Detecting Metaphors in Textual Data

Kartik Goyal
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
kartikgo@cs.cmu.edu

Shriphani Palakodety
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
spalakod@cs.cmu.edu

Thom Popovici
ECE
Carnegie Mellon University
Pittsburgh, PA 15213
dpopovic@andrew.cmu.edu

1 Project Idea

Our project deals with detecting metaphors in textual data. A metaphor is a figure of speech in which a word or phrase is applied to an object or action to which it is not literally applicable. In this project, our hypothesis is that metaphors used in language are generated from a thought process much similar to the one behind generating semantics and topics responsible for the surface form of text. Previous generative models propose that there exists a background distribution for the text and each latent topic modifies this background distribution. In addition, a perspective which influences these topic distrbutions can be modelled as a further perturbation to these distributions Eisenstein et al. [2011]. We aim to model metaphors as such a perspective. We also draw analogies from Eisenstein et al. [2010], which attempts to model the lexical variations according to geography as a multi-generative model whose observables are both text and the geotags, where the metaphors are analogous to the geotags.
The area of metaphor detection using topic models has not received a lot of attention. Klebanov et al. [2009] proposed that the words representing topics are generally not metaphorical and their experiments validate this hypothesis. Li et al. [2010] use topic models for word-sense disambiguation and Idiom detection using a corpus of paraphrases.

2 Dataset

We plan to use two kinds of data sources:

- Annotated Data: We have 1300 annotated sentences, which identify the exact words which play a metaphorical role in the corresponding sentence. This data set will aid in supervised settings and evaluation.
- Unannotated Corpus: We plan to obtain metaphor-rich content from the World Wide Web. We are currently targetting various blogs, public domain literature and movie scripts. This dataset will be used extensively in unsupervised settings.

3 Software

We will write code for implementing the generative model we formulate for metaphor detection. In addition,we will also write a web crawler and a scraper for collecting data to build our unannotated corpus.

4 Midterm Milestone

We plan to implement a basic generative model based on our readings and approach. We will also set up a test bench for evaluation of our model and its comparison with a baseline approach for our task. Our whole team will contribute equally, to all the aspects of the project.

References

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. Sparse additive generative models of text. In *Proceedings of ICML*, pages 1041–1048, 2011.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. Discourse topics and metaphors. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 1–8. Association for Computational Linguistics, 2009.

Linlin Li, Benjamin Roth, and Caroline Sporleder. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1138–1147. Association for Computational Linguistics, 2010.