A Geometric Measure of Polysemy in Hindi Language

Anmol Goel
GGSIPU, Delhi
agoel00@gmail.com

Ponnurangam Kumaraguru
IIIT-Delhi
pk@iiitd.ac.in

ABSTRACT

A word referring to two or more different meanings is called *polysemous*. In this study, we introduce a geometric method to estimate the polysemy of words using the discrete Ollivier-Ricci curvature of a graph of synonyms in Hindi. We show that this approach can effectively measure the polysemy of words and is strongly correlated with theoretical interpretations of polysemy.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**.

ACM Reference Format:
Anmol Goel and Ponnurangam Kumaraguru. 2021. A Geometric Measure of Polysemy in Hindi Language. In *8th ACM IKDD CODS and 26th COMAD (CODS COMAD 2021), January 2–4, 2021, Bangalore, India.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3430984.3431044

1 INTRODUCTION

The different senses a word can have is known as its polysemy. Detecting and measuring polysemous words is an important task in NLP. Such information helps interpret the evolution and characteristics of a language. Previous works have focused on detecting different word senses in high resource languages like English, French, etc. The contributions of this study are twofold - a) We introduce a method to quantify the polysemy of a word instead of just detecting whether a word is polysemous or not; b) we study polysemy in the context of Hindi which has received relatively less attention than works involving English, French, etc.

2 RICCI CURVATURE

We borrow the concept of curvature from Riemannian geometry to study the graph of synonyms. In particular, we utilise the **discrete Ollivier-Ricci curvature** [1] which involves solving an optimal transport problem of Wasserstein distances between probability measures on the vertices of an edge. An interesting property of this measure is that edges with positive Ricci curvature are likely to be in dense communities while negatively curved edges connect distinct communities.

3 APPROACH

We first construct a graph of synonyms by accessing the Hindi WordNet using the pyiwn library. We denote words as nodes and connect two words with an unweighted edge if they occur in the same synset of the WordNet similar to [2]. To remove sparsity from the graph, we delete the words which have less than 5 synonyms. More formally, we obtain a graph $G(V,E)$ where V is the set of unique words and E is the set of edges connecting synonyms. To study the properties of this graph using its geometry, we obtain the Ricci curvature of each edge $e \in E$. The dataset used in this study is described in Table 1.

Nodes	Edges	Average Degree
9224	34335	7.44

Table 1: Details of the graph of synonyms

4 DISCUSSION

We propose to consider the average Ricci curvature of all edges incident on a word as a measure of its polysemy. A negatively curved incident edge on this word connects it to a synonym with a different semantic group while a positive incident edge connects synonyms which have similar semantic context and thus belong to the same semantic group. This intuition provides a simple way to measure the polysemy of words and trace their evolution in a language. Interestingly, Figure 1 shows that the degree of a word is correlated with its polysemy (average curvature in this case). In our experiments, we found that their **Spearman correlation is -0.84**. A word with a higher degree in the graph of synonyms tends to be more polysemous. This also implies that the higher frequency words acquire more meanings in a language [2] which is popularly known as **Zipf's Law**. This shows that our approach describes linguistic phenomena reasonably well and is in line with previous knowledge about language polysemy.

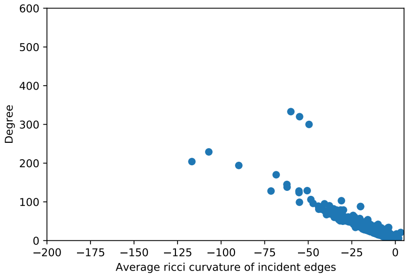

Figure 1: Average curvature and Degree correlation

Polysemy is an important linguistic phenomenon and to the best of our knowledge, this study is the first to empirically measure it instead of just detecting it in Hindi.

REFERENCES

[1] C. Ni, Y. Lin, J. Gao, X. David Gu, and E. Saucan. 2015. Ricci curvature of the Internet topology. In *2015 IEEE Conference on Computer Communications (INFOCOM)*. 2758–2766.

[2] Ivan P. Yamshchikov, Cyrille Merleau Nono Saha, Igor Samenko, and Jürgen Jost. 2020. It Means More if It Sounds Good: Yet Another Hypotheses Concerning the Evolution of Polysemous Words. arXiv:2003.05758 [cs.CL]

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CODS COMAD 2021, January 2–4, 2021, Bangalore, India
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8817-7/21/01.
https://doi.org/10.1145/3430984.3431044