Semi-Automatic Definition Phrasing

Master Thesis Proposal

Anna-Janina Goecke
Universität Potsdam
Matriculation number: 777707
goecke@uni-potsdam.de

October 14, 2022

1 Topic

The discourse of climate change is formed by various opinions ranging from the strong belief of need for action, to neglect and the active denial of the existence of climate change resulting from human activities. Even though dividing the discourse of climate change into two subgroups, namely *activists* and *skeptics*, is a very simplistic approach, it however reflects the two opposing attitudes towards climate change. This separation has been proven crucial for the task of generating a climate change discourse glossary consisting of climate change compounds carrying an intense connotation.

In previous projects, we created text corpora from websites that reflect the current discourse of climate change and which we identify as representative for the two subgroups. This was done with the scope of building a discourse-oriented glossary. On the basis of the corpora we automatically retrieved words connected to climate change of the pattern ”KlimaX” (en. ”climateX”) and saved these words to lists. In a next step, these lists were semi-automatically cleaned and now compose a final list of 248 glossary terms. Some of these terms were already defined manually with respect to their function and use in discourse and their general meaning. Meanwhile, the terms were gathered in a web app and connected to Twitter and to DWDS entries to create a first attempt of a discourse-oriented glossary of climate change compounds.

2 Motivation

So far, we only provide definition texts for a very small subset of compound words, since those texts are phrased manually. To efficiently enrich our glossary I would now like to explore various computer-based methods to semi-automatically generate definition texts for our compound words. With the help of text mining tools and corpus-based approaches we seek to extract information from the corpus data which can be used for the final definition text of each compound. Furthermore it would be interesting to see if we can not only retrieve straightforward observations such as word relations and dependencies from the corpora. By exploiting the knowledge that we get from text mining and corpus-based approaches, we also wish to derive implied information about a term’s use in discourse. For the demand of our glossary, namely to display the discourse of a term, I attempt to include judgments about the use of the compounds as a self-attribution or by means of an external attribution to refer to the opposing subgroup.

3 Approach

To be able to capture information for the definition texts of the glossary terms, I want to make use of various text mining methods. A useful tool for extracting further information of the glossary terms is the extraction of hypernyms. With hypernyms and the detection of semantically related words through the application of word similarity measures, the relations of the glossary terms to each other can be displayed and further explored. To give an example, while compound words such as *Klimalügner* (en: "climate liar") and *Klimalüge* (en: "climate lie") can be associated by applying a stemmer which truncates the words to its, in this case, common stem, we also wish to connect compounds with a close semantic representation. When applying similarity measures to the compounds, terms such as *Klimaverbrechen* (en: "climate crime") and *Klimabetrug* (en: "climate fraud") display connectivity given that they are semantically related in so far as they both refer to a criminal action. Also, the analysis of the polarity of the context of a compound will provide further information about the specific connotation a word can be associated with. Even though one may expect the texts to reveal a polarity shift to either negative or positive sentiment, most of the corpus data will be tagged as neutral since many texts are technical articles and do not carry any intense polarity. Nevertheless, even this fact can be used for finding a definition of the glossary terms. Additional methods such as Named Entity Recognition will be applied to explore whether specific persons, organizations and locations can be associated with the compound. For instance, very commonly the corpus texts consist of phrases such as "Die Klimaaktivistin Greta Thunberg ..." (en: "The climate activist Greta Thunberg ...") from which we could extract the proper name via entity extraction methods. By the implementation of Dependency Parsing we seek to examine the dependents of the compound words. I.e. which words do have the compound word as their syntactic head. Here, we are particularly interested in the investigation of adjectives and adverbs that are dependent on the glossary term given that they act as modifiers of the compound. Very frequent modifiers will then be used for the definition phrasing part of this work. For the corpus-based methods, we rely on collocation measures to identify very frequent words co-occurring with the compound. Collocation analysis is a quantitative approach of information extraction since we can simply obtain frequency counts without the need of manual annotation. Extracting the collocations of the compound *Klimaskeptiker* (en: "climate skeptic"), for instance, shows the common use of the adjective "sogenannter" (en: "so-called") to the left of the compound in texts of the climate skeptics corpus. This suggests the rather sarcastically influenced use of this compound word. Concordance extraction, i.e. the retrieval of key word phrases, will be used to determine the immediate and broader context of a term. This context is then being used as an input for the various text mining techniques, mentioned before. As part of the exploration of the data set we also attempt to determine whether a compound is used in terms of a self-attribution or to externally refer to the opposing discourse actor. For the compound *Klimaleugner* (en: "climate denier") frequency analyses revealed that it is actually more frequently used in the discourse of climate skeptics. This is a surprising observation since this word is commonly used to negatively refer to climate skeptics. The manual analysis of the concordances of this compound then uncovered that the term is indeed used by climate skeptics in terms of a sarcastic or hyperbolical self-attribution. This observation is drawn from the immediate context and from the fact that the term primarily appears in quotation marks. Accordingly, a manual annotation of a subset of the keyword phrases to look for specific patterns that help with the assignment of attribution will be carried out. An annotation is crucial to find discourse cues which we can use as on indicator of self- or external attribution. On the basis of all the extracted knowledge a universal definition text will be phrased and filled up with the according information for each compound. The text pattern is created in a way that it fits to each compound word and consist of placeholders which will be specified by the observations we draw for each glossary term in the course of this work.

The intention of this work is to use the approaches we elaborated above to semi-automatically generate definitions for the present glossary terms. Also, we seek to develop a universal definition pattern that can be used in the future for new glossary terms. Thus, we not only wish to enrich the glossary in its current status but also want to simplify the addition of new terms by applying a predefined pipeline of techniques to extract information that is necessary for generating a final definition. The examples we referred to before suggest that this approach may be very fruitful and straightforward when being applied to all glossary terms and will result into valuable definition texts for our glossary.