asian Federation of Natural Language Processing.