Matching Poems in a Parallel Corpus using Concept Networks

Aubrey Barnard

Abstract

Many natural language processing approaches, and machine learning approaches in general, focus on learning "rich" concepts from "poor" representations. An example would be detecting humor in bag-of-word vectors. In this work, I investigate enriching the representation of text using concept networks, and apply the approach to matching poems between languages in a parallel corpus. The initial results show promise, but there is much more to investigate.

Introduction

In the quest for better and better natural language processing (NLP) techniques, researchers explore increasingly sophisticated computational models. Yet, the ways in which they represent text are often limited. I like to think of this situation as trying to learn "rich" concepts from "poor" representations. That is, typical approaches place the burden of learning and inference on the computational model rather than the text representation. These approaches seem somewhat unnatural to me when I compare them to my understanding of how the brain works, that humans learn simple concepts from rich representations. Therefore, I was curious to see what would happen when trying to reverse this trend and learn a poor concept from a rich representation.

The Corpus

The Concept Network

Approach

Distances

$$d_{cosine}(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{\|\mathbf{x}_1\|\|\mathbf{x}_2\|}$$

$$d_{Hamming}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{V} |\text{sign}(x_{1,i}) - \text{sign}(x_{2,i})|$$

(von Goethe 2009a) (von Goethe 2009b) (Murthy, Keerthi, and Murty 2007) (Gregorowicz and Kramer 2006) (Steyvers and Tenenbaum 2005) (Google 2009)

Poem Set	*Method*	*Accuracy*	
Google-En	Cosine	53/127	41.7%
Google-En	Euclidean	58/127	45.7%
German	Concept-Euclidean	71/127	55.9%
German	Concept-L1Norm	74/127	58.2%
Google-En	Hamming	91/127	71.7%
Google-En	L1Norm	101/127	79.5%

Table 2: Matching accuracy by poem set and method.

Results

Evaluation

Conclusion

References

Google. 2009. Google language tools.

Gregorowicz, A., and Kramer, M. A. 2006. Mining a large-scale term-concept network from wikipedia. Technical report, Mitre.

Murthy, K. R. K.; Keerthi, S. S.; and Murty, M. N. 2007. Concept network: A structure for context sensitive document representation. Technical report, Indian Institute of Science.

Steyvers, M., and Tenenbaum, J. B. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science* 29:41–78.

von Goethe, J. W. 2009a. Johann Wolfgang von Goethe: Gedichte. `http://www.wissen-im-netz.info/literatur/goethe/gedichte/index.htm`. Published by Jürgen Kühnle.

von Goethe, J. W. 2009b. The poems of Goethe. `http://www.gutenberg.org/etext/1287`. Translated by Edgar Alred Bowring. Published by Project Gutenberg.

Und pflanzt’ es wieder	In silent corner	And planted it again
Am stillen Ort;	Soon it was set;	On quiet place;
Nun zweigt es immer	There grows it ever,	Now branches are always
Und blüht so fort.	There blooms it yet.	And so forth blossoms.
German	*English*	*Google-English*

Table 1: The last stanza of “Gefunden” (“Found”) in the three languages.