SSIE 519: Applied Soft Computing

Project Proposal

Project Title: Investigating Homographic Pun Distances via Semantic Embeddings and NLP
Team: Srikanth Iyer, Annajirao Challa.

Project Objective

In this project, we would like to evaluate soft-computing mechanisms to map semantic spaces of different bodies of literature. The aim is to quantify (or at least, try to) the semantic distance between different polysemous interpretations of words within a novel, particularly focusing on how this distance impacts reader comprehension and engagement. By comparing the semantic spaces of well-known bodies of comedy (E.g. P.G. Wodehouse, Douglas Adams, Shakespeare), we wish to use known methods of identification of puns and attempt to calculate the semantic leaps made by the puns and evaluate the nature of comedy to other non-comedy pieces of work using this semantic distance.

Introduction

Words can have multiple meanings (polysemy), and the specific meaning intended by the author is crucial for interpreting a text. Most words in natural languages are polysemous; that is, they have related but different meanings in different contexts.[1] Traditional approaches to automatic word sense disambiguation (WSD) rest on the assumption that there exists a single, unambiguous communicative intention underlying every word in a document. [2][3] This one-to-many mapping of form to meaning presents a challenge to understanding how word meanings are learned, represented, and processed. Lexical polysemy, a fundamental characteristic of all human languages, has long been regarded as a major challenge to machine translation, human–computer interaction, and other applications of computational natural language processing (NLP).

However, writers sometimes intend for a word to be interpreted as simultaneously carrying multiple distinct meanings. In literature, polysemy is a powerful tool for authors to create ambiguity, suspense, and layered meaning. This deliberate use of lexical ambiguity — i.e. punning — is a particularly common source of humor, and therefore has important implications for how NLP systems process documents and interact with users.[4] Measuring the semantic distance between polysemous interpretations can help us evaluate authors who frequently employ polysemy and explore how the distance between meanings affects reader experience. For this project, we will focus primarily on Homographic Puns. Homographic puns are puns that exploit words that are spelled the same (homographs) but possess different meanings and sounds.

Homographic pun examples include:

- After hours of waiting for the bowling alley to open, we finally got the ball **rolling**.
- Always trust a glue salesman. They tend to **stick** to their word.
- Every calendar's days are **numbered**.
- A dog gave birth to puppies near the road and was cited for **littering**.
- If you don't pay your exorcist, you will get **repossessed**.

Traditional humor theory points out that laughter is the emotion produced by a sudden change from nervous expectation to failure. The argument is called the incongruity theory of humor. The incongruity theory of humor implies that the punchline of humor is caused by two or more situations that are inconsistent but unified with special connections. Since the pun is one of the most important forms of humor, the incongruity theory can also be applied to puns. Humor and pun are closely related, but not exactly the same.[5]

Methodology

Our aim is to use tools introduced to us in SSIE 519: Soft Computing as well as other methods found in our research.

Latent Semantic Analysis (LSA) can uncover the underlying structure in word usage across large text corpora, which can be useful for identifying semantic incoherence in puns. This method can calculate semantic distance differences by comparing word embedding and language models.

Pre-trained word embeddings (e.g., Word2Vec, GloVe) capture semantic relationships but may not account for context-specific polysemy. We wish to utilize contextualized word embeddings (e.g., ELMo, BERT [6]) that consider the surrounding words to capture the specific meaning intended in the context of the novel.

Another option we wish to evaluate is using the SemD measure of semantic ambiguity.[7] SemD, or Semantic Density, is a measure used to evaluate the richness of word meanings in a text. To calculate SemD for a particular word, the authors of the paper examined all of the contexts in which the word appeared and calculated their average similarity to one another. When the contexts were very similar to one another on average, this suggested that the word was associated with a fairly restricted set of meanings and was relatively unambiguous. When the contexts associated with a given word were quite dissimilar to one another, this suggested that the meaning of the word was more ambiguous.

This method, however, may not be useful in calculating specific “semantic leaps” of puns, but more in the semantic diversity of the whole corpus of an author's work. Do comedies have higher or lower semantic density compared to non-comedies?

In addition to this, we will continue to explore mechanisms of Word Sense Disambiguation (WSD). The landscape of NLP and Semantic structures is rapidly evolving. By the time of finishing this sentence, we are certain there will have been published newer and more effective models of evaluating semantic relations.

Approaches to evaluate Semantic distances:

Cosine Similarity: This measures the angle between the two word embedding vectors. A smaller angle indicates closer semantic meaning.[8]
Euclidean Distance: This calculates the straight-line distance between the two word embedding vectors. A smaller distance implies higher semantic similarity. By taking known Pun-based word pairs from Pun datasets (E.g. SemEval-2017 [9]) we can measure the average semantic distances of these known words within the specific body of an author’s work.

Data Sources

1. **Project Gutenberg:** We will utilize novels from Project Gutenberg, a vast collection of public domain ebooks, offering a diverse range of writing styles and genres. The rationale is to leverage freely available data and focus on established literary works where polysemy is likely to be a significant stylistic feature.
2. **Open Library:** We also have an Open Library as a backup that provides a wealth of information on novels and other ebooks, including metadata such as author, title, publication date, and more. It's a promising resource due to its extensive catalog and open-access nature.

References

[1] Li, J., & Joanisse, M. F. (2021). Word Senses as Clusters of Meaning Modulations: A Computational Model of Polysemy. Cognitive Science, 45(4), e12955. https://doi.org/10.1111/cogs.12955

[2] Miller, T., & Turković, M. (2016). Towards the Automatic Detection and Identification of English Puns. The European Journal of Humour Research, 4(1), 59–75. https://doi.org/10.7592/EJHR2016.4.1.miller

[3] Li, J., & Joanisse, M. F. (2021). Word Senses as Clusters of Meaning Modulations: A Computational Model of Polysemy. Cognitive Science, 45, e12955. https://doi.org/10.1111/cogs.12955

[4] Carston, R. (2021). Polysemy: Pragmatics and sense conventions. Mind & Language, 36, 108–133. https://doi.org/10.1111/mila.12329

[5] Ren, L., Xu, B., Lin, H., et al. (2021). ABML: attention-based multi-task learning for jointly humor recognition and pun detection. Soft Computing, 25, 14109–14118. https://doi.org/10.1007/s00500-021-06136-y

[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv. https://doi.org/10.48550/arXiv.1810.04805

[7] Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic Diversity: A Measure of Semantic Ambiguity Based on Variability in the Contextual Usage of Words. Behavior Research Methods, 45(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

[8] Diao, Y., Yang, L., Zhang, D., Xu, L., Fan, X., Wu, D., & Lin, H. (2018). Homographic Puns Recognition Based on Latent Semantic Structures. In X. Huang, J. Jiang, D. Zhao, Y. Feng, & Y. Hong (Eds.), Natural Language Processing and Chinese Computing (pp. 565–576). Springer International Publishing. https://doi.org/10.1007/978-3-319-73618-1_47

[9] Miller, T., Hempelmann, C., & Gurevych, I. (2017). SemEval-2017 Task 7: Detection and Interpretation of English Puns. In S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, & D. Jurgens (Eds.), Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 58–68). Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-2005