Towards Evaluating Creativity in Language

Matey Krastev

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Bachelor of Science
of the
University of Aberdeen.

Department of Computing Science

2022

Declaration

No portion of the work contained in this document has been submitted in support of an application for a degree or qualification of this or any other university or other institution of learning. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Signed:

Date: 2022

Abstract

An expansion of the title and contraction of the thesis.

Acknowledgements

Much stuff borrowed from elsewhere.

Contents

Chapter 1

Introduction

1.1 Inspiration

At the heart of the field of artificial intelligence is the concept of reproducing aspects defining human intelligence through rigourous examination and replication of the mechanisms that drive progress. This is a uniquely multi-faceted problem with a multitude of approaches, each tailored to a very specific aspect or manifestations of intelligence. Naturally, intelligence assumes following a logical pathway to arrive at sensible conclusions that interact with the real world in a beneficial way. This can be viewed through the lens of methodical, defined process that always follows a certain formula. An organism, assumed to be intelligent, might always follow such formulaic actions given a set of prerequisite conditions to accomplish a defined goal. Certain schools of thought theorise that there is such an order in every little action, and such thread of logicality interweaves every law of nature, known or not. Then, follows the question, can we recognise and define such a thread for the ambitious field of creativity? We seek not to properly define or constrain the subject of creativity, rather, we explore markers of what could only be a subset of the very broad field of human creativity.

The subject of the present document is exclusively the study of linguistic creativity. Henceforth, we seek to: confirm prior results of the research of psycholinguistics, affirm that hypotheses and conclusions drawn from them correlate highly with certain manifestations of creativity, and make firm the subject matter of creativity, that is, provide tools that may be used for exploration and analysis of specific creative features found in text.

Chapter 2

Methodology

2.1 Datasets

Some introductory text on the purpose and uses of the datasets...

2.1.1 Brown Corpus

The Brown Corpus (Francis and Kucera, 1979) is a widely used corpus in the field of computational linguistics, noted for the small variety of genres of literature it contains. The Corpus itself is founded on a compilation of American English literature from the year 1961. It is also small in terms of size, totalling around one million words, at least compared to modern corpora, which we also explore later on. The corpus also suffers from the issue of recency, as the works and language may be outdated for modern speakers of English.

Of interest is the fact that the corpus has been manually tagged for parts of speech, a process that tends to be error-prone. As we will see later on, this fact has implications in terms of the supervised learning algorithms we implement for creativity evaluation. Still, we opt to utilise it primarily for prototyping purposes and drawing preliminary conclusions about the effectiveness of the implemented algorithms, rather than in-depth analysis and publication of results.

2.1.2 Project Gutenberg

Project Gutenberg[1] is a large collection of more than 50,000 works available in the public domain. The collection contains literature from various years and various genres and thus is suitable for training and evaluation of the developed benchmarks in the context of creativity study.

As the Project does not offer an easy to process copy of its collection, we turn to the work of Gerlach and Font-Clos (2018). The team developed a catalogue for on-demand download of the entire set of books available on the Project Gutenberg website, intended for use in the study of computational linguistics. The tool avoids the overhead of writing a web-scraper or a manual parser for the downloadable collections of Project Gutenberg books made available by third parties, as well as enables easy synchronization of newly released literature. Instead, we are only required to develop a simple pipeline for the data to be fed into the utilised systems.

2.1.3 Hierarchical Neural Story Generation

Fan et al. (2018)

[1] `https://www.gutenberg.org/`

2.1.4 WordNet

WordNet(Fellbaum, 1998) is a lexical database of semantic relations between words that links words into semantic relations including synonyms, hyponyms, and meronyms. The synonyms are grouped into synsets with short definitions and usage examples. It can thus be seen as a combination and extension of a dictionary and thesaurus (Wikipedia contributors, 2023).

For our specific use cases, we have identified it as a valuable resource in terms of relational representation of words in semantic space. In the given context, this enables us to traverse a semantic graph for synonyms and related words for the goal of enriching potential similarity between the set of creative parts of speech (i.e., nouns, adjectives, adverbs), which we narrow down our scope to in particular.

2.1.5 Word2Vec

Mikolov et al. (2013) show in their work that words may be represented as dense vectors in N-dimensional space, and we can perform mathematical operations on them that may yield effective results in terms of word representation.

Measuring distance in vector representations of semantic tokens

Intuition tells us that the dot product of vectors in N-dimensional space will grow when the set of vectors has similar values and decrease when the values are not similar. Thus, we can then construct the following metric for semantic similarity between vector representations of words:

$$D(v,w) = v \times w = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + \cdots + v_N w_N$$

The current metric, however, suffers from the problem that vectors of higher dimensions will inevitably be larger than vectors with lower dimensions. Furthermore, embedding vectors for words that occur frequently in text, tend to have high values in more dimensions, that is, they correlate with more words. The proposed solution is to normalize using the **vector length** as defined:

$$|v| = \sqrt{\sum_{i=1}^{N} v_i^2}$$

Therefore, we obtain the following:

$$\text{Similarity}(v,w) = \frac{v \times w}{|v||w|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

This product turns out to be the same as the cosine of the angle between two vectors:

$$\frac{a \times b}{|a||b|} = \cos(\theta)$$

Therefore, we will call this metric the **cosine similarity** of two words. As mentioned, the similarity grows for vectors with similar features along the same dimensions. Note the boundaries of said cosine metric: we get -1 for vectors which are polar opposites, 0 for orthogonal vectors, and 1 for equivalent vectors. Of note is the fact that such learned vector embeddings only have values in the positive ranges, thus, it is impossible to have negative values for the cosine similarity

(Similarity$(a,b) \in [0,1]$).

Contrary to it, we also identify the metric of **cosine distance** between two vectors, as one minus the similarity of the vectors, or:

$$\text{Distance}(v,w) = 1 - \text{Similarity}(v,w)$$

The cosine distance may prove useful when dealing with minimisation problems as is often the case with machine learning.

2.2 Metrics

2.3 Functional Requirements

2.4 Non-functional Requirements

-

Bibliography

Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical Neural Story Generation. arXiv:1805.04833 [cs].

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Francis, W. N. and Kucera, H. (1979). Brown corpus manual. *Letters to the Editor*, 5(2):7.

Gerlach, M. and Font-Clos, F. (2018). A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *CoRR*, abs/1812.08092.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs].

Wikipedia contributors (2023). Wordnet — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=WordNet&oldid=1143619785`. [Online; accessed 14-March-2023].