Literature Review and Data Plan

Rishabh Choudhary, Mithun Bharadwaj Michael Rawson, and Samuel Dooley

Abstract

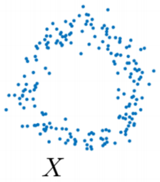
Recall that the crux of our project is to explore word sense disambiguation (WSD) with topological data analysis (TDA) in word embeddings. We believe that TDA can help identify provide a mathematically rigorous unsupervised clustering algorithm to disambiguate the different senses of a word. Note: the literature review is long to accommodate the background on topology – Hal approved.

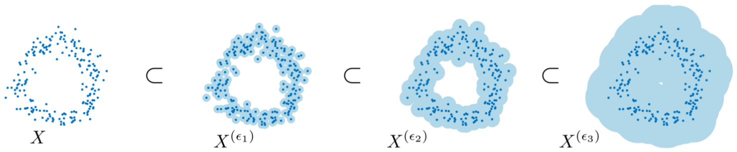
1 Literature Review

1.1 What is TDA and how did it develop?

We often want to understand the shape or topology of any dataset we are given. We may think there is a linear relationship between the variables and fit a linear model. We may believe that the data comes from a sphere and visualize them to test this theory. Often, we do not know what the shape of the data is, but believe the data has been sampled from some underlying surface or manifold lesser in dimension than the ambient space. The shape of this surface is interesting, but difficult to analyze so we focus on topological features which are much simpler. For example we check if the surface has holes or even multiple connected components which will correspond to clusters of the data.

To formalize this question, suppose you are given some (finite) data X which you believe come from some surface[1] $\mathbb{X}$. The critical question is whether we can infer the shape[2] (number of holes, compontents, etc.) of $\mathbb{X}$ from the given data X. Consider the data to the right[3]. Intuitively we see that this data comes from noisily sampling a circle because there is some hole in the middle of the data. But how would one algorithmically detect this for any given data? The most common way to do this is to use a concept called persistence. The core idea of persistence is to see how the union of balls around each point becomes connected as the balls grow in radius.



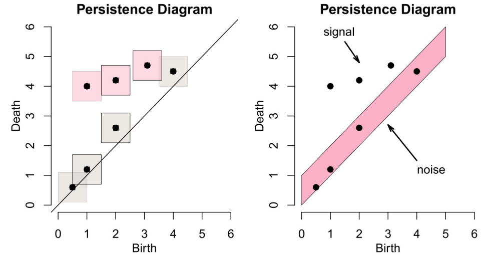


We see that as we grow an ϵ-ball around each point, the union of balls becomes more connected until we have all balls connected and 1 hole in the middle of the union. In persistent homology, we examine how these connections evolve overtime (time being the ball radius increasing). For the different discriminators of different shapes (connected components, holes, etc), we can observe their birth and death as we increase ϵ. By focusing on those discriminators which have a long life, i.e., persist, we can begin to understand the shape of our data.

[1] technically some embedded manifold

[2] technically the homology

[3] Figures taken from reference [1]



This field has been studied for a long time, dating back to the 1940s ([2], [3], and [4]). These works focus on the theory of topology, but they were restricted in their applicability due to their lack of computation. Then [5] introduced a fast algorithm to compute a surface's topology, and this led to the focus of computational techniques in the field of topology. After the publication of this work, many subsequent work was focused on advancing the theory and computability of the shape of surfaces. For an abstract overview of these works see [6]. For a history of persistent homology in TDA see [1]. For more mathematical and technical descriptions of the field see [7], [8], and [9].

There are only six published works at the intersection of NLP and TDA ([10], [11], [12], [13], [14], [15]). The first concrete example of a successful application of TDA to NLP comes from [10] which demonstrates the difficulties and possibilities for computational topology to analyze the similarities within a collection of text documents. [14] argues for applicability of persistent homology to lexical analysis using word embeddings. This paper aims towards the same goal as ours, but does so using a slightly different TDA method, which we believe we can improve upon. [15] applies topological data analysis (TDA) to entailment, with an improvement of accuracy over the baseline without persistence.

2 Word Sense Disambiguation

WSD was first framed as a computational task in the 1940s with Zipf's power-law theory. Since then, the types of solutions to WSD can be binned into three categories: knowledge-based, supervised, and unsupervised. An example of the knowledge-based approach is [16] where semantic similarity is used to measure distances between words which theoretically "measure" how much words are semantically similar. The approach taken in this example is different from our unsupervised approach, but the end goal is the same. [16] support their claim of the appropriateness of (their specific) similarity measures by using real time implicit feedback from user. An example of the supervised approach is [17] which uses a combination of different models: Naive Bayes, maximum entropy model, boosting, and kernel-PCA. This paper is the first of its kind to provide evidence that WSD can be used to improve the performance of statistical machine translation (SMT) tasks. Like the previous example, the goal is the same as ours but the approach is different. [17] justify their claim using experiments comparing SMT against SMT + WSD across a variety of tasks and a variety of performance measures. The final approach is an unsupervised approach, like [18]. These authors posit that graph-based centrality measures for word sense disambiguation can capture the necessary information. This approach is similar to ours in that there is an assumption that the geometry of a space carries some information about word senses. However, their work uses different similarity measures directly, while ours will use inferred measures through word embeddings.

3 Data Plan

The training and evaluation datasets have been taken from Linguistic Computing Laboratory group at the Sapienza University of Rome. The training data consists of two sense annotated training corpora, SemCor and OMSTI. Each dataset consists of a data file [dataset].data.xml and a gold file [dataset].gold.key.txt. All senses are annotated with WordNet 3.0.

The XML data file contains the following tags: corpus → text → sentence. Then, each sentences consists of "wf" (non-disambiguated) and "instance" (disambiguated) tags. Both types contain two mandatory attributes ("lemma" and "pos"). "instance" should also contains an id, which is in the gold file.

The .txt gold file contains all the disambiguation instances included in the XML. Each line is space-separated, where the first column corresponds to the instance id and the remaining columns correspond to the gold key/s. The given dataset is used to extract the corpus of words and a list of senses for each word.

References

[1] Jose A Perea. A brief history of persistence. *arXiv preprint arXiv:1809.03624*, 2018.

[2] Marston Morse. Rank and span in functional topology. *Annals of Mathematics*, pages 419–454, 1940.

[3] Patrizio Frosini. A distance for similarity classes of submanifolds of a euclidean space. *Bulletin of the Australian Mathematical Society*, 42(3):407–415, 1990.

[4] Vanessa Robins. Towards computing homology from finite approximations. In *Topology proceedings*, volume 24, pages 503–532, 1999.

[5] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 454–463. IEEE, 2000.

[6] Michael Lesnick. Studying the shape of data using topology.

[7] Robert Ghrist. Homological algebra and data. *Math. Data*, 25:273, 2018.

[8] Herbert Edelsbrunner and Dmitriy Morozov. Persistent homology: theory and practice. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2012.

[9] Frédéric Chazal. High-dimensional topological data analysis, 2016.

[10] Hubert Wagner, Paweł Dłotko, and Marian Mrozek. Computational topology in text mining. In *Computational Topology in Image Context*, pages 68–78. Springer, 2012.

[11] Xiaojin Zhu. Persistent homology: An introduction and a new text representation for natural language processing. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[12] Paul Michel, Abhilasha Ravichander, and Shruti Rijhwani. Does the geometry of word embeddings help document classification? a case study on persistent homology based representations. *arXiv preprint arXiv:1705.10900*, 2017.

[13] Ishrat Rahman Sami and Katayoun Farrahi. A simplified topological representation of text for local and global context. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1451–1456. ACM, 2017.

[14] Tadas Temčinas. Local homology of word embeddings. *arXiv preprint arXiv:1810.10136*, 2018.

[15] Ketki Savle, Wlodek Zadrozny, and Minwoo Lee. Topological data analysis for discourse semantics? In *Proceedings of the 13th International Conference on Computational Semantics-Student Papers*, pages 34–43, 2019.

[16] Kanika Mittal and Amita Jain. Word sense disambiguation method using semantic similarity measures and owa operator. *ICTACT Journal on Soft Computing*, 5(2), 2015.

[17] Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[18] Ravi Sinha and Rada Mihalcea. Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In *International Conference on Semantic Computing (ICSC 2007)*, pages 363–369. IEEE, 2007.