COMPUTATIONAL LINGUISTICS I PROJECT
TOPOLOGICAL DATA ANALYSIS FOR WORD SENSE DISAMBIGUATION

INTRODUCTION:

Topological Data Analysis (TDA) provides a general framework to extract information from high dimensional, incomplete and noisy datasets. Persistent homology is an algebraic tool for measuring topological features of data. The aim of this project is to use persistent homology to find cycles in the local homology of embedding representations of words and develop it into an unsupervised clustering algorithm to disambiguate the different senses of a word.

MOTIVATION:

TDA has become a very active and broad area of research but its application in natural language processing (NLP) is in its infancy. Using persistent homology to analyze the distribution of word embeddings is a challenging and a relatively unexplored area. Further, the algorithm using the local homology of data-points is unsupervised and implementations of such algorithms run in polynomial time.

WHY TDA?

Word embedding is a collective term for ways to represent words of a natural language as vectors in a high-dimensional real vector space. They have interesting topology and hence NLP seems to be a natural domain of TDA applications. Despite this fact, only a few attempts at using TDA techniques to analyze language data have been published.

Intuitively, stratification is a decomposition of a topological space into manifold like pieces. When thinking about stratification learning and word embeddings, it seems intuitive that vectors of words corresponding to the same broad topic would constitute a structure, which we might hope to be a manifold. Hence, for example, by looking at the intersections between those manifolds one might hope to algorithmically find vectors of homonyms like ‘bank’ (which can mean either a river bank, or a financial institution) or vectors of words with very different meanings like ‘cancer’ (which can refer to the Cancer constellation or to the illness). This, in turn, has the potential to help solve the word sense disambiguation (WSD) problem in NLP

PERSISTENT HOMOLOGY

The central object in algebraic topology is a simplicial complex K, e.g. an undirected and weighted connected graph. Persistent homology captures topological features of K on different scales by defining a filtration on it. Figure 1 illustrates this growth process for a 2-dimensional point cloud. With a defined metric in our space (e.g. the Euclidean distance), we keep track of the generation and destruction of connected components (CC) while increasing the distance threshold (i.e. a point’s range of vision). Note that in this example each point in the 0^{th} filtration-step (left most panel in figure 1) is considered as its own connected component. Increasing the threshold therefore leads to a decreasing number of connected components. To be more precise, we start with 16 CCs, observe 11 CCs at step 1, and arrive at 1 CC at step 2 and 3

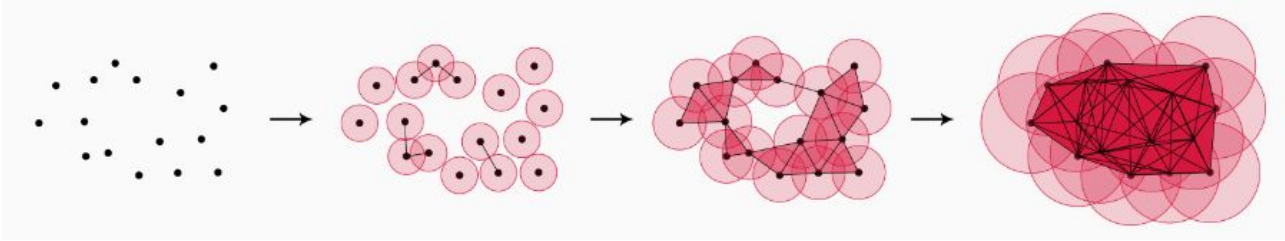
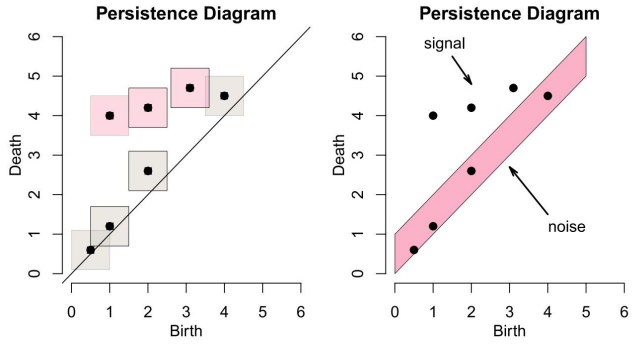


Fig1: The filtration process applied to a 2-dimensional point cloud. At each step m, we increase the distance threshold ϵ and therefore destroy existing CCs. We can define the edges of the graph as $E:=\{(u,v)|dist(u,v)\leq\epsilon_m\}$. As you can see in the second filtration step, we can also keep track of higher dimensional topological features like cycles and voids.

IMPLEMENTATION:

- Cluster the words into different connected groups using persistent homology.
- Persistent homology gives the birth and death time for the clusters. The death time is just the ϵ at which clusters merge. Since clusters are contractible, the birth time is always ϵ=0. On a persistence diagram (Birth time vs Death time), any cluster close to the $x = y$ line i.e the difference between the birth and the death time is below some threshold is considered noise. This is because by sampling a manifold, we expect points near each other to correspond to one piece of the manifold. So the short lifespan of these points does not correspond to pieces of the manifold and thus should be ignored as noise.
- The remaining clusters for the word can be considered as an estimate to the number of different meanings and the other words in each of the clusters are associated with different meanings.



WHAT DIFFERENCE WILL IT MAKE IF WE'RE SUCCESSFUL?

If we're successful it'll result in the development of an unsupervised algorithm to cluster the embeddings and infer the different senses of a given word. It will also explore the topology of the embedding space which seem to exhibit some interesting properties. We will be using TDA and Persistence Homology to identify these qualities which is a relatively unexplored and promising area in NLP.

EVALUATION:

Evaluation can be done in two ways against any word sense disambiguation (WSD) dataset:

1. Compare the number of clusters from the algorithm to the number of meanings for the word according to the dataset.
2. Evaluating the sense assigned to the word by the algorithm to be disambiguated against the dataset.

FEEDBACK SUMMARY

Although almost all the comments about the main goal of the project were accurate, understanding the algorithm including topological data analysis and persistent homology which is the crux of the project is not easy. Understanding the birth and death time for connected components and how that relates to filtering noisy data is hard to grasp with little knowledge about persistent homology.

Some of the additional feedback received were to evaluate the performance of the algorithm for different embeddings (such as GloVe, Word2Vec, etc) and the senses predicted by the algorithm and compare them. Exploring the extension of this algorithm into a multilingual setting. Comparing the performance of the algorithm (evaluation metric and runtime) to some of the state of the art unsupervised WSD algorithms. Additional inputs to evaluating the algorithm using some of the standard WSD tasks using SemEval dataset. Input regarding reference materials where TDA is used for classification tasks, multilingual embeddings and looking at translation as providing word senses.

TEAM MEMBERS: • Michael Rawson • Samuel Dooley • Rishabh Choudhary • Mithun Bharadwaj