Project 1. Basic Level Concepts and Conceptual vs. Perceptual Features

Supervisors: Simone Teufel and Yiwen Chen

Project Description

Rosch (1970) defined "basic level concepts" (BLC) as the most abstract categories which maximise the number of perceptual and functional attributes that distinguish these categories from their distractors. Examples for basic level objects are "chair" or "car"; they contain many attributes which distinguish them from basic level objects at the same level ("table" or "bus"). In contrast, more abstract categories ("furniture", "vehicle") share fewer attributes with each other, and less abstract categories ("kitchen chair", "limousine") have few additional attributes not shared with their respective basic level category. Basic level categories are also those that are learned earliest during child language acquisition. Rosch advanced a method for identifying basic level categories, based on two properties: "cue validity", a probabilistic property of attributes (cues), and "category resemblance" (originally described by Tversky). Both rely on the ready availability of symbolic features for each concept.

This project will explore to which degree modern distributional techniques can be used to shed light onto whether Rosch's theory can be substantiated based on naturally occurring text. Distributional techniques such as transformers and LSTMs compile a large set of features of a concept as high-dimensional vectors; Frassinelli and Keller (2012) have shown that high-dimensional vectors can be used as a stand-in for human-elicited features, but they operate with interpretable dimensions rather than uninterpretable ones.

This project will first test the hypothesis of cue validity. As the source of its features, it will use information from Wordnet, a large taxonomy of word senses. The student will study how features can be attached to concepts, their hyponyms and hypernyms, study the overlap of features and possibly learn how word sense disambiguation relates to our problem. On the basis of those concept -- feature associations it should then be possible to quantify the ability of cue validity and category resemblance to identify basic level categories. It will then distinguish conceptual features from visual features. For this part of the project, visual resources will be included, in a manner which is open to research.

Resources and Practicalities

An experiment similar to that of Mills et al (2018) can be set up and used for evaluation. A separate, independent data set with around 500 basic level categories exists, which the student can use as a gold standard for testing; feature norms from various sources are available (e.g., McRae 2005, Buchanan et al 2019). BERT-Vectors trained on Wikipedia are also available. Extensions to the project are possible and might involve the use of visual features for BLC identification, e.g. using Google Quick! Draw data (human sketching), YOLO9000 data or imagenet (image recognition). This project is open-ended and requires the student to be independent and curious. The student would be co-supervised by my PhD student Yiwen Chen, who studied BLC in the framework of her own PhD, in connection with the generation of song lyrics.

Literature

- Tversky (1977), Features of Similarity, Psychological Review.
• Buchanan, Valentine, Maxwell: English semantic feature production norms: An extended database of 4436 concepts, Behavioral Research Methods
• Chen, Teufel (2021). Synthetic textual features for the large-scale detection of basic-level categories in English and Mandarin. EMNLP.
Conceptual metaphors is a theory due to Lakoff that describes metaphors as connecting a target domain (e.g., arguments) with a source domain (e.g., war). This means that many concepts originally and literally describing acts of war can be metaphorically used to describe arguments. To which degree is it possible to find evidence for this theory in real-world texts? To this end, this project will look at different models for clustering real world occurrences of metaphors. Students would start by familiarising themselves with metaphor identification as in the 2018 VUA Shared Task of Metaphor Detection, possibly reimplement and train a simple metaphor identifier and then study different clustering techniques for metaphors according to their conceptual groups. As the desired algorithm is unsupervised, a model needs to be designed that uses outside information, for instance modelling of target and source domain s by topics. The baseline that the system runs against is the SOA in topic-based clustering, as a conceptual metaphor finder should go beyond simply clustering by keywords or even topics. The Shutova algorithm from 2017 that is able to find other instances of the same conceptual metaphor, rather than all conceptual metaphors, also acts as baseline.

Data Set

The VUA corpus for metaphor identification is available. A separate new conceptual-metaphor based corpus is also available for evaluation. This corpus has been newly created in the NLP group and has not been published with.

Literature

- Shutova, Sun, Korohonen (2010). Metaphor identification using verb and noun clustering. COLING.
Project 3: Automatic Identification of Creativity and Innovativeness in Scientific Writing

Supervisor: Simone Teufel

Project Description

This project proposes the development of an indicator of innovativeness, in order to improve bibliometric assessment of science. Bibliometrics is the science of assessing the quality of the research output of researchers or universities on the basis of their research output, for instance the UK's Research Excellence Framework [1]. A related task is IARPA's FUSE program [1], which seeks to detect emerging opportunities in science and technology as early as possible. Its fundamental hypothesis is that real-world processes of technical emergence leaves discernible traces in the public scientific, technical and patent literature. Most of the current science indicators are citation-based. The degree of innovativeness of a paper is an aspect of emergence that is closely related to this idea.

It is commonly believed that high impact papers are innovative. However, some highly cited papers are conforming, document incremental research and tend to reinforce the status quo [3]. Innovativeness cannot therefore be assessed by purely looking at citation count. One can try to approach the problem of identifying innovative scientific papers using citation networks to [3, 4]. This approach is based on the idea that innovative papers maximally disrupt the existing citation structure of the topic.

It has also been long assumed that access to full text would result in better innovation finding. This is exemplified by a related problem of identifying "paradigm shifts" [5]. The current project follows along this research avenue, and attempts to add information about sentences such as the following to the search for innovativeness:

*This result challenges the claims of recent discourse theories (Grosz and Sidner 1986; Reichman 1985) which argue for the close relation between cue words and discourse structure.*

A gold standard about innovativeness, collected by my US collaborators Richard Klavans and Kevin Boyack, is available. Klavans and Boyak asked highly influential biomedical scientists to rate 10 of their high-cited papers. Despite the deeply subjective nature of innovativeness, the authors themselves are certainly in the best position to assess how innovative their own papers are, if the self-elicitation is performed in an honest, trusted manner, where the reputation of the informant is not threatened. Klavans and Boyak achieved this by asking only about those papers which are of high-impact anyway. Ranking data for about 400 papers is available. One possible methodology for this project relies on performing Argumentative Zoning classification on the corpus of full-text for the 1200 papers, and to find a correlation between the rhetorical "footprint" of a paper (derived via AZ) and its level of innovation. The "rhetorical footprint" will be based on AZ-based features, which are fed into a machine learning system that correlates these features to the papers' innovativeness status. Other features based on simple sentiment detection are equally possible. On the other end of the feature space, one might use the bibliometric intuition that innovative papers cite papers that "lie far away from each other in conventional citation space"; it would be up to the student to operationalize this notion. The best possible project I could imagine would combine such bibliometric and textbased features.

Practicalities
The student would run AZ on a large biology dataset in SciXML format. They would create a citation network for this corpus, which does not exist yet. Alternatively, they could reverse engineer a full text corpus starting from an existing citation network. In the end, a parallel dataset must be created. The student would then, starting from their own intuition of creativity, study and fine-tune simple NLP and bibliometric features, each of which might be the outcome of its own ML task.

**Literature**