An Unsupervised, Geometric and Syntax-aware Quantification of Polysemy

Anmol Goel[†], **Charu Sharma**[†], **Ponnurangam Kumaraguru**[†]
[†]International Institute of Information Technology, Hyderabad
anmol.goel@research.iiit.ac.in
{charu.sharma, pk.guru}@iiit.ac.in

Abstract

Polysemy is the phenomenon where a single word form possesses two or more related senses. It is an extremely ubiquitous part of natural language and analyzing it has sparked rich discussions in the linguistics, psychology and philosophy communities alike. With scarce attention paid to polysemy in computational linguistics, and even scarcer attention toward quantifying polysemy, in this paper, we propose a novel, unsupervised framework to compute and estimate polysemy scores for words in multiple languages. We infuse our proposed quantification with syntactic knowledge in the form of dependency structures. This informs the final polysemy scores of the lexicon motivated by recent linguistic findings that suggest there is an implicit relation between syntax and ambiguity/polysemy. We adopt a graph based approach by computing the discrete Ollivier Ricci curvature on a graph of the contextual nearest neighbors. We test our framework on curated datasets controlling for different sense distributions of words in 3 typologically diverse languages - English, French and Spanish. The effectiveness of our framework is demonstrated by significant correlations of our quantification with expert human annotated language resources like WordNet. We observe a 0.3 point increase in the correlation coefficient as compared to previous quantification studies in English. Our research leverages contextual language models and syntactic structures to empirically support the widely held theoretical linguistic notion that syntax is intricately linked to ambiguity/polysemy.

https://github.com/agoel00/polysemy

1 Introduction

Polysemy is a phenomenon prevalent in everyday language use where the same lexical unit (or word form) is associated with multiple distinct yet *related* meanings (or senses). Determining which words are polysemous can help in filtering data for linguist studies, creation of sense corpora and the anthropological study of language. Consider the following sentences:

1a His **aunt** is his legal guardian.

2a The dog would always **bark** at mailmen.

2b The tree's **bark** was rusty brown.

3a The **mouth** of the wine was dry.

3b I have three **mouths** to feed.

3c You can see the **mouth** of the river from here.

Polysemy is distinct from monosemy (a word form with only one meaning; 1a) and homonymy (multiple *unrelated* senses of the same word form; 2a-b). The polysemous senses of a word often have metonymic (3a-b) or metaphorical (3c) relations among them (Vicente and Falkum, 2017). Polysemy is a central feature of natural languages and proliferates almost every word to varying degrees in the lexicon of a language. Attempts (Piantadosi et al., 2012) at explaining the presence of ambiguity[1] in language suggest that polysemy is a desirable property for language systems since it allows efficient communication by allowing simpler units to be reused. Ambiguity and polysemy have sparked debate among linguists and philosophers for decades but relatively little attention has been paid to analyze and measure polysemy in language by computational linguists. While a human listener is easily able to disambiguate the specific sense of the word being used in context, it is notoriously difficult for NLP systems to separate the distinct senses of a word being used (Yenicelik et al., 2020).

Recently, there has been widespread attention on including syntactic knowledge in various computational linguistic systems and studies - ranging from

[1]In the context of this paper, we use ambiguity and polysemy of a word form interchangeably.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10565 - 10574

syntax aware language models (Zhou et al., 2020) to syntax informed sentiment analysis (Hou et al., 2021). Recent works have identified (Čech et al., 2017) an intricate link between the syntactic properties of a lexical unit and its ambiguity (or lack thereof) since the meaning of a word is influenced by its syntactic as well as semantic context. The fact that most open class word forms are associated with multiple related senses hints at the possible role that syntax plays in influencing polysemy. Syntactic structures can constrain the possible contexts a word form may be used in, thus there is an implicit relation between the semantics of a lexical unit and its associated polysemy. Motivated by these recent linguistic findings, we operationalize the polysemy of a word form as being influenced by both - its semantic variability and its importance in the syntactic network.

The level of polysemy a word possesses is highly subjective and varies widely across annotators (Artstein and Poesio, 2008). To aid annotators in creating, validating and qualitatively analysing sense inventories, having an estimate of the ambiguity a word possesses could be very helpful. This measure then acts as a proxy to how many (or how few) senses a word in a certain language possesses. A quantification of polysemy is also helpful in Information Retrieval systems as they can be used to rank more relevant results (Krovetz, 1997). Polysemic knowledge can also help improve cross-lingual alignment of embedding spaces and cross-lingual transfer (Garí Soler and Apidianaki, 2021).

While recent contextual embedding models like BERT, XLM and RoBERTa have been shown to possess the ability of distinguishing between different senses of a word (Garí Soler and Apidianaki, 2021), less attention has been paid towards quantifying the level of polysemy that a word represents - a measure which is continuous and can be compared across lexica. Attempts at quantifying polysemy either rely on large amounts of data (Pimentel et al., 2020) and/or on carefully tuned hyperparameters and embedding distortion due to dimensionality reduction of the contextual space of language models (Xypolopoulos et al., 2021).

We operationalize polysemy of a word form as a quantity influenced by its contextual semantic neighbors and its syntactic role in a syntactic network. In particular, we construct a contextual nearest-neighbor graph of lexical units using a pretrained language model like BERT (Devlin et al., 2019). We leverage the discrete Ricci curvature (Ni et al., 2015) measure defined on graph edges as an indicator of ambiguity of a word form. The Ricci curvature can be used to determine edge roles like bridge, cliques, etc. in a graph as illustrated in Figure 1. Additionally, we construct a syntactic network for the (ambiguous) word form based on the dependency trees of the randomly sampled contexts in which the word has occurred. This network acts as another linguistic signal guiding the polysemy measure. We rely on the ability of pretrained language models to distinguish between word senses (Garí Soler and Apidianaki, 2021) and the power of graph entropy methods to identify syntactic importance of word forms in the syntactic network. We propose a syntax-aware and fully unsupervised approach leveraging the discrete Ollivier-Ricci curvature of a graph to quantify the polysemy of a word. Our contributions can be summarized as follows:

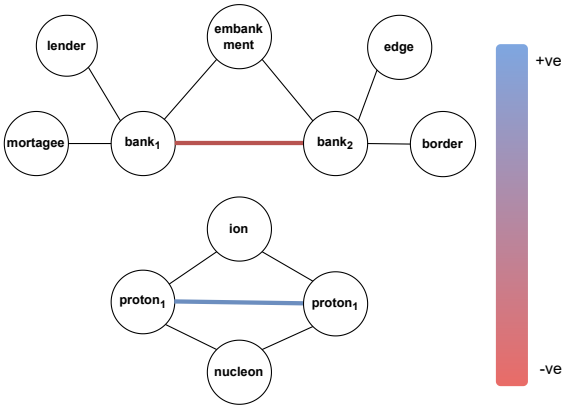

Figure 1: An illustrative example of Ricci curvature. The red edge (more negative) acts as a bridge connecting two distinct neighborhoods (distinct senses of the word *bank*) while the blue edge (more positive) is an edge within the same neighborhood (sense cluster of the monosemous word *proton*).

- We propose an approach which is fully unsupervised and parameter-free. To the best of our knowledge, we are the first to introduce a polysemy quantification method that leverages syntactic signals in the form of dependency trees.
- We show the effectiveness of network based approaches in linguistic research investigating polysemy and validate previous findings on the relationship between syntax and polysemy.
- We test our method on a set of typologically

diverse languages - English, French and Spanish and find significant correlations.

The rest of the paper is organized as follows. Section §2 discusses the background on ambiguity in linguistics and related topics. Mathematical preliminaries and the proposed method are detailed in Section §3 and Section §4 respectively. Section §5 discusses the implementation, data, experimental setup and the results. Limitations and future work are discussed in Section §6.

2 Background & Related Work

Lexical Ambiguity Ambiguity of language has been addressed as early as in the writings of Aristotle but relatively recent linguistic research in the form of Zipf's Principle of Least Effort (Bain, 1950) heralded a new understanding of human cognition and language systems positing the tradeoff between efficiency and brevity in communication systems (Piantadosi et al., 2012). Polysemy is a natural outcome of lexical semantic change (Bréal, 1904) by virtue of words gaining new meanings over time.

Recent works in computational linguistics for ambiguity mostly deal with word sense disambiguation (Pasini et al., 2021; Wiedemann et al., 2019), word-in-context tasks (Pilehvar and Camacho-Collados, 2019) and analyzing polysemy in language models like BERT (Garí Soler and Apidianaki, 2021). While some previous works (Erk and McCarthy, 2009; Friedrich et al., 2012) acknowledge polysemy even in particular instances, relatively less attention has been paid towards quantifying polysemy using current NLP tools. (Pimentel et al., 2020) measure ambiguity in language from an information-theoretic lens but their approach requires a large number of sentences to give a good upper bound on ambiguity estimates. (Xypolopoulos et al., 2021) leveraged contextual language models like BERT to estimate polysemy but they rely on dimensionality reduction and sensitive hyperparameters.

Works like (Reif et al., 2019; Haber and Poesio, 2021) have explored the geometry of BERT embeddings and their relation to polysemy levels thus highlighting the importance of neural embeddings in the quantification of polysemy levels of lexicons.

Lexical Substitution Lexical substitution is the task of finding relevant contextual replacements of a word given its context. To generate good quality contextual replacements, previous works have relied heavily on distributional semantic models like word2vec (Mikolov et al., 2013) and specialized language models like context2vec (Melamud et al., 2016). In all models, the generated substitutes are ranked based on some relation with the target word to be replaced. Recent advances in language models like the Transformer-based BERT (Devlin et al., 2019) and XLNet (Yang et al., 2020) rely on the bidirectional context and the special [MASK] token based training to generate contextual substitutes. (Zhou et al., 2019) showed that BERT performs poorly on lexical substitution and proposed a dropout based approach which is even more computationally expensive due to the large number of forward passes required. Supervised approaches (Lacerra et al., 2021) often rely on manually curated databases and sense inventories like WordNet, Wikipedia or BabelNet. (Arefyev et al., 2020) is a recent neural lexical substitution method which injects information about the target word in the form of probability distribution of possible word subtitutes based on word frequencies.

Graphs and NLP Traditional works in linguistics have used language networks and graphs for analyzing morphological complexity (Inglese and Brigada Villa, 2021), ambiguity (Čech et al., 2017) and phonetics (Yamshchikov et al., 2020). Recent advances in Graph Neural Networks (GNNs) has opened new avenues to apply network based approaches to language problems. While language networks have been analysed before, GNNs provide an alternative to traditional methods with more natural inductive biases for syntactic models to work with. The combination of graphs and language models has proved to be effective in incorporating semantics and syntax in language problems (Marcheggiani and Titov, 2020; Ahmad et al., 2021; Xu et al., 2021).

3 Preliminaries

3.1 Notations

Given a set of vertices $\mathcal{V}$ and set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, an undirected graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. For each node $v \in \mathcal{V}$, $\mathcal{N}(v)$ denotes the set of its 1-hop neighbors and $k_v = |\mathcal{N}(v)|$ denotes its degree.

3.2 Ricci Curvature

Traditionally, curvature is the geometric characteristic that measures how flat or curved an object is. The discrete Ollivier Ricci curvature (Ni et al., 2015) is the coarse graph generalization of curva-

ture measures usually defined on smooth surfaces or manifolds. For $u, v \in \mathcal{V}$, m_u and m_v are probability measures of total mass 1 each centered at u and v respectively. The Wasserstein (Earth Mover) distance $W(m_u, m_v)$ finds the optimal transportation plan ξ between probability distributions m_u and m_v.

$$W(m_u, m_v) = \inf_{\xi} \int \int d(u, v) d\xi(u, v) \quad (1)$$

It gives a metric to measure the minimum amount of work required to transform one probability distribution into another. The Ricci curvature thus becomes

$$\kappa_{uv} = 1 - \frac{W(m_u, m_v)}{d(u, v)} \quad (2)$$

where $d(u, v)$ is the number of edges in the shortest path between u and v.

Based on (Lin et al., 2011), we define the probability measure on node $v \in \mathcal{V}$ with $\alpha \in [0, 1]$ as:

$$m_v^{\alpha}(v_i) = \begin{cases} \alpha & \text{if } v_i = v \\ (1\text{-}\alpha)/k_v & \text{if } v_i \in \mathcal{N}(v) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

We use $\alpha = 0.5$ based on previous literature since it assigns equal weights to the node and its neighbors. The edge Ricci curvature acts as an indicator of the importance and structural role of an edge in a graph. It is representative of the intrinsic geometry and local topology of the edges in a graph. This property of the discrete Ricci curvature has been used to analyze the geometry of the Internet topology (Ni et al., 2015), Graph Neural Networks (Luo et al., 2021) and community detection (Sia et al., 2019).

4 Proposed Approach

In this section, we introduce our proposed approach to quantify polysemy as illustrated in Figure 2.

4.1 Semantic Module

Given a word w and its list of sentences (contexts where it occurs), $S = \{s_1, s_2, \cdots, s_k\}$, we consider each instance of the word w in its corresponding sentence s_i as a separate lemma w_i. For example, if we have two contexts for the word *bank*: 1) I went to the *bank*$_1$ to deposit money, and 2) Flowers grow along the river *bank*$_2$, we consider *bank*$_1$ and *bank*$_2$ as two separate lemmas during the graph construction.

We add each instance w_i of the word as a node to a graph $\mathcal{G}$ and pass each sentence $s_i \in S$ through a lexical substitution system (Arefyev et al., 2020) to retrieve the top contextual neighbors C_k of the word w_i, adding an edge between the nearest neighbor word $c_k^i \in C_k$ and the lemma w_i. The lexical substitution model gives the most appropriate contextual replacement of a word in the input sentence, thus we can derive the semantic neighbors of a word given its context which renders the construction of the graph $\mathcal{G}$ possible.

We now efficiently compute the Ricci curvature on each edge of the graph $\mathcal{G}$ based on the linear programming method introduced by (Ni et al., 2015) and Equations 2 and 3:

$$\begin{aligned} \min \quad & \textstyle\sum_{y \in V} \sum_{x \in V} d(x, y) \rho_{xy} m_u^{\alpha}(x), \\ \text{s.t.} \quad & 0 \leq \rho_{xy} \leq 1 \quad \forall x, y \in V, \\ & \textstyle\sum_{y \in V} \rho_{xy} = 1 \quad \forall x \in V, \\ & \textstyle\sum_{x \in V} \rho_{xy} m_u^{\alpha}(x) = m_v^{\alpha}(y) \quad \forall y \in V, \end{aligned} \quad (4)$$

where ρ is the transportation plan matrix.

For the graph $\mathcal{G}$, we now have the edge feature matrix $E \in \mathbb{R}^{\mathcal{E} \times 1}$. Based on the intuition that negative edges act as bridge across clusters, we hypothesize that negatively curved edges connect distinct senses of the same word w. We derive the negative edges normalized by total edges in the graph as:

$$\mathcal{P}_1 = \frac{|E^-|}{|E|} \quad (5)$$

where $|E^-|$ is the number of negative edges in the graph. This formulation describes the variation of the curved edges in the graph. While we describe here a ratio-based definition of $\mathcal{P}_1$, it can also be operationalised as the variation of edge weights in the graph, with similar results.

4.2 Syntactic Module

For the given word w and its list of contexts, we derive the syntactic dependency trees of each sentence $s_i \in S$. Note that, here we do not make any distinctions between the instances of the word w unlike in the case of the Semantic Module. The obtained dependency trees are converted to their corresponding adjacency matrix A with $A_{ij} = 1$ if there is a dependency relation between tokens i and j. Each adjacency matrix corresponding to each sentence s_i can be converted to an unweighted, undirected graph D_i.

We then construct a single, global syntactic graph $\mathcal{D} = \{D_1 \bigcup D_2 \cdots \bigcup D_k\}$ where $\bigcup$ is

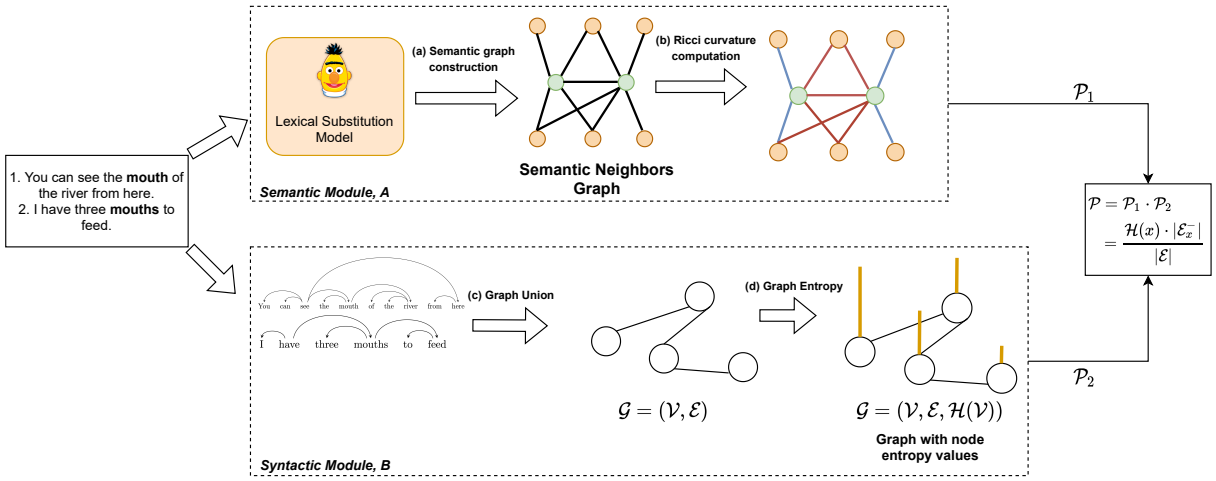

Figure 2: Proposed Approach for Polysemy Quantification. The set of contexts of the polysemous word is passed through: the Semantic Module (A) and the Syntactic Module (B). In the semantic module, (a) a contextual semantic graph is constructed with the help of a lexical substitution model and (b) Ricci curvature is computed on the graph edges. In the syntactic module, (c) dependency trees of the input sentences are constructed and combined in a global syntactic network using the Graph Union operator, and (d) the graph entropy is computed on the syntactic network. Both semantic (A) and syntactic (B) modules are then combined to derive the final measure of polysemy.

the graph union operator, i.e., for two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, $G_1 \bigcup G_2 = (V_1 \bigcup V_2, E_1 \bigcup E_2)$. The global syntactic graph $\mathcal{D}$ contains tokens as nodes and edges as syntactic dependencies between the tokens. It thus represents the syntactic relations between input word w and the tokens of contexts. Based on ideas proposed by previous work (Čech et al., 2017) that syntactic relations influence polysemy of words across languages, we utilize the relations encoded in this graph as a signal to our polysemy measure. Inspired by recent advances in graph signal processing (Wijesinghe et al., 2021; Nouranizadeh et al., 2021; Luo et al., 2021), we compute the node entropy of the word w to quantify the importance of a node as a function of its structure.

The adjacency matrix A of the global graph $\mathcal{D} = (V, E)$ contains first order links of the graph. We define $A^2 = A^T A$ to study second order links. D represents the degree vector of the graph. We define D_r as the normalized degree vector which contains information about first and second order links.

$$D_r = D^T A_r^2 \tag{6}$$

Here A_r^2 is the normalized second order adjacency matrix defined as,

$$A_r^2[i,j] = \frac{A^2[i,j]}{\sum_j A^2[i,j]} \tag{7}$$

Here, $A^2[i,j]$ is the i-th row and $j-th$ column of the second order adjacency matrix.

Following principles of information theory, the entropy of a node, x is thus defined as

$$\begin{aligned} \mathcal{P}_2 = H(x) &= -P_x \log P_x \\ &= -\frac{D_r[x]}{\sum_x D_r[x]} \log \frac{D_r[x]}{\sum_x D_r[x]} \end{aligned} \tag{8}$$

4.3 Polysemy Quantification

To derive the quantification for polysemy, we combine Equations 5 and 8 as:

$$\begin{aligned} \mathcal{P} &= \mathcal{P}_1 \cdot \mathcal{P}_2 \\ \mathcal{P} &= \frac{H(x) \cdot |E^-|}{|E|} \end{aligned} \tag{9}$$

We thus derive the final measure of polysemy as desrcibed in Equation 9. This operationalization of polysemy in a graph-based measure incorporates syntactic signals as well as semantic structural variation.

5 Experiments

In this section, we first describe the data used in the current study (§5.1) followed by a description of the flow of the proposed approach (§5.2). Next we describe the evaluation metrics used (§5.3) and the implementation details of the current study (§5.4). Finally, we discuss the results of the proposed approach (§5.5) and perform an ablation study of investigating the individual contribution of semantics and syntax towards polysemy (§5.6).

5.1 Data

We utilize the data introduced by (Garí Soler and Apidianaki, 2021). Sentences were sampled from SemCor 3.0 (Miller et al., 1993) dataset controlling for sense distributions in polysemous words that occur at least ten times in the corpus. For each polysemous word, we have 2 sets of sentences:

- **Random senses (poly-rand)**: Randomly sampling 10 sentences which captures the natural distribution of the senses of a word.
- **Balanced senses (poly-bal)**: 10 sentences of the word containing distinct senses. This is a controlled setting where the variation in the senses of the word is maximized.

The original English dataset is composed of 836 polysemous words, and their corresponding 8,195 unique sentences. For French and Spanish, the sentences are taken from the Eurosense corpus (Delli Bovi et al., 2017) which contains texts from Europarl automatically annotated with BabelNet word senses (Navigli and Ponzetto, 2012). In the multilingual corpus, we have 418 polysemous words.

We use the Frequency and Random baselines as described by (Xypolopoulos et al., 2021). In the frequency baseline, words are ranked in decreasing order of their frequency in the Wikipedia dump. The random baseline assigns scores by sampling from a Log Normal distribution.

5.2 Setup

We pass each sentence in the sentence pool (poly-bal or poly-rand) through the semantic module (§4.1) to get a contextual nearest neighbor graph and compute $\mathcal{P}_1$ (Equation 5) via the Ricci curvature. Parallel to this, the sentences are also passed through the syntactic module (§4.2) to build a global syntactic network to compute $\mathcal{P}_2$ (Equation 8). Finally, based on Equation 9, we compute the polysemy score for the input word.

5.3 Evaluation

Following previous literature in polysemy quantification (Xypolopoulos et al., 2021), we utilised Spearman correlation as our evaluation metric. We also perform significance tests of the correlation across all languages tested.

5.4 Implementation Details

We use the Stanford Stanza library (Qi et al., 2020) to build the dependency trees of sentences. We use the author's implementation of LexSubGen (Arefyev et al., 2020) as the lexical substitution module in our framework. To compute the Ricci curvature on graphs, we used the implementation based on (Ni et al., 2015).[2] All other code is written in PyTorch and uses Huggingface Transformers library (Wolf et al., 2020).

We use language-specific models for each of the language tested in our study. For English we use the state-of-the-art Lexical Substitution system described by (Arefyev et al., 2020). For languages other than English, we rely on the Masked Language Model prediction of the model which has been shown to be effective for lexical substitution by (Qiang et al., 2021). We use *bert-base-uncased* (Devlin et al., 2019) for English, *flaubert-base-uncased* (Le et al., 2020) for French and *bert-base-spanish-wwm-uncased* (Cañete et al., 2020). We compare our results with the model based on dimensionality reduction and multiresolution grids on the reported hyperparameters proposed by (Xypolopoulos et al., 2021).

5.5 Results

In this section, we discuss the results of the proposed quantification measure. We assume the number of senses of a word in the WordNet is a good representative of the ambiguity it possesses (Pimentel et al., 2020) and calculate its correlation with our proposed metric. Prior work like Xypolopoulos et al. (2021) have used WordNet as ground truth and empirically demonstrated that WordNet, WordNet-reduced and domain-specific WordNet all produce highly similar polysemy rankings despite the different sense granularities they have. Hence we report our results on the classic WordNet data. Henceforth, we refer to the approach proposed by (Xypolopoulos et al., 2021) as D2L8.

In Table 1, we observe that our measure shows higher significant correlations with the WordNet rankings on English data. For poly-rand setting, where the natural sense distribution of a word is captured, we observe an increment of 0.3 points in the correlation as compared to the D2L8 baseline which is based on the notion of multiresolution

[2] https://github.com/saibalmars/GraphRicciCurvature

Method	poly-bal	poly-rand
Random	0.11	0.15
Frequency	0.18	0.20
(Garí Soler and Apidianaki, 2021)	0.29	0.32
D2L8	0.30	0.27
Ours	**0.62**	**0.60**

Table 1: Spearman correlation of WordNet senses and polysemy scores on English data. Our approach improves the correlation by 0.3 points over D2L8. Numbers in bold are statistically significant ($p < 0.05$)

grids where volume is approximated hierarchical discretization of the embedding space (Nikolentzos et al., 2017). The poly-bal data is a controlled setting where the number of contexts is balanced. Although the baseline was described to work on randomly sampled sentences in English, we apply it to the controlled setting where it achieves a much better correlation of 0.3 and comparable to ours.

	French		Spanish	
	D2L8	Ours	D2L8	Ours
poly-bal	0.48	0.45	0.48	0.62
poly-rand	0.19	0.43	0.14	0.20

Table 2: Spearman correlation of the proposed polysemy quantification with WordNet number of senses across different languages.

We apply our measure in a cross-lingual setting to measure polysemy across 2 diverse languages other than English - French and Spanish. We also extend the baseline D2L8 to our cross-lingual setting.[3] Table 2 reports the Spearman correlations of the number of senses of a word in the Multilingual WordNet (Bond and Paik, 2012) of the language with our proposed quantification. We observe significant correlations across all languages and all settings (poly-bal and poly-rand). The poly-bal data setting shows consistently strong correlations as compared to poly-rand setting which is quite intuitive due to the carefully controlled sense distribution in poly-bal sentences. We note here that since we only take 10 sentences in each context pool (poly-bal and poly-rand), it is a highly constrainted setting as compared to previous works (Xypolopoulos et al., 2021; Pimentel et al., 2020) which randomly sampled greater than 10,000 sentences for each word. Our motivation behind taking this constrained approach is to enable our method to perform even for low-resource languages.

[3] https://github.com/ksipos/polysemy-assessment

5.6 Ablation study

We perform an ablation study in order to investigate the individual contribution of Semantic and Syntactic Module. In Table 3, we report the Spearman correlations of polysemy measure taken from each module with the English WordNet rankings.

	Syntax Module	Semantic Module
poly-bal	0.28	0.33
poly-rand	0.48	0.46

Table 3: Spearman correlation of individual measures from syntax and semantic modules with English WordNet ground truth rankings.

We observe that both semantic and syntactic module are positively correlated with the number of senses a word possesses. This result validates previous findings linking syntax and polysemy (Čech et al., 2017). These results suggest that studies in ambiguity should investigate syntax along with semantics of an utterance.

5.7 Error Analysis

Since we rely on a lexical substitution module (Arefyev et al., 2020), the errors in this model might propagate into the final score. For example, in some cases, the substitution model fails to generate enough number of word substitutions given the context, thus resulting in a sparse graph where Ricci curvature might not be a good metric to compute polysemy.

In some cases, the model also generates variations of the same semantic word, *home* and *homes*, which can further reduce the important signals required for the model to compute a good polysemy score.

6 Discussion

Since our method aims to quantify the tendency of a word to have more meanings, words assigned

higher values are assumed to be more polysemous. While this operationalisation does not explicitly allow for the discovery of new polysemy relations, we observe, for example, that "accord" (6 ground truth senses) is assigned a higher polysemy score relative to a word like "maximum" (4 ground truth sense). This case is interesting since WordNet provides very similar ratings for both while our method accentuates the difference between the two, intuitively, giving "accord" much higher score.

Conclusion In this study, based on previous linguistic evidence, we posit that including syntactic information in the form of dependency structural knowledge can help in the quantification of lexical ambiguity or polysemy of a wordform. To investigate this, we propose a simple operationalization of polysemy based on the Ricci curvature of the contextual nearest neighbors graph of a word and the entropy of its combined syntactic network.

We show that our proposed measure shows high correlations with number of word senses in WordNet across multiple languages. Our approach is fully unsupervised, simple and grounded in previously established linguistic theories. We hope that similar graph-based approaches can help in creation and validation of sense inventories across languages.

Limitations Our work acts as a proxy for the ambiguity of a word form and the scores are continuous but it does not quantify the discrete counts of the senses of a word. We rely on the availability of good quality language-specific language models which can be used as the lexical substitution model in the Semantic Module. Any errors in the language model may propagate into our score.

We tested our framework on sentences sampled from the SemCor 3.0 dataset which is a good resource for sense analysis in NLP but is naturally limited to sentences in formal English. A lack of diversely sourced corpora for a study in polysemy may limit the generalizability of a quantification measure to other domains.

Future Work We leave the utility of polysemy quantification to improve extrinsic tasks Word Sense Disambiguation or Word In Context for future work. We hope that works in polysemy quantification also lead to interesting linguistic analyses about the nature of ambiguity in natural languages and the relationship between syntactic information like Part-Of-Speech Tags and polysemy scores of word units.

Ethical Considerations

This study investigates polysemy and its quantification with the help of graph algorithms and language models which can be useful to many disciplines in the NLP and linguistics community. We are aware of the fact that any biases of the language model may creep into the proposed approach.

References

Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554.

Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. 2020. Always keep your target in mind: Studying semantics and improving performance of neural lexical substitution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1242–1255.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Read Bain. 1950. Human behavior and the principle of least effort: An introduction to human ecology. by george kingsley zipf. cambridge, mass.: Addison-wesley press, inc., 1949. 573 pp.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71.

Michel Bréal. 1904. *Essai de sémantique (science des significations).* Hachette.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Radek Čech, Ján Mačutek, Zdeněk Žabokrtskỳ, and Aleš Horák. 2017. Polysemy and synonymy in syntactic dependency networks. *Digital Scholarship in the Humanities*, 32(1):36–49.

Claudio Delli Bovi, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. 2017. EuroSense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–600.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449, Singapore. Association for Computational Linguistics.

Annemarie Friedrich, Nikos Engonopoulos, Stefan Thater, and Manfred Pinkal. 2012. A comparison of knowledge-based algorithms for graded word sense assignment. In *Proceedings of COLING 2012: Posters*, pages 329–338, Mumbai, India. The COLING 2012 Organizing Committee.

Aina Garí Soler and Marianna Apidianaki. 2021. Let's play mono-poly: Bert can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Janosch Haber and Massimo Poesio. 2021. Patterns of polysemy and homonymy in contextualised language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2663–2676, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaochen Hou, Peng Qi, Guangtao Wang, Rex Ying, Jing Huang, Xiaodong He, and Bowen Zhou. 2021. Graph ensemble learning over multiple dependency trees for aspect-level sentiment classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2884–2894.

Guglielmo Inglese and Luca Brigada Villa. 2021. Inferring morphological complexity from syntactic dependency networks: A test. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 10–22.

Robert Krovetz. 1997. Homonymy and polysemy in information retrieval. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 72–79.

Caterina Lacerra, Rocco Tripodi, and Roberto Navigli. 2021. GeneSis: A Generative Approach to Substitutes in Context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10810–10823.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490.

Yong Lin, Linyuan Lu, and Shing-Tung Yau. 2011. Ricci curvature of graphs. *Tohoku Mathematical Journal, Second Series*, 63(4):605–627.

Gongxu Luo, Jianxin Li, Jianlin Su, Hao Peng, Carl Yang, Lichao Sun, Philip S Yu, and Lifang He. 2021. Graph entropy guided node embedding dimension selection for graph neural networks. *arXiv preprint arXiv:2105.03178*.

Diego Marcheggiani and Ivan Titov. 2020. Graph convolutions over constituent trees for syntax-aware semantic role labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3915–3928.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Chien-Chun Ni, Yu-Yao Lin, Jie Gao, Xianfeng David Gu, and Emil Saucan. 2015. Ricci curvature of the internet topology. In *2015 IEEE conference on computer communications (INFOCOM)*, pages 2758–2766. IEEE.

Giannis Nikolentzos, Polykarpos Meladianos, and Michalis Vazirgiannis. 2017. Matching node embeddings for graph similarity. In *Thirty-first AAAI conference on artificial intelligence*.

Amirhossein Nouranizadeh, Mohammadjavad Matinkia, Mohammad Rahmati, and Reza Safabakhsh. 2021. Maximum entropy weighted independent set pooling for graph neural networks. *arXiv preprint arXiv:2107.01410*.

Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273.

Tiago Pimentel, Rowan Hall Maudslay, Damian Blasi, and Ryan Cotterell. 2020. Speakers fill lexical semantic gaps with context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4004–4015.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, Yang Shi, and Xindong Wu. 2021. Lsbert: Lexical simplification based on bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3064–3076.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

Jayson Sia, Edmond Jonckheere, and Paul Bogdan. 2019. Ollivier-ricci curvature-based method to community detection in complex networks. *Scientific reports*, 9(1):1–12.

Agustín Vicente and Ingrid L Falkum. 2017. Polysemy. In *Oxford Research Encyclopedia of Linguistics*.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.

Asiri Wijesinghe, Qing Wang, and Stephen Gould. 2021. A regularized wasserstein framework for graph kernels. *arXiv preprint arXiv:2110.02554*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjun Zhong, Xiaojun Quan, Daxin Jiang, and Nan Duan. 2021. Syntax-enhanced pre-trained model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5412–5422.

Christos Xypolopoulos, Antoine Tixier, and Michalis Vazirgiannis. 2021. Unsupervised word polysemy quantification with multiresolution grids of contextual embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3391–3401.

Ivan P Yamshchikov, Cyrille Merleau Nono Saha, Igor Samenko, and Jürgen Jost. 2020. It means more if it sounds good: Yet another hypothesis concerning the evolution of polysemous words. *arXiv preprint arXiv:2003.05758*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.

David Yenicelik, Florian Schmidt, and Yannic Kilcher. 2020. How does BERT capture semantics? a closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Junru Zhou, Zhuosheng Zhang, Hai Zhao, and Shuailiang Zhang. 2020. LIMIT-BERT : Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373.