MEASURING AND IMPROVING INTERPRETABILITY OF WORD EMBEDDINGS USING LEXICAL RESOURCES

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

ELECTRICAL AND ELECTRONICS ENGINEERING

By

Lütfi Kerem Şenel

August 2019

Measuring and Improving Interpretability of Word Embeddings Using Lexical Resources
By Lütfi Kerem Şenel
August 2019

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Tolga Çukur (Advisor)

Aykut Koç (Co-Advisor)

Varol Akman

Aykut Erdem

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

MEASURING AND IMPROVING INTERPRETABILITY OF WORD EMBEDDINGS USING LEXICAL RESOURCES

Lütfi Kerem Şenel
M.S. in Electrical and Electronics Engineering
Advisor: Tolga Çukur
Co-Advisor: Aykut Koç
August 2019

As an ubiquitous method in natural language processing, word embeddings are extensively employed to map semantic properties of words into a dense vector representations. They have become increasingly popular due to their state-of-the-art performances in many natural language processing (NLP) tasks. Word embeddings are substantially successful in capturing semantic relations among words, so a meaningful semantic structure must be present in the respective vector spaces. However, in many cases, this semantic structure is broadly and heterogeneously distributed across the embedding dimensions. In other words, vectors corresponding to the words are only meaningful relative to each other. Neither the vector nor its dimensions have any absolute meaning, making interpretation of dimensions a big challenge. We propose a statistical method to uncover the underlying latent semantic structure in the dense word embeddings. To perform our analysis, we introduce a new dataset (SEMCAT) that contains more than 6,500 words semantically grouped under 110 categories. We further propose a method to quantify the interpretability of the word embeddings that is a practical alternative to the classical word intrusion test that requires human intervention. Moreover, in order to improve the interpretability of word embeddings while leaving the original semantic learning mechanism mostly unaffected, we introduce an additive modification to the objective function of the embedding learning algorithm, GloVe, that promotes the vectors of words that are semantically related to a predefined concept to take larger values along a specified dimension. We use Roget's Thesaurus to extract concept groups and align the words in these groups with embedding dimensions using modified objective function. By performing detailed evaluations, we show that proposed method improves interpretability drastically while preserving the semantic structure. We also demonstrate that imparting method

with suitable concept groups can be used to significantly improve performance on benchmark tests and to measure and reduce gender bias present in the word embeddings.

Keywords: word embeddings, interpretability, semantics.

ÖZET

SÖZCÜKSEL KAYNAKLAR KULLANARAK KELİME TEMSİLLERİNİN YORUMLANABİLİRLİKLERİNİN ÖLÇÜLMESİ VE İYİLEŞTİRİLMESİ

Lütfi Kerem Şenel
Elektrik ve Elektronik Mühendisliği, Yüksek Lisans
Tez Danışmanı: Tolga Çukur
İkinci Tez Danışmanı: Aykut Koç
Ağustos 2019

Doğal dil işlemede (DDİ) yaygın bir yöntem olan kelime temsilleri, kelimelerin anlamsal özelliklerini yoğun vektörler kullanarak temsil etmek için sıklıkla kullanılmaktadır. Çok sayıda DDİ uygulamasında elde edilen en iyi performansları sağladıklarından popülerlikleri giderek artmıştır. Kelime temsilleri özellikle kelimeler arasındaki anlamsal ilişkileri yakalamakta başarılı olduklarından, bu temsil uzayları içlerinde anlamlı bir semantik yapı barındırmalıdırlar. Ancak genellikle bu anlamsal yapı uzayın boyutları arasında heterojen bir şekilde dağılmaktadır. Başka bir ifadeyle, kelimelere karşılık gelen vektörler sadece birbirlerine göre anlam taşırlar. Bir kelime vektörünün ve bu vektörün boyutlarının tek başına mutlak bir anlamı yoktur ve bu durum boyutların yorumlanmasını zorlaştırmaktadır. Bu tezde, yoğun kelime temsil uzaylarında altta yatan saklı anlamsal yapıyı ortaya çıkarmak için istatistiksel bir yöntem önerilmiştir. Buna ek olarak, kelime temsil uzaylarının yorumlanabilirlik düzeylerini sayısal olarak ölçmeye yarayan bir yöntem önerilmiştir. Önerilen yöntem, literatürde yorumlanabilirliği ölçmek için kullanılan ve insan değerlendirmesine gereksinim duyan kelime ihlal testine pratik bir alternatif olma potansiyeline sahiptir. Ayrıca, orijinal öğrenme mekanizmasını etkilemeden kelime temsillerinin yorumlanabilirliklerini arttırmak amacıyla, GloVe kelime temsil algoritmalasının amaç fonksiyonuna yeni bir terim eklenmiştir. Eklenen terim, önceden tanımlanan konular ile anlamsal olarak ilişkili olan kelimelerin vektörlerinin temsil uzayının belirli boyutlarında yüksek değerler almasını sağlamaktadır. Kavram gruplarını oluşturmak amacıyla Roget's Thesaurus kaynak olarak kullanılmıştır. Elde edilen kavram gruplarının içerisindeki kelimelerin vektörlerinin temsil uzayının belirli boyutlarında yüksek

değerler almaları sağlanmıştır. Önerilen yöntemin kelime temsil uzayının yorumlanabilirliğini, uzayın anlamsal yapısına zarar vermeden, önemli derecede arttırdığı yapılan ayrıntılı değerlendirme ve ölçümler ile gösterilmiştir. Ayrıca önerilen yöntemin uygun kavram grupları ile beraber kullanıldığında denektaşı sınamalarında önemli performans artışı sağladığı ve kelime temsillerinde bulunan cinsiyet önyargısını düşürdüğü gösterilmiştir.

Anahtar sözcükler: kelime temsilleri, yorumlanabilirlik, semantik.

Acknowledgement

I want to acknowledge the support of TÜBİTAK (The Scientific and Technological Research Council of Turkey) BİDEB 2210 graduate student fellowship.

I would like to express my thanks to my advisor Assoc. Prof. Dr. Tolga Çukur. He has been an excellent advisor throughout my master's studies and his expertise was invaluable in the writing of this dissertation.

I would like to thank to my co-advisor Assist. Prof. Dr. Aykut Koç who introduced me to natural language processing. His unending motivation and tenacity helped me greatly to give my full effort during my research.

I want to thank my colleagues at ASELSAN Research Center, especially Dr. Veysel Yücesoy, for their wonderful collaboration. They supported me greatly and were willing to help me whenever I needed.

I thank all of my friends, especially Furkan Güç and Çağlar Öksüz, for providing happy distraction to rest my mind outside of my research.

Finally, I would like to thank my parents for their moral support and wise counsel and also my wife, Gonca, for her understanding and patience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Language is one of the unique attributes of the human species. There are other species that can communicate in different and limited ways. However, their communication is limited to what is physically present around them. Only humans can use a creative and narrative language to communicate events that go beyond here and now. Until the development of language, knowledge was mainly transferred from one generation to the next through the genes slowly with the course of natural selection. However, with language, which is an exceptionally effective tool to transfer knowledge, information spread and accumulated rapidly. We use language for cooperation, coordination and planning as large groups. With the accumulating knowledge and ability to coordinate in large groups, humans became the dominant species on our world and established control over all other creatures and even nature to some extend [1].

For a long time, language existed in the sign language and verbal language (i.e. speech) form. Invention of writing allowed us to store information in a physical form outside of our forgetful minds that can stay unchanged over time and it had a great impact on human civilization. Throughout the history of writing, we have used many different materials to write on such as stone, metal, clay or wooden tablets, leather, papyrus, parchment and paper. However, with the advancing technology during 20^{th} century, we began to store data in electronic format and

computers began to emerge. In 1950, Alan Turing wrote a paper in which he proposed a criterion for machine intelligence what is now called the *Turing test* [2]. He stated that a machine can be called intelligent if it can imitate a human in a written conversation with a person sufficiently well so that the person cannot distinguish the program from a real human. Although there have been some work from earlier periods regarding structure of language and source of meaning [3], this test, along with the studies showing that brain is composed of neurons forming an electrical network [4], motivated the idea of *artificial intelligence (AI)* and *natural language processing (NLP)*,

Until 1980s several NLP systems have been developed with limited success, primarily for the machine translation task [5, 6], based on complex hand-written rules. Starting from the late 1980s, with the increasing computational power, *machine learning* algorithms have been introduced for NLP, shifting the research focus to statistical models that can make probabilistic decisions based on real valued input data. However, since language is composed of discrete units, these units must be represented by real valued vectors before they taken as input to such models.

Words are the smallest elements of a language with a practical meaning. Hence, they are commonly used as input units and are represented by vectors. The simplest approach to represent a word as a vector is one hot encoding. One-hot vectors are immensely long due to large vocabulary size which increases computation and memory requirements, and they do not provide any information about meaning of a word or semantic relations between words. This limits the performance of the overall model. Therefore, it is necessary to have models that map words to effective vectors. Researchers from diverse fields including linguistics [7], computer science [8] and statistics [9] have developed models that seek to capture "word meaning" so that these models can accomplish various NLP tasks such as parsing, word-sense disambiguation and machine translation. Most of the effort in this field is based on the distributional hypothesis [10] which claims that a word is characterized by the company it keeps [11]. Building on this idea, several vector space models such as the well known Latent Semantic Analysis (LSA) [12] and Latent Dirichlet Allocation (LDA) [13] that make use of word distribution statistics

have been proposed in distributional semantics. Although these methods have been commonly used in NLP, more recent techniques that generate dense, continuous valued vectors, called *embeddings*, have been receiving increasing interest in NLP research. Approaches that learn embeddings include neural network based predictive methods [8, 14, 15] and count-based matrix-factorization methods [16]. Word embeddings brought about significant performance improvements in many intrinsic NLP tasks such as analogy or semantic textual similarity tasks, as well as downstream NLP tasks such as part-of-speech (POS) tagging [17], parsing [18], named entity recognition [19], word sense disambiguation [20], sentiment analysis [21, 22] and cross-lingual studies [23] where they generally serve as elementary building blocks in the course of algorithm design.

Empirical utility of word embeddings as an unsupervised method for capturing the semantic and syntactic features of a certain word as it is used in a given lexical resource is well-established [24, 25, 26]. However, an understanding of what these features mean remains an open problem [27, 28] and as such word embeddings mostly remain a black box. It is desirable to be able to develop insight into this black box and to interpret what it means, while retaining the utility of word embeddings as semantically-rich intermediate representations. A systematic assessment of the semantic structure intrinsic to word embeddings would enable an improved understanding of how NLP algorithms work [29], would allow for comparisons among different embeddings in terms of interpretability, set a ground that would facilitate the design of new algorithms in a more deliberate way and potentially motivate new research directions.

Generally, the learned embeddings make sense only in relation to each other and their specific dimensions do not carry explicit information that can be interpreted. However, being directly able to interpret a word embedding would illuminate the semantic concepts implicitly represented along the various dimensions of the embedding, and reveal which information is covered by the embedding. Knowing what information is captured by which dimensions, unnecessary dimensions can be removed from the embedding for a specific task, reducing the computation and memory requirements. Moreover, interpretable word embeddings can be crucial for achieving interpretable deep learning models for natural language

processing. Most of the current deep learning models lack interpretability which is one of their most significant shortcomings.

In the literature, researchers tackled interpretability problem of the word embeddings using different approaches. Several researchers [30, 31, 32] proposed algorithms based on non-negative matrix factorization (NMF) applied to co-occurrence variant matrices. Other researchers suggested to obtain interpretable word vectors from existing uninterpretable word vectors by applying sparse coding [33, 34], by training a sparse auto-encoder to transform the embedding space [35], by rotating the original embeddings [36, 37].

Although the aforementioned approaches provide better interpretability that is measured using a particular method such as word intrusion test, usually the improved interpretability comes with a cost of performance in the benchmark tests such as word similarity or word analogy. One possible explanation for this performance degradation is that the proposed transformations from the original embedding space distort the underlying semantic structure constructed by the original embedding algorithm. Therefore, it can be claimed that a method that learns dense and interpretable word embeddings without inflicting any damage to the underlying semantic learning mechanism is the key for achieving both high performing and interpretable word embeddings.

This thesis is organized as follows: In Chapter 2, we propose a method to investigate the semantic structure of the word embeddings based on a category dataset we introduce, called *SEMCAT*, and validate our findings by various tests. In that chapter, we also propose a method to quantify the interpretability of word embeddings without requiring any human effort and we introduce a new Turkish category dataset, *ANKAT*, to evaluate interpretability of Turkish word embeddings. In Chapter 3, we propose a modification to the cost function of the popular embedding algorithm, *GloVe*, so that it makes use of an external lexical resource and the resulting embedding space is highly interpretable while preserving the semantic structure learned by the original algorithm. In Chapter 4, we show that by proper selection of the external resource, the proposed method can also be used to greatly improve performance of the resulting embedding on

intrinsic evaluations and to reduce the gender bias present in the embedding space. Finally, in Chapter 5 we summarize our contributions and discuss possible future studies.

Chapter 2

Semantic Structure and Interpretability

In this chapter, we aim to bring to light the semantic concepts implicitly represented by various dimensions of a word embedding. To explore these hidden semantic structures, we leverage the category theory [38] that defines a category as a grouping of concepts with similar properties. We use human-designed category labels to ensure that our results and interpretations closely reflect human judgements. Human interpretation can make use of any kind of semantic relation among words to form a semantic group (category). This does not only significantly increase the number of possible categories but also makes it difficult and subjective to define a category. Although several lexical databases such as WordNet [7] have a representation for relations among words, they do not provide categories as needed for this study. Since there is no gold standard for semantic word categories to the best of our knowledge, we introduce a new category dataset where more than 6.500 different words are grouped into 110 semantic categories. Then, we propose a method based on distribution statistics of category words within the embedding space in order to uncover the semantic structure of dense word vectors. We apply quantitative and qualitative tests to substantiate our method. Finally, we claim that the semantic decomposition of the embedding space can be used to quantify the interpretability of the word embeddings

without requiring any human effort unlike the word intrusion test [39].

This chapter is organized as follows: Following a discussion of related work in Section 2.1, we investigate semantic structure of word embeddings in Section 2.2. In that section, we introduce our dataset and present the methods we used to investigate word embeddings and to validate our findings. We also present the results for our experiments in this chapter. In Section 2.3, we propose a new method to quantify the interpretability of the word embeddings. We test our method on various word embeddings. In that section, we also introduce a more sophisticated version of the proposed method and a new category dataset for Turkish that is used to measure the interpretability of Turkish word embeddings. Finally, we conclude the chapter in Section 2.4 with a discussion of our findings.

2.1 Related Work

In the word embedding literature, the problem of interpretability has been approached via several routes. For learning sparse, interpretable word representations from co-occurrence variant matrices, [30] suggested algorithms based on non-negative matrix factorization (NMF) and the resulting representations are called non-negative sparse embeddings (NNSE). To address memory and scale issues of the algorithms in [30], [31] proposed an online method of learning interpretable word embeddings. In both studies, interpretability was evaluated using a word intrusion test introduced in [39]. The word intrusion test costly since it requires manual evaluations by human observers separately for each embedding dimension. As an alternative method to incorporate human judgement, [32] proposed joint non-negative sparse embedding (JNNSE), where the aim is to combine text-based similarity information among words with brain activity based similarity information to improve interpretability. Yet, this approach still requires labor-intensive collection of neuroimaging data from multiple subjects.

Instead of learning interpretable word representations directly from co-occurrence matrices, [33] and [34] proposed to use sparse coding techniques on

conventional dense word embeddings to obtain sparse, higher dimensional and more interpretable vector spaces. However, since the projection vectors that are used for the transformation are learned from the word embeddings in an unsupervised manner, they do not have labels describing the corresponding semantic categories. Moreover, these studies did not attempt to enlighten the dense word embedding dimensions, rather they learned new high dimensional sparse vectors that perform well on specific tests such as word similarity and polysemy detection. In [34], interpretability of the obtained vector space was evaluated using the word intrusion test. An alternative approach was proposed in [36], where interpretability was quantified by the degree of clustering around embedding dimensions and orthogonal transformations were examined to increase interpretability while preserving the performance of the embedding. Note, however, that it was shown in [36] that total interpretability of an embedding is constant under any orthogonal transformation and can only be redistributed across the dimensions. With a similar motivation to [36], [37] proposed rotation algorithms based on exploratory factor analysis (EFA) to preserve the expressive performance of the original word embeddings while improving their interpretability. In [37]. interpretability was calculated using a distance ratio (DR) metric that is effectively proportional to the metric used in [36]. Although interpretability evaluations used in [36] and [37] are free of human effort, they do not necessarily reflect human interpretations since they are directly calculated from the embeddings.

Taking a different perspective, a recent study [40] attempted to elucidate the semantic structure within NNSE space by using categorized words from the HyperLex dataset [41]. The interpretability levels of embedding dimensions were quantified based on the average values of word vectors within categories. However, HyperLex is based on a single type of semantic relation (hypernym) and average number of words representing a category is small (≈ 2) making it challenging to conduct a comprehensive analysis.

2.2 Semantic Structure Analysis

2.2.1 Methods

To address the limitations of the approaches discussed in Section 2.1, we introduce a new conceptual category dataset. Based on this dataset, we propose statistical methods to capture the hidden semantic concepts in word embeddings.

2.2.1.1 Dataset

Understanding the hidden semantic structure in dense word embeddings and providing insights on interpretation of their dimensions are the main objectives of this study. Since embeddings are formed via unsupervised learning on unannotated large corpora, some conceptual relationships that humans anticipate may be missing and some that humans do not anticipate may be present in the embedding space [42]. Thus, not all clusters obtained from a word embedding space will be interpretable. Therefore, using the clusters in the dense embedding space might not take us far towards interpretation. This observation is also rooted in the need for human judgement in evaluating interpretability.

To provide meaningful interpretations for embedding dimensions, we refer to the category theory [38] where concepts with similar semantic properties are grouped under a common category. As mentioned earlier, using clusters from the embedding space as categories may not reflect human expectations accurately. Hence, having a basis founded on human judgements is essential for evaluating interpretability. In that sense, semantic categories as dictated by humans can be considered a gold standard for categorization tasks since they directly reflect human expectations. Therefore, using supervised categories can enable a proper investigation of the word embedding dimensions. In addition, by comparing the human-categorized semantic concepts with the unsupervised word embeddings, one can acquire an understanding of what kind of concepts can or cannot be captured by the current state-of-the-art embedding algorithms.

Table 2.1: Summary Statistics of SEMCAT and HyperLex

	SEMCAT	HyperLex
Number of Categories	110	1399
Number of Unique Words	6559	1752
Average Word Count per Category	91	2
Standard Deviation of Word Counts	56	3

In the literature, the concept of category is commonly used to indicate super-subordinate (hyperonym-hyponym) relations, where words within a category are types or examples of that category. For instance, the furniture category includes words for items such as bed or table. The HyperLex category dataset [41]. which was used in [40] to investigate embedding dimensions, is constructed based on this type of relation that is also the most frequently encoded relation among sets of synonymous words in the WordNet database [7]. However, there are many other types of semantic relations such as meronymy (part-whole relations). antonymy. synonymy and cross-Part of Speech (POS) relations (i.e, lexical entailments). Although WordNet provides representations for a subset of these relations, there is no clear guideline for constructing unified categories based on multiple different types of relations. It remains unclear what should be considered as a category, how many categories there should be, how narrow or broad they might be, and which words they should contain. Furthermore, humans can group words by inference, based on various physical or numerical properties such as color, shape, material, size or speed, increasing the number of possible groups almost unboundedly. For instance, words that may not be related according to classical hypernym or synonym relations might still be grouped under a category due to shared physical properties: sun, lemon and honey are similar in terms of color; spaghetti, limousine and sky-scrapper are considered as long; snail, tractor and tortoise are slow.

In sum, diverse types of semantic relationships or properties can be leveraged by humans for semantic interpretation. Therefore, to investigate the semantic

structure of the word embedding space using categorized words, we need categories that represent a broad variety of distinct concepts and distinct types of relations. To the best of our knowledge, there is no comprehensive word category dataset that captures the many diverse types of relations mentioned above. What we have found the closest to the required dataset are the online categorized word-lists[1] that were constructed for educational purposes. There are a total of 168 categories on these word-lists. To build a word-category dataset suited for assessing the semantic structure in word embeddings, we took these word-lists as a foundation. We filtered out words that are not semantically related but share a common syntactic property such as their POS tagging (verbs. adverbs. adjectives etc.) or being compound words. Several categories containing proper words or word phrases such as "chinese new year" and "good luck symbols". which we consider too specific. are also removed from the dataset. The vocabulary is limited to the most frequent 50,000 words, where frequencies are calculated from Wikipedia (English), and words that are not contained in this vocabulary are removed from the dataset. We call the resulting semantically grouped word dataset "SEMCAT" (SEMantic CATegories[2]). Summary statistics of SEMCAT and HyperLex datasets are given in Table 2.1. Ten sample words from each of six representative SEMCAT categories are given in Table 2.2.

2.2.1.2 Semantic Decomposition

For semantic decomposition, we use GloVe [16] as the source algorithm for learning dense word vectors. The entire content of Wikipedia is utilized as the corpus. In the preprocessing step, all non-alphabetic characters (punctuation, digits, etc.) are removed from the corpus and all letters are converted to lowercase. Letters coming after apostrophes are taken as separate words (`she'll` becomes `she ll`). The resulting corpus is input to the GloVe algorithm. Window size is set to 15, vector length is chosen to be 300 and minimum occurrence count is set to 20 for the words in the corpus. Default values are used for the remaining parameters.

[1] www.enchantedlearning.com/wordlist/

[2] github.com/avaapm/SEMCATdataset2018

Table 2.2: Ten sample words from each of the six representative SEMCAT categories.

Science	**Sciences**	**Art**	**Car**	**Cooking**	**Geography**
atom	astronomy	abstract	auto	bake	africa
cell	botany	artist	car	barbecue	border
chemical	economics	brush	coupe	boil	capital
data	genetics	composition	hybrid	dough	city
element	linguistics	draw	jeep	grill	continent
evolution	neuroscience	masterpiece	limo	juice	earth
laboratory	psychology	photograph	runabout	marinate	east
microscope	sociology	perspective	rv	oil	gps
scientist	taxonomy	sketch	taxi	roast	river
theory	zoology	style	van	serve	sea

The word embedding matrix, $\mathcal{E}$, is obtained from GloVe after limiting vocabulary to the most frequent 50,000 words in the corpus (i.e, $\mathcal{E}$ is 50,000$\times$300). The GloVe algorithm is again used for the second time on the same corpus generating a second embedding space, $\mathcal{E}^2$, to examine the effects of different initializations of the word vectors prior to training.

To quantify the significance of word embedding dimensions for a given semantic category, one should first understand how a semantic concept can be captured by a dimension, and then find a suitable metric to measure it. [40] assumed that a dimension represents a semantic category if the average value of the category words for that dimension is above an empirical threshold, and therefore took that average value as the representational power of the dimension for the category. Although this approach may be convenient for NNSE, directly using the average values of category words is not suitable for well-known dense word embeddings due to several reasons. First, in dense embeddings it is possible to encode in both positive and negative directions of the dimensions, making a single threshold insufficient. In addition, different embedding dimensions may have different

statistical characteristics. For instance, average value of the words from the jobs category of SEMCAT is around 0.38 and 0.44 in the 221^{st} and 57^{th} dimensions of $\mathcal{E}$. respectively; and the average values across all vocabulary are around 0.37 and -0.05, respectively, for the two dimensions. Therefore, the average value of 0.38 for the jobs category may not represent any encoding in the 221^{st} dimension since it is very close to the average of any random set of words in that dimension. In contrast, an average of similar value 0.44 for the jobs category may be highly significant for the 57^{th} dimension. Note that focusing solely on average values might be insufficient to measure the encoding strength of a dimension for a semantic category. For instance, words from the car category have an average of -0.08 that is close to the average across all vocabulary. -0.04. for the 133^{th} embedding dimension. However, standard deviation of the words within the car category is 0.15 which is significantly lower than the standard deviation of all vocabulary (0.35) for this particular dimension. In other words, although the average of words from the car category is very close to the overall mean. category words are more tightly grouped compared to other vocabulary words in the 133^{th} embedding dimension, potentially implying significant encoding.

From a statistical perspective, the question of "How strong a particular concept is encoded in an embedding dimension?" can be interpreted as "How much information can be extracted from a word embedding dimension regarding a particular concept?". If the words representing a concept (i.e, words in a SEMCAT category) are sampled from the same distribution with all vocabulary words, then the answer would be zero since the category would be statistically equivalent to a random selection of words. For dimension i and category j. if $\mathcal{P}_{i,j}$ denotes the distribution from which words of that category are sampled and $\mathcal{Q}_{i,j}$ denotes the distribution from which all other vocabulary words are sampled, then the distance between distributions $\mathcal{P}_{i,j}$ and $\mathcal{Q}_{i,j}$ will be proportional to the information that can be extracted from dimension i regarding category j. Based on this argument, Bhattacharya distance [43] with normal distribution assumption is a suitable metric, which is given in (2.1), to quantify the level of encoding in the word embedding dimensions. Normality of the embedding dimensions is tested using one-sample Kolmogorov-Smirnov test (KS test, Bonferroni corrected for

multiple comparisons).

$$\mathcal{W}_B(i,j) = \frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{\sigma^2_{p_{i,j}}}{\sigma^2_{q_{i,j}}} + \frac{\sigma^2_{q_{i,j}}}{\sigma^2_{p_{i,j}}} + 2\right)\right) + \frac{1}{4}\left(\frac{\left(\mu_{p_{i,j}} - \mu_{q_{i,j}}\right)^2}{\sigma^2_{p_{i,j}} + \sigma^2_{q_{i,j}}}\right) \tag{2.1}$$

In (2.1), $\mathcal{W}_B$ is a 300×110 Bhattacharya distance matrix, which can also be considered as a category weight matrix; i is the dimension index ($i \in \{1.2.....300\}$) and j is the category index ($j \in \{1.2.....110\}$). $p_{i,j}$ is the vector of the i^{th} dimension of each word in j^{th} category and $q_{i,j}$ is the vector of the i^{th} dimension of all other vocabulary words ($p_{i,j}$ is of length n_j and $q_{i,j}$ is of length $(50000 - n_j)$ where n_j is the number of words in the j^{th} category), μ and σ are the mean and the standard deviation operations, respectively. Values in $\mathcal{W}_B$ can range from 0 (if $p_{i,j}$ and $q_{i,j}$ have the same means and variances) to ∞. In general, a better separation of category words from remaining vocabulary words in a dimension results in larger $\mathcal{W}_B$ elements for the corresponding dimension.

Based on SEMCAT categories, for the learned embedding matrices $\mathcal{E}$ and $\mathcal{E}^2$. the category weight matrices ($\mathcal{W}_B$ and $\mathcal{W}_B^2$) are calculated using Bhattacharya distance metric (2.1).

2.2.1.3 Mapping Word Vectors to SEMCAT

If the weights in $\mathcal{W}_B$ truly correspond to the categorical decomposition of the semantic concepts in the dense embedding space, then $\mathcal{W}_B$ can also be considered as a transformation matrix that can be used to map word embeddings to a semantic space where each dimension is a semantic category. However, it would be erroneous to directly multiply the word embeddings with category weights. The following steps should be performed in order to map word embeddings to a semantic space where dimensions are interpretable:

1. To make word embeddings compatible in scale with the category weights. word embedding dimensions are standardized ($\mathcal{E}_S$) such that each dimension

has zero mean and unit variance since category weights have been calculated based on the deviations from the general mean (the second term in (2.1)) and standard deviations (the first term in (2.1)).

2. Category weights are normalized across dimensions such that each category has a total weight of 1 ($\mathcal{W}_{NB}$). This is necessary since some columns of $\mathcal{W}_B$ dominate others in terms of representation strength (to be discussed in Section 2.2.2 in more detail). This inequality across semantic categories can cause an undesired bias towards categories with larger total weights in the new vector space. ℓ_1 normalization of the category weights across dimensions is performed to prevent bias.

3. Word embedding dimensions can encode semantic categories in both positive and negative directions ($\mu_{p_{i,j}} - \mu_{q_{i,j}}$ can be positive or negative) that contribute equally to the Bhattacharya distance. However, since encoding directions are important for the mapping of the word embeddings, $\mathcal{W}_{NB}$ is replaced with its signed version $\mathcal{W}_{NSB}$ (if $\mu_{p_{i,j}} - \mu_{q_{i,j}}$ is negative, then $\mathcal{W}_{NSB}(i,j) = -\mathcal{W}_{NB}(i,j)$. otherwise $\mathcal{W}_{NSB}(i,j) = \mathcal{W}_{NB}(i,j)$) where negative weights correspond to encoding in the negative direction.

Then, interpretable semantic vectors ($\mathcal{I}_{50000\times110}$) are obtained by multiplying $\mathcal{E}_S$ with $\mathcal{W}_{NSB}$.

One can reasonably suggest to alternatively use the centers of the vectors of the category words as the weights for the corresponding category as given in (2.2).

$$\mathcal{W}_C(i,j) = \mu_{p_{i,j}} \tag{2.2}$$

A second interpretable embedding space, $\mathcal{I}^*$, is then obtained by simply projecting the word vectors in $\mathcal{E}$ to the category centers. (2.3) and (2.4) show the calculation of $\mathcal{I}$ and $\mathcal{I}^*$, respectively. Figure 2.1 shows the procedure for generation of interpretable embedding spaces $\mathcal{I}$ and $\mathcal{I}^*$.

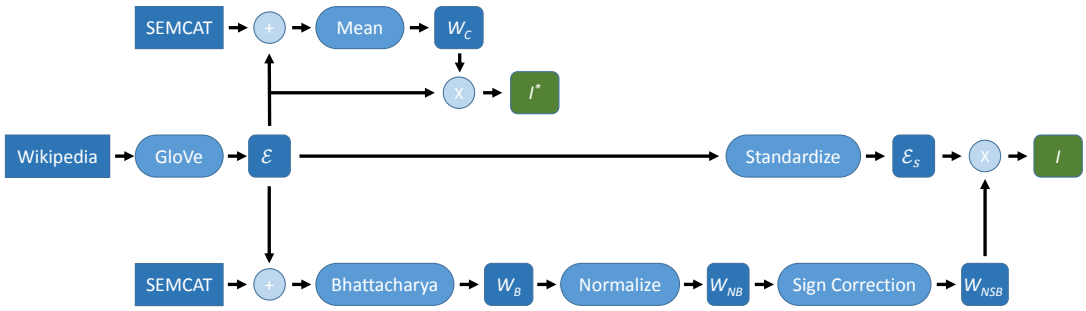

Figure 2.1: Flow chart for the generation of the interpretable embedding spaces $\mathcal{I}$ and $\mathcal{I}^*$. First, word vectors are obtained using the GloVe algorithm on Wikipedia corpus. To obtain $\mathcal{I}^*$, weight matrix $\mathcal{W}_C$ is generated by calculating the means of the words from each category for each embedding dimension and then $\mathcal{W}_C$ is multiplied by the embedding matrix (see Section 2.2.1.3). To obtain $\mathcal{I}$, weight matrix $\mathcal{W}_B$ is generated by calculating the Bhattacharya distance between category words and remaining vocabulary for each category and dimension. Then, $\mathcal{W}_B$ is normalized (see Section 2.2.1.3, item 2), sign corrected (see Section 2.2.1.3, item 3) and finally multiplied with standardized word embedding ($\mathcal{E}_s$. see Section 2.2.1.3, item 1)

$$\mathcal{I} = \mathcal{E}_S \mathcal{W}_{NSB} \tag{2.3}$$

$$\mathcal{I}^* = \mathcal{E} \mathcal{W}_C \tag{2.4}$$

2.2.1.4 Validation

$\mathcal{I}$ and $\mathcal{I}^*$ are further investigated via qualitative and quantitative approaches in order to confirm that $\mathcal{W}_B$ is a reasonable semantic decomposition of the dense word embedding dimensions, that $\mathcal{I}$ is indeed an interpretable semantic space and that our proposed method produces better representations for the categories than their center vectors.

If $\mathcal{W}_B$ and $\mathcal{W}_C$ represent the semantic distribution of the word embedding dimensions. then columns of $\mathcal{I}$ and $\mathcal{I}^*$ should correspond to semantic categories. Therefore, each word vector in $\mathcal{I}$ and $\mathcal{I}^*$ should represent the semantic decomposition of the respective word in terms of the SEMCAT categories. To test this

prediction, word vectors from the two semantic spaces ($\mathcal{I}$ and $\mathcal{I}^*$) are qualitatively investigated.

To compare $\mathcal{I}$ and $\mathcal{I}^*$, we also define a quantitative test that aims to measure how well the category weights represent the corresponding categories. Since weights are calculated directly from word vectors, it is natural to expect that words should have high values in dimensions that correspond to the categories they belong to. However, using words that are included in the categories for investigating the performance of the calculated weights is similar to using training accuracy to evaluate model performance in machine learning. Using validation accuracy is more adequate to see how well the model generalizes to new, unseen data which, in our case, correspond to words that do not belong to any category. During validation, we randomly select 60% of the words for training and use the remaining 40% for testing for each category. From the training words we obtain the weight matrix $\mathcal{W}_B$ using Bhattacharya distance and the weight matrix $\mathcal{W}_C$ using the category centers. We select the largest k weights ($k \in \{5, 7, 10, 15, 25, 50, 100, 200, 300\}$) for each category (i.e, the largest k elements for each column of $\mathcal{W}_B$ and $\mathcal{W}_C$) and replace the other weights with 0 to obtain sparse category weight matrices $\mathcal{W}_B^s$ and $\mathcal{W}_C^s$. Then, projecting dense word vectors onto the sparse weights from $\mathcal{W}_B^s$ and $\mathcal{W}_C^s$, we obtain interpretable semantic spaces $\mathcal{I}_k$ and $\mathcal{I}_k^*$. Afterwards, for each category, we calculate the percentages of the unseen test words that are among the top n, $3n$ and $5n$ words (excluding the training words) in their corresponding dimensions in the new spaces, where n is the number of test words that varies across categories. We calculate the final accuracy as the weighted average of the accuracies across the dimensions in the new spaces, where the weighting is proportional to the number of test words within the categories. We repeat the same procedure for 10 independent random selections of the training words.

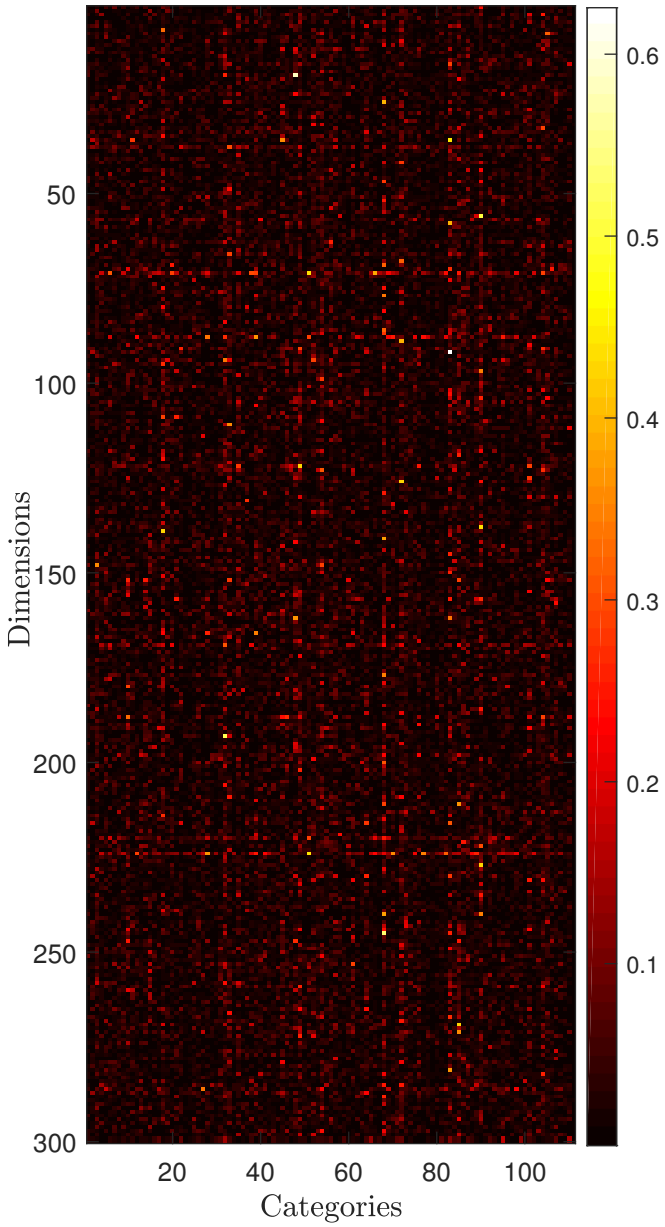

Figure 2.2: Semantic category weights ($\mathcal{W}_{B\ 300\times110}$) for 110 categories and 300 embedding dimensions obtained using Bhattacharya distance. Weights vary between 0 (represented by black) and 0.63 (represented by white). It can be noticed that some dimensions represent larger number of categories than others do and also some categories are represented strongly by more dimensions than others.

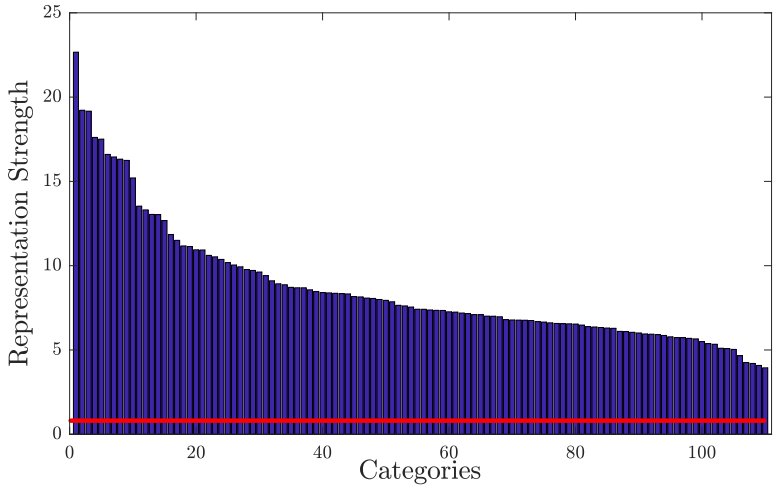

Figure 2.3: Total representation strengths of 110 semantic categories from SEMCAT. Bhattacharya distance scores are summed across dimensions and then sorted. Red horizontal line represents the baseline strength level obtained for a category composed of 91 randomly selected words from the vocabulary. The metals category has the strongest total representation among SEMCAT categories due to relatively few and well clustered words it contains while the pirate category has the lowest total representation due to widespread words it contains.

2.2.2 Results

2.2.2.1 Semantic Decomposition

The semantic category weights calculated using the method introduced in Section 2.2.1.2 are displayed in Figure 2.2. A close examination of the distribution of category weights indicates that the representation of semantic concepts is broadly distributed across many dimensions of the GloVe embedding space. This suggests that the raw space output by the GloVe algorithm has poor interpretability.

In addition, it can be observed that the total representation strength summed across dimensions varies significantly across categories: some columns in the category weight matrix contain much higher values than others. In fact, total representation strength of a category greatly depends on its word distribution. If a particular category reflects a highly specific semantic concept with relatively few words such as the metals category, category words tend to be well clustered

in the embedding space. This tight grouping of category words results in large Bhattacharya distances in most dimensions, indicating stronger representation of the category. On the other hand, if words from a semantic category are weakly related, it is more difficult for the word embedding to encode their relations. In this case, word vectors are relatively more widespread in the embedding space, and this leads to smaller Bhattacharya distances, indicating that the semantic category does not have a strong representation across embedding dimensions. The total representation strengths of the 110 semantic categories in SEMCAT are shown in Figure 2.3, along with the baseline strength level obtained for a category composed of 91 randomly selected words (91 is the average word count across categories in SEMCAT). The metals category has the strongest total representation among SEMCAT categories due to relatively few and well clustered words it contains, whereas the pirate category has the lowest total representation due to widespread words it contains.

To closely inspect the semantic structure of dimensions and categories, let us investigate the decompositions of three sample dimensions and three specific semantic categories (math, animal and tools). The left column of Figure 2.4 displays the categorical decomposition of the 2^{nd}, 6^{th} and 45^{th} dimensions of the word embedding. While the 2^{nd} dimension selectively represents a particular category (sciences), the 45^{th} dimension focuses on 3 different categories (housing, rooms and sciences) and the 6^{th} dimension has a distributed and relatively uniform representation of many different categories. These distinct distributional properties can also be observed in terms of categories as shown in the right column of Figure 2.4. While only few dimensions are dominant for representing the math category, semantic encodings of the tools and animals categories are distributed across many embedding dimensions.

Note that these results are valid regardless of the random initialization of the GloVe algorithm while learning the embedding space. For the weights calculated for our second GloVe embedding space $\mathcal{E}^2$, where the only difference between $\mathcal{E}$ and $\mathcal{E}^2$ is the independent random initializations of the word vectors before training, we observe nearly identical decompositions for the categories ignoring the order of the dimensions (similar number of peaks and similar total representation

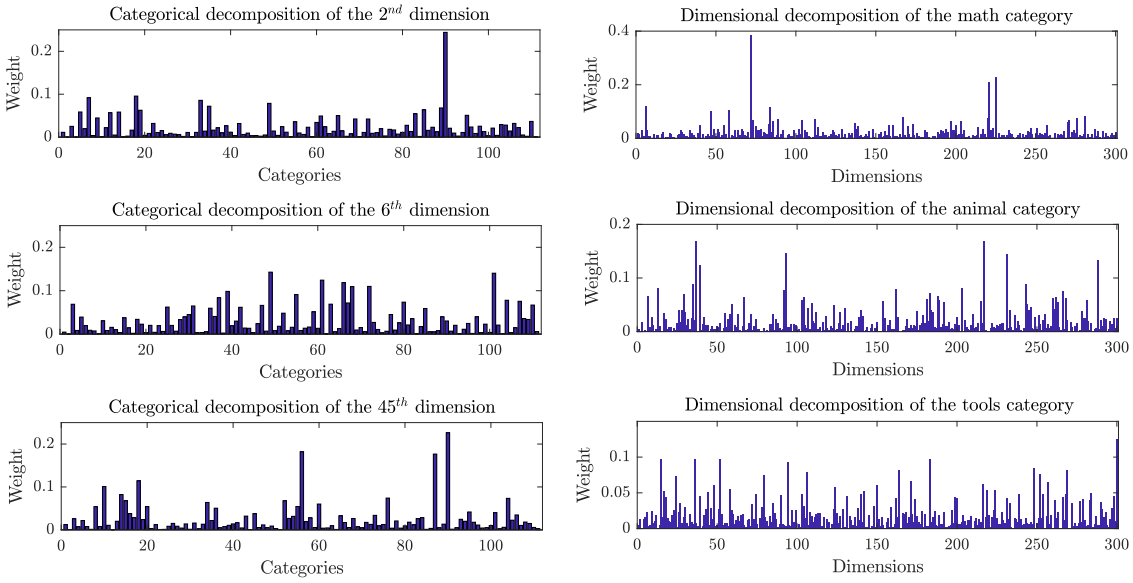

Figure 2.4: Categorical decompositions of the 2^{nd}, 6^{th} and 45^{th} word embedding dimensions are given in the left column. A dense word embedding dimension may focus on a single category (top row), may represent a few different categories (bottom row) or may represent many different categories with low strength (middle row). Dimensional decompositions of the math, animal and tools categories are shown in the right column. Semantic information about a category may be encoded in a few word embedding dimensions (top row) or it can be distributed across many of the dimensions (bottom row).

strength; not shown).

2.2.2.2 Validation

A representative investigation of the semantic space $\mathcal{I}$ is presented in Figure 2.5, where semantic decompositions of four different words, *window*, *bus*, *soldier* and *article*, are displayed using 20 dimensions of $\mathcal{I}$ with largest values for each word. These words are expected to have high values in the dimensions that encode the categories to which they belong. However, we can clearly see from Figure 2.5 that additional categories such as jobs, people, pirate and weapons that are semantically related to *soldier* but that do not contain the word also have high values. Similar observations can be made for *window*, *bus*, and *article*, supporting the conclusion that the category weights spread broadly to many non-category words.

Figure 2.6 presents the semantic decompositions of *window*, *bus*, *soldier* and *article* obtained form $\mathcal{I}^*$ that is calculated using the category centers. Similar to the distributions obtained in $\mathcal{I}$, words have high values for semantically-related categories even when these categories do not contain the words. In contrast to $\mathcal{I}$, however, scores for words are much more uniformly distributed across categories, implying that this alternative approach is less discriminative for categories than the proposed method.

To quantitatively compare $\mathcal{I}$ and $\mathcal{I}^*$, a category word retrieval test is applied and the results are presented in Figure 2.7. As depicted in Figure 2.7, the weights calculated using our method ($\mathcal{W}_B$) significantly outperform the weights from the category centers ($\mathcal{W}_C$). It can be noticed that, using only 25 largest weights from $\mathcal{W}_B$ for each category ($k = 25$) yields higher accuracy in word retrieval compared to the alternative $\mathcal{W}_C$ with any k. This result confirms the prediction that the vectors that we obtain for each category (i.e, columns of $\mathcal{W}_B$) distinguish categories better than their average vectors (i.e, columns of $\mathcal{W}_C$).

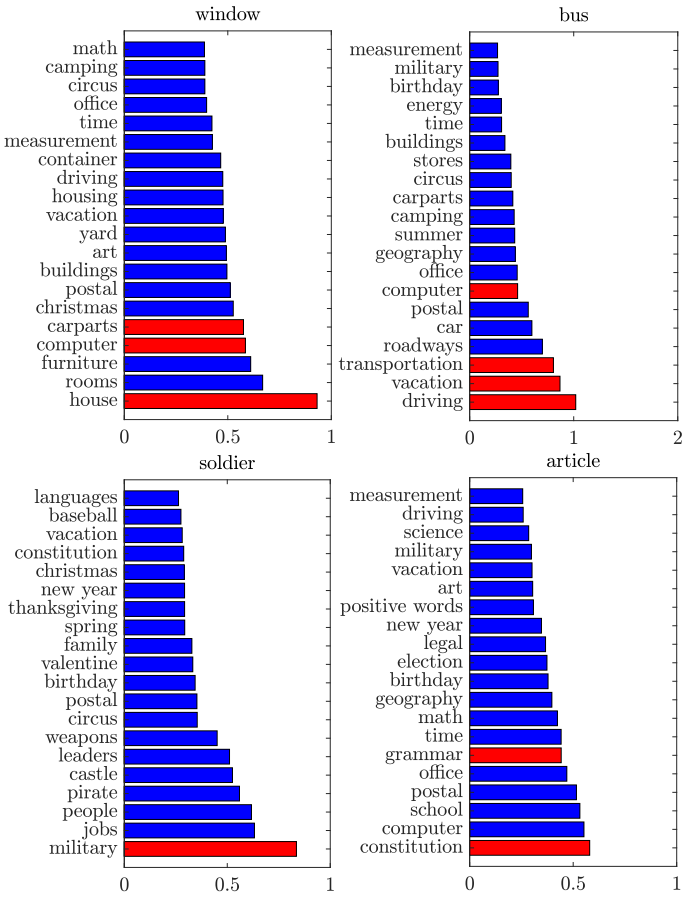

Figure 2.5: Semantic decompositions of the words *window*, *bus*, *soldier* and *article* for 20 highest scoring SEMCAT categories obtained from vectors in $\mathcal{I}$. Red bars indicate the categories that contain the word, blue bars indicate the categories that do not contain the word.

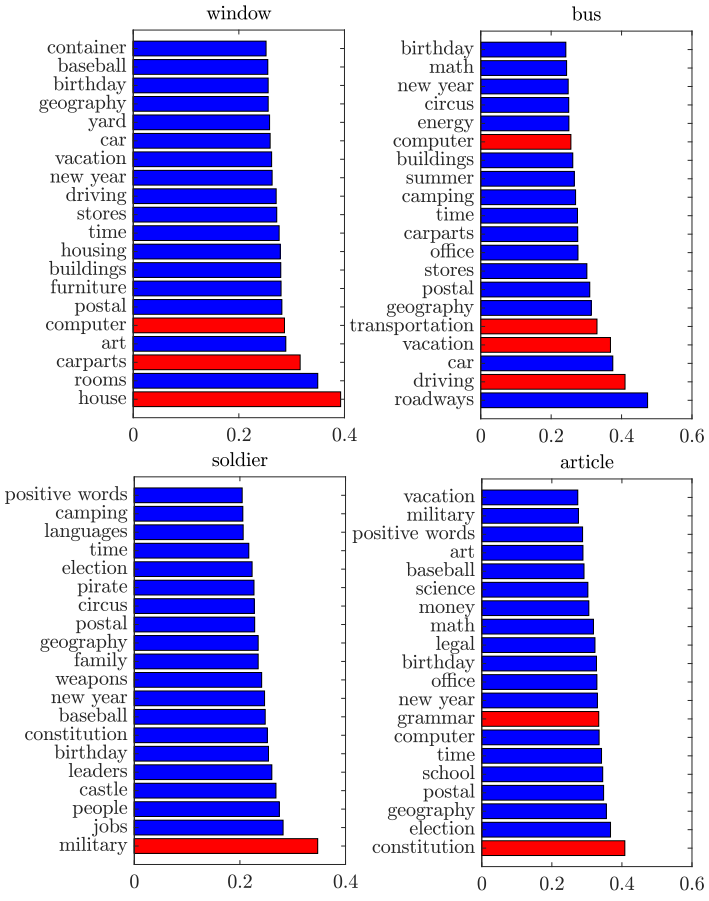

Figure 2.6: Categorical decompositions of the words *window*, *bus*, *soldier* and *article* for 20 highest scoring categories obtained from vectors in $\mathcal{I}^*$. Red bars indicate the categories that contain the word, blue bars indicate the categories that do not contain the word.

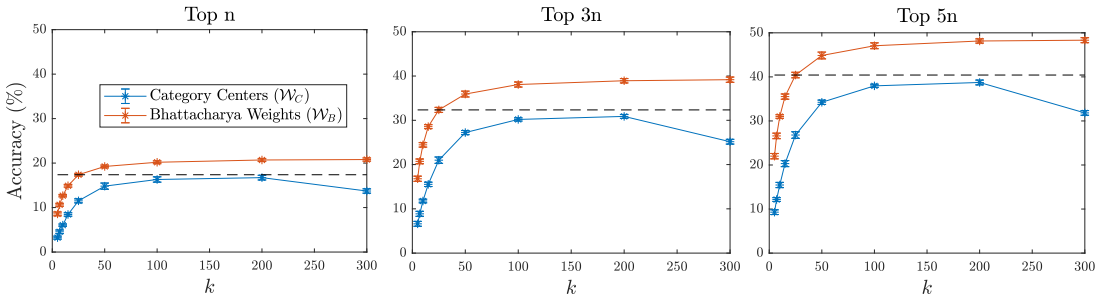

Figure 2.7: Category word retrieval performances for top n, $3n$ and $5n$ words where n is the number of test words varying across categories. Category weights obtained using Bhattacharya distance represent categories better than the center of the category words. Using only 25 largest weights from $\mathcal{W}_B$ for each category ($k = 25$) gives better performance than using category centers with any k (shown with dashed line).

2.3 Measuring Interpretability

2.3.1 Methods

In several studies [30, 31, 39] interpretability is evaluated using the word intrusion test. In the word intrusion test, for each embedding dimension, a word set is generated including the top 5 words in the top ranks and a noisy word (intruder) in the bottom ranks of that dimension. The intruder is selected such that it is in the top ranks of a separate dimension. Then, human editors are asked to determine the intruder word within the generated set. The editors' performances are used to quantify the interpretability of the embedding. Although evaluating interpretability based on human judgement is an effective approach, word intrusion is an expensive method since it requires human effort for each evaluation. Furthermore, the word intrusion test does not quantify the interpretability levels of the embedding dimensions; instead, it yields a binary decision as to whether a dimension is interpretable or not. However, using continuous values is more adequate than making binary evaluations since interpretability levels may vary gradually across dimensions.

2.3.1.1 Measuring Interpretability using SEMCAT

We propose a framework that addresses both of these issues by providing automated, continuous valued evaluations of interpretability while keeping the basis of the evaluations as human judgement. The basic intuition behind our framework is that humans interpret dimensions by trying to group the most distinctive words in the dimensions (i.e, top or bottom rank words), an idea also leveraged by the word intrusion test. Based on this key intuition, it can be noted that if a dataset represents all the possible groups humans can form, instead of relying on human evaluations, one can simply check whether the distinctive words of the embedding dimensions are present together in any of these groups. As discussed earlier, the number of groups humans can form is theoretically unbounded; therefore, it is not possible to compile a comprehensive dataset for all potential groups. However, we claim that a dataset with a sufficiently large number of categories can still provide a good approximation to human judgement. Based on this claim, we propose a simple method to quantify the interpretability of the embedding dimensions.

We define two interpretability scores for an embedding dimension-category pair as:

$$\begin{aligned} IS^{+}_{i,j} &= \frac{|S_j \cap V_i^{+}(\lambda \times n_j)|}{n_j} \times 100 \\ IS^{-}_{i,j} &= \frac{|S_j \cap V_i^{-}(\lambda \times n_j)|}{n_j} \times 100 \end{aligned} \tag{2.5}$$

where $IS^{+}_{i,j}$ is the interpretability score for the positive direction and $IS^{-}_{i,j}$ is the interpretability score for the negative direction for the i^{th} dimension ($i \in \{1, 2, ..., D\}$, where D is the dimensionality of the embedding) and j^{th} category ($j \in \{1, 2, ..., K\}$, where K is the number of categories in the dataset). S_j is the set representing the words in the j^{th} category, n_j is the number of the words in the j^{th} category and $V_i^{+}(\lambda \times n_j)$, $V_i^{-}(\lambda \times n_j)$ refer to the distinctive words located at the top and bottom ranks of the i^{th} embedding dimension, respectively. $\lambda \times n_j$ is the number of words taken from the upper and bottom ranks, where λ is the

parameter determining how strict the interpretability definition is. The smallest value for λ is 1 and it corresponds to the most strict definition; larger λ values relax the definition by increasing the range for selected category words. $\cap$ is the intersection operator between category words and top and bottom ranks words, $|.|$ is the cardinality operator.

We take the maximum of scores in the positive and negative directions as the overall interpretability score for a category ($IS_{i,j}$). The interpretability score of a dimension is then taken as the maximum of individual category interpretability scores across that dimension (IS_i). Finally, we calculate the overall interpretability score of the embedding (IS) as the average of the dimension interpretability scores:

$$\begin{aligned} IS_{i,j} &= \max(IS^+_{i,j}, IS^-_{i,j}) \\ IS_i &= \max_j IS_{i,j} \\ IS &= \frac{1}{D}\sum_{i=1}^{D} IS_i \end{aligned} \tag{2.6}$$

We test our method on the GloVe embedding space, on the semantic spaces $\mathcal{I}$ and $\mathcal{I}^*$, and on a random space where word vectors are generated by randomly sampling from a zero mean. unit variance normal distribution. Interpretability scores for the random space are taken as our baseline. We measure the interpretability scores while λ values are varied from 1 (strict interpretability) to 10 (relaxed interpretability).

Our interpretability measurements are based on our proposed dataset SEMCAT, which was designed to be a comprehensive dataset that contains a diverse set of word categories. Yet, it is possible that the precise interpretability scores that are measured here are biased by the dataset used. In general, two main properties of the dataset can affect the results: category selection and within-category word selection. To examine the effects of these properties on interpretability evaluations, we create alternative datasets by varying both category selection

and word selection for SEMCAT. Since SEMCAT is comprehensive in terms of the words it contains for the categories, these datasets are created by subsampling the categories and words included in SEMCAT. Since random sampling of words within a category may perturb the capacity of the dataset in reflecting human judgement, we subsample r% of the words that are closest to category centers within each category, where $r \in \{40, 60, 80, 100\}$. To examine the importance of number of categories in the dataset we randomly select m categories from SEMCAT, where $m \in \{30, 50, 70, 90, 110\}$. We repeat the selection 10 times, independently for each m.

For an embedding dimension i to have a large interpretability value based on a category j ($IS_{i,j}$), most of the words in category j must be among the most active words in that dimension. In other words, the dimension has to encode the entire category. However, most of the categories in SEMCAT are large (i.e, more than 100 words), making the concepts represented by these categories relatively broad. An embedding dimension may encode a specific concept that is represented by only a small part of a category, instead of entire category, and still be interpretable. Based on this argument. we propose the following alternative to (2.5) that takes possible subcategories into account.

$$\begin{aligned} IS^{+}_{i,j} &= \max_{n_{min} \leq n \leq n_j} \frac{|S_j \cap V^{+}_i(\lambda \times n)|}{n} \times 100 \\ IS^{-}_{i,j} &= \max_{n_{min} \leq n \leq n_j} \frac{|S_j \cap V^{-}_i(\lambda \times n)|}{n} \times 100 \end{aligned} \tag{2.7}$$

where n_{min} is the minimum number of words required to represent a concept. (2.7) can be considered as an extension to (2.5) since they are equal for $n = n_j$. For $n_{min} < n_j$, (2.7) also calculates interpretability scores for all possible subcategories of any size down to n_{min} within the given category and takes the maximum of them as $IS^{+}_{i,j}$ and $IS^{-}_{i,j}$ for the positive and negative directions, respectively.

Note that (2.7) may result in $IS^{+}_{i,j}$ and $IS^{-}_{i,j}$ values that are greater than 100. These values are taken as 100 for the rest of the calculations (i.e, (2.6)).

Table 2.3: Comparison of ANKAT and SEMCAT

	ANKAT	SEMCAT
Number of Categories	62	110
Number of Unique Words	4096	6559
Average Word Count	79	91
Minimum Word Count	22	20
Maximum word count	201	276

2.3.1.2 Interpretability for Turkish Word Embeddings

Since the proposed interpretability measurement method is based on a category dataset, it can only evaluate interpretability of word embeddings for the same language with the dataset. In order to evaluate interpretability level of Turkish word embeddings, we propose a new Turkish category dataset called ANKAT (ANlamsal KATegori) composed of 62 different semantic categories. SEMCAT and ANKAT are compared in Table 2.3 and ten sample words from six representative ANKAT categories are given in Table 2.4.

Turkish Wikipedia is taken as the source corpus to learn Turkish word embeddings. As pre-processing, infleational suffixes are removed from the words using the *zemberek*[3] natural language processing library for Turkish language. Moreover, non-alphanumeric characters are removed from the corpus and all letters are converted to lowercase. The resulting corpus contains 50,855,950 tokens with 820,446 unique tokens.

The resulting Turkish corpus is used to train GloVe and word2vec (skip-gram with negative sampling) embedding algorithms. Then, I and I^* interprertable embedding spaces are obtained by mapping the GloVe embedding space to ANKAT categories following the same steps given in Section 2.2.1.3. Finally, the vocabulary is restricted to most frequent 50,000 words from the corpus for all embedding spaces.

[3]https://github.com/ahmetaa/zemberek-nlp

Table 2.4: Ten sample words from each of the six representative ANKAT categories.

Aile	**Duygular**	**Mutfak**	**Müzik Aletleri**	**Seyahat**	**Zaman**
akraba	acıma	bardak	arp	acenta	ağustos
bacanak	bıkkınlık	bulaşık	bateri	bavul	asır
bebek	düşmanlık	deterjan	çan	bilet	ay
boşanmak	gurur	fırın	flüt	gümrük	çarşamba
damat	korku	kavanoz	gitar	harita	dakika
düğün	merak	oklava	mandolin	liman	hafta
elti	sabır	maşa	obua	pasaport	öğlen
eş	suçluluk	sürahi	piyano	rezervasyon	sonbahar
kayınpeder	umut	tava	tuba	varış	yakında
torun	yalnızlık	tepsi	viyola	yolcu	yıl

2.3.2 Results

Figure (2.8) displays the interpretability scores calculated from SEMCAT using (2.5) and 2.6 for the GloVe embedding, $\mathcal{I}$, $\mathcal{I}^*$ and the random embedding for varying λ values. λ can be considered as a design parameter adjusted according to the interpretability definition. Increasing λ relaxes the interpretability definition by allowing category words to be distributed on a wider range around the top ranks of a dimension. We observe that $\lambda = 5$ is an adequate choice that yields a similar evaluation to measuring the top-5 error in category word retrieval tests. As clearly depicted, semantic space $\mathcal{I}$ is significantly more interpretable than the GloVe embedding, as justified in Section 2.2.2.2. We can also see that interpretability score of the GloVe embedding is close to the random embedding representing the baseline interpretability level.

Interpretability scores for datasets constructed by sub-sampling SEMCAT are given in Table 2.5 for the GloVe, $\mathcal{I}$, $\mathcal{I}^*$ and random embedding spaces for $\lambda = 5$. Interpretability scores for all embeddings increase as the number of categories in

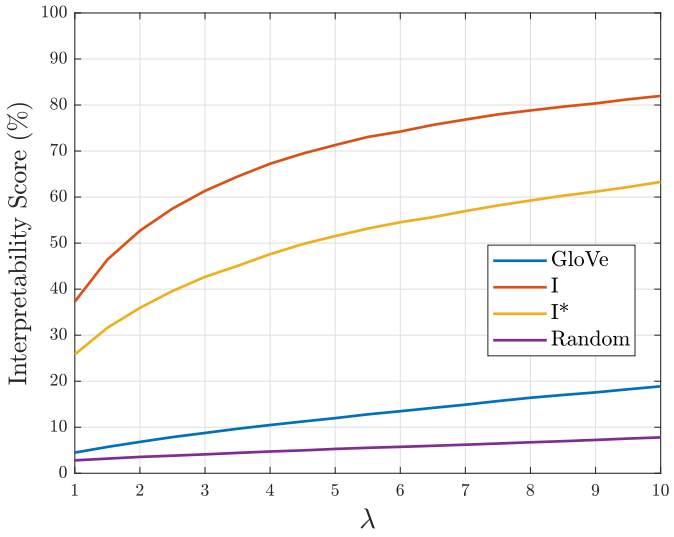

Figure 2.8: Interpretability scores for GloVe, $\mathcal{I}$, $\mathcal{I}^*$ and random embeddings for varying λ values where λ is the parameter determining how strict the interpretability definition is ($\lambda = 1$ is the most strict definition. $\lambda = 10$ is a relaxed definition). Semantic spaces $\mathcal{I}$ and $\mathcal{I}^*$ are significantly more interpretable than GloVe, as expected. $\mathcal{I}$ outperforms $\mathcal{I}^*$ suggesting that weights calculated with our proposed method represent categories more distinctively as opposed to the weights calculated as the category centers. Interpretability scores of Glove are close to the baseline (Random) implying that the dense word embedding has poor interpretability.

the dataset increase (30, 50, 70, 90, 110) for each category coverage (40%, 60%, 80%, 100%). This is expected since increasing the number of categories corresponds to taking into account human interpretations more substantially during evaluation. One can further argue that the true interpretability scores of the embeddings (i.e, scores from an all-comprehensive dataset) should be even larger than those presented in Table 2.5. However, it can also be noticed that the increase in the interpretability scores of the GloVe and random embedding spaces gets smaller for larger number of categories. Thus, there are diminishing returns to increasing number of categories in terms of interpretability. Another important observation is that the interpretability scores of $\mathcal{I}$ and $\mathcal{I}^*$ are more sensitive to number of categories in the dataset than the GloVe or random embeddings. This can be attributed to the fact that $\mathcal{I}$ and $\mathcal{I}^*$ comprise dimensions that correspond to SEMCAT categories, and that inclusion or exclusion of these categories more

Table 2.5: Average Interpretability Scores (%) for $\lambda = 5$. Results are averaged across 10 independent selections of categories for each category coverage.

Category Coverage (%)		Number of Categories				
		30	50	70	90	110
40	Random	4.9	5.5	6.0	6.4	6.7
	GloVe	5.6	6.8	7.7	8.3	8.9
	$\mathcal{I}^*$	25.9	33.6	40.2	44.8	49.1
	$\mathcal{I}$	34.2	45.2	55.5	62.9	69.2
60	Random	4.5	4.9	5.3	5.6	5.8
	GloVe	6.7	7.8	9.0	9.7	10.2
	$\mathcal{I}^*$	27.6	35.8	42.4	47.7	51.6
	$\mathcal{I}$	36.1	48.4	59.0	67.0	72.8
80	Random	4.2	4.6	4.9	5.1	5.3
	GloVe	7.6	8.9	9.7	10.4	11.0
	$\mathcal{I}^*$	30.2	31.1	43.2	48.1	52.0
	$\mathcal{I}$	39.8	50.7	60.1	67.4	73.2
100	Random	4.3	4.6	4.8	5.0	5.1
	GloVe	8.4	9.8	10.8	11.4	12.0
	$\mathcal{I}^*$	30.3	37.7	43.4	48.1	51.5
	$\mathcal{I}$	38.9	49.9	59.0	65.7	71.3

directly affects interpretability.

In contrast to the category coverage, the effects of within-category word coverage on interpretability scores can be more complex. Starting with few words within each category, increasing the number of words is expected to sample more uniformly from the word distribution, reflect more accurately the semantic relations within each category and thereby enhance interpretability scores. However, having categories over-abundant in words might inevitably weaken semantic correlations among them, reducing the discriminability of the categories and interpretability of the embedding. Interestingly, Table 2.5 shows that changing the category coverage has different effects on the interpretability scores of different types of embeddings. As category word coverage increases, interpretability scores for random embedding gradually decrease while they monotonically increase for the GloVe embedding. For semantic spaces $\mathcal{I}$ and $\mathcal{I}^*$. interpretability scores increase as the category coverage increases up to 80% of that of SEMCAT, then

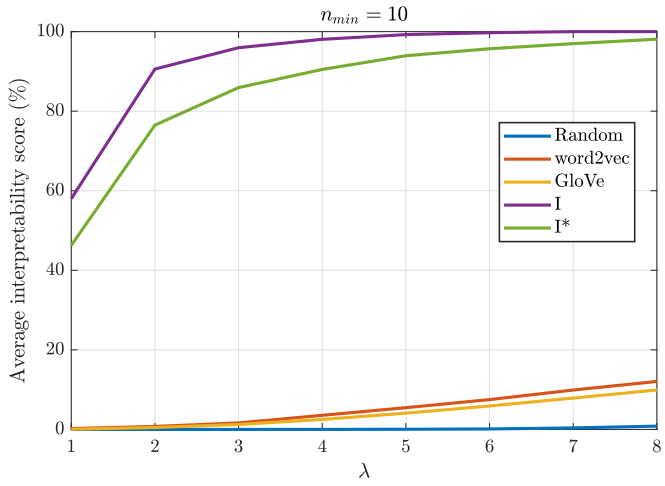

Figure 2.9: Average interpretability scores of four Turkish embedding spaces along with a random baseline for $n_{min} = 10$ and $\lambda \in \{1, 2, ..., 8\}$.

the scores decrease. This may be a result of having too comprehensive categories (as argued earlier), implying that categories with coverage of around 80% of SEMCAT are better suited for measuring interpretability. However, it should be noted that the change in the interpretability scores for different word coverages might be effected by non-ideal subsampling of category words. Although our word sampling method, based on words' distances to category centers, is expected to generate categories that are represented better compared to random sampling of category words. category representations might be suboptimal compared to human designed categories.

Interpretability of the Turkish word embedding spaces are evaluated using ANKAT dataset and the alternative evaluation method, given in (2.7), that considers possible subcategories for varying λ and n_{min}. Figure 2.9 demonstrates interpretability scores for the four Turkish word embedding spaces along with the random embedding space for $n_{min} = 10$ and $\lambda \in \{1, 2, ..., 8\}$. It can be seen that interpretability scores of both GloVe and word2vec are close to random baseline implying their uninterpretable structure. On the other hand, interpretability scores of semantic spaces I and I^* are significantly higher (even close to 100, maximum interpretability) for large λ. This result was expected since dimensions of I and I^* directly correspond to ANKAT categories.

Table 2.6: Average interpretability scores (%) of the five embedding spaces for different λ and n_{min} values.

		λ							
		1	2	3	4	5	6	7	8
$n_{min} = 5$	Random	0.0	0.2	0.9	2.2	3.8	5.2	6.5	7.5
	word2vec	1.1	3.7	7.1	10.4	13.3	16.3	18.9	21.4
	GloVe	0.8	3.0	6.1	9.1	12.0	15.0	17.5	19.9
	$\mathcal{I}$	71.4	96.3	99.3	99.7	100	100	100	100
	$\mathcal{I}^*$	57.4	86.9	93.8	96.9	98.6	98.8	99.2	99.5
$n_{min} = 10$	Random	0.0	0.0	0.0	0.0	0.0	0.1	0.4	0.8
	word2vec	0.2	0.8	1.6	3.5	5.5	7.5	9.9	12.1
	GloVe	0.0	0.5	1.2	2.5	4.1	5.9	7.9	9.9
	$\mathcal{I}$	58.0	90.6	96.0	98.1	99.3	99.7	100	100
	$\mathcal{I}^*$	46.3	76.5	85.9	90.5	93.9	95.7	97.0	98.1
$n_{min} = 15$	Random	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	word2vec	0.1	0.1	0.5	0.9	1.7	2.9	4.2	5.4
	GloVe	0.0	0.0	0.4	0.9	1.5	2.5	3.2	4.5
	$\mathcal{I}$	42.9	79.5	92.1	96.2	97.8	98.3	98.8	99.3
	$\mathcal{I}^*$	31.9	63.0	78.8	85.6	89.8	93.0	94.7	96.0

Table 2.6 presents the measured interpretability levels of the five embedding spaces for $n_{min} \in \{5, 10, 15\}$ and $\lambda \in \{1, 2, ..., 8\}$. It can be observed that interpretability levels increase with decreasing n_{min}, as expected. This is because the number of possible subcategories significantly increases for lower n_{min}, making it easier to get larger interpretability values. These results suggest that for $\lambda = 5$ (as stated earlier), n_{min} values between 10 and 15 are adequate choices since interpretability levels for the random embeddings are around 0 while they are around 100 for I (a perfectly interpretable embedding space).

2.4 Discussion

In this chapter, we proposed a statistical method to uncover the latent semantic structure in dense word embeddings. Based on a new dataset (SEMCAT) we introduced that contains more than 6,500 words semantically grouped under 110

categories, we provided a semantic decomposition of the word embedding dimensions and verify our findings using qualitative and quantitative tests. We also introduced a method to quantify the interpretability of word embeddings based on SEMCAT that can replace the word intrusion test that relies heavily on human effort while keeping the basis of the interpretations as human judgement. Our proposed method to investigate the hidden semantic structure in the embedding space is based on calculation of category weights using a Bhattacharya distance metric. This metric implicitly assumes that the distribution of words within each embedding dimension is normal. Our statistical assessments indicate that the GloVe embedding space considered here closely follows this assumption. In applications where the embedding method yields distributions that significantly deviate from a normal distribution, nonparametric distribution metrics such as Spearman's correlation could be leveraged as an alternative. The resulting category weights can be input seamlessly to the remaining components of our framework.

Since our proposed framework for measuring interpretability depends solely on the selection of the category words dataset, it can be used to directly compare different word embedding methods (e.g, GloVe, word2vec, fastText) in terms of the interpretability of the resulting embedding spaces. A straightforward way to do this is to compare the category weights calculated for embedding dimensions across various embedding spaces. Note, however, that the Bhattacharya distance metric for measuring the category weights does not follow a linear scale and is unbounded. For instance, consider a pair of embeddings with category weights 10 and 30 versus another pair with weights 30 and 50. For both pairs, the latter embedding can be deemed more interpretable than the former. Yet, due to the gross nonlinearity of the distance metric, it is challenging to infer whether a 20-unit improvement in the category weights corresponds to similar levels of improvement in interpretability across the two pairs. To alleviate these issues, here we propose an improved method that assigns normalized interpretability scores with an upper bound of 100%. This method facilitates interpretability assessments and comparisons among embedding spaces.

The results reported in this chapter for semantic analysis and interpretability

assessment of embeddings are based on SEMCAT. SEMCAT contains 110 different semantic categories where average number of words per category is 91, rendering SEMCAT categories quite comprehensive. Although the HyperLex dataset contains a relatively larger number of categories (1399), the average number of words per category is only 2, insufficient to accurately represent semantic categories. Furthermore, while HyperLex categories are constructed based on a single type of relation among words (hyponym). SEMCAT is significantly more comprehensive since many categories include words that are grouped based on diverse types of relationships that go beyond hypernym-hyponym relations. Meanwhile, the relatively smaller number of categories in SEMCAT is not considered a strong limitation, as our analyses indicate that the interpretability levels exhibit diminishing returns when the number of categories in the dataset are increased and that SEMCAT is readily yielding near optimal performance. Moreover, our proposed alternative interpretability measurement method considers possible subgroups in large SEMCAT categories which effectively corresponds to dividing SEMCAT categories to smaller concept groups of all possible sizes. That said, extended datasets with improved coverage and expert labeling by multiple observers would further improve the reliability of the proposed approach. To do this, a synergistic merge with existing lexical databases such as WordNet might prove useful.

As an extension of our interpretability measurement studies, we introduced a new category dataset (ANKAT) for Turkish language and measured interpretability of Turkish word embeddings. Extension of our proposed method to other languages is expected to be straightforward. One needs to only construct a category dataset in the desired language which is a one-time effort.

Methods for learning dense word embeddings remain an active area of NLP research. The framework proposed in this chapter enables quantitative assessments on the intrinsic semantic structure and interpretability of word embeddings. Therefore, the proposed framework can be a valuable tool in guiding future research on obtaining interpretable yet effective embedding spaces for many NLP tasks that critically rely on semantic information. For instance, performance evaluation of more interpretable word embeddings on higher level NLP tasks (i.e, sentiment analysis, named entity recognition, question answering) and the

relation between interpretability and NLP performance can be a future project.

Chapter 3

Imparting Interpretability

After the introduction of the *word2vec* algorithm by Mikolov [14, 8], there has been a growing interest in algorithms that generate improved word representations under some performance metric. Significant effort is spent in appropriately modifying the objective functions of the algorithms in order to incorporate knowledge from external resources, with the purpose of increasing the performance of the resulting word representations [44, 7, 45, 46, 47, 48, 49, 50, 51, 52]. Inspired by the line of work reported in these studies, in this chapter we propose to use modified objective functions for a different purpose: learning more interpretable dense word embeddings. By doing this, we aim to incorporate semantic information from an external lexical resource into the word embedding so that the embedding dimensions are *aligned* along predefined concepts. This alignment is achieved by introducing a modification to learning process of embeddings. In our proposed method, which is built on top of the GloVe algorithm [16], the cost function for any one of the words of concept word-groups is modified by the introduction of an additive term to the cost function. Each embedding vector dimension is first associated with a concept. For a word belonging to any one of the word-groups representing these concepts, the modified cost term favors an increase for the value of this word's embedding vector dimension corresponding to the concept that the particular word belongs to. For words that do not belong to any one of the word-groups, the cost term is left untouched. Specifically, *Roget's*

Thesaurus [53, 54] is used to derive the concepts and concept word-groups to be used as the external lexical resource for our proposed method. We quantitatively demonstrate the increase in interpretability by using the measure given (2.7) and (2.6), as well as demonstrating qualitative results. We also show that the semantic structure of the original embedding has not been harmed in the process since there is no performance loss in standard word-similarity or word-analogy tests.

This chapter is organized as follows: In Section 3.1, we discuss previous studies related to our work under two main categories: interpretability of word embeddings and joint-learning frameworks where the objective function is modified. In Section 3.2, we present the problem framework and provide the formulation within the GloVe [16] algorithm setting. In Section 3.3, where our approach is proposed, we motivate and develop a modification to the original objective function with the aim of increasing representation interpretability. In Section 3.4, experimental results are provided and the proposed method is quantitatively and qualitatively evaluated. Additionally, in Section 3.4, results demonstrating the extent to which the original semantic structure of the embedding space is affected are presented by using word-analogy and word-similarity tests. We conclude the chapter in Section 3.5 with a discussion.

3.1 Related Work

Methodologically, our work is related to prior studies that aim to obtain 'improved' word embeddings using external lexical resources, under some performance metric. Previous work in this area can be divided into two main categories: i) works that *modify* the word embedding learning algorithm to incorporate lexical information, ii) works that operate on pre-trained embeddings with a *post-processing* step.

Among works that follow the first approach, [44] extends the Skip-Gram model by incorporating the word similarity relations extracted from the Paraphrase Database (PPDB) and WordNet [7], into the Skip-Gram predictive model as an

additional cost term. In [45], the authors extend the CBOW model by considering two types of semantic information, called *relational* and *categorical*, to be incorporated into the embeddings during training. For the former type of semantic information, the authors propose the learning of explicit vectors for the different relations extracted from a semantic lexicon such that the word pairs that satisfy the same relation are distributed more homogeneously. For the latter, the authors modify the learning objective such that some weighted average distance is minimized for words under the same semantic category. In [46], the authors represent the synonymy and hypernymy-hyponymy relations in terms of *inequality constraints*, where the pairwise similarity rankings over word triplets are forced to follow an order extracted from a lexical resource. Following their extraction from WordNet, the authors impose these constraints in the form of an additive cost term to the Skip-Gram formulation. Finally, [47] builds on top of the GloVe algorithm by introducing a regularization term to the objective function that facilitates the vector representations of similar words as dictated by WordNet to be similar as well.

Turning our attention to the post-processing approach for enriching word embeddings with external lexical knowledge, [48] has introduced the *retrofitting* algorithm that acts on pre-trained embeddings such as Skip-Gram or GloVe. The authors propose an objective function that aims to balance out the semantic information captured in the pre-trained embeddings with the constraints derived from lexical resources such as WordNet, PPDB and FrameNet. One of the models proposed in [49] extends the retrofitting approach to incorporate the word sense information from WordNet. Similarly, [50] creates multi-sense embeddings by gathering the word sense information from a lexical resource and by learning to decompose the pre-trained embeddings into a convex combination of sense embeddings. In [51], the authors focus on improving word embeddings for capturing word similarity, as opposed to mere relatedness. To this end, they introduce the *counter-fitting* technique which acts on the input word vectors such that synonymous words are brought closer to one another whereas antonymous words are distanced from each other, where the synonymy-antonymy relations are extracted from a lexical resource. More recently, the *ATTRACT-REPEL* algorithm

proposed by [52] improves on counter-fitting by a formulation which imparts the word vectors with external lexical information in *mini-batches*.

Most of the studies discussed above [45, 46, 47, 48, 49, 51, 52] report performance improvements in benchmark tests such as word similarity or word analogy, while [7] uses a different analysis method (mean reciprocal rank). In sum, the literature is rich with studies aiming to obtain word embeddings that perform better under specific performance metrics. However, less attention has been directed to the issue of interpretability of the word embeddings. In the literature, the problem of interpretability has been tackled using different approaches. [30] proposed non-negative matrix factorization (NMF) for learning sparse, interpretable word vectors from co-occurrence variant matrices, where the resulting vector space is called non-negative sparse embeddigns (NNSE). However, since NMF methods require maintaining a global matrix for learning, they suffer from memory and scalability issues. This problem has been addressed in [31], where an online method of learning interpretable word embeddings from corpora using a modified version of skip-gram model [14] is proposed. As a different approach, [32] combined text-based similarity information among words with brain activity based similarity information to improve interpretability using joint non-negative sparse embedding (JNNSE).

A common alternative approach for learning interpretable embeddings is to learn transformations that map pre-trained state-of-the-art embeddings to new interpretable semantic spaces. To obtain sparse, higher dimensional and more interpretable vector spaces, [33] and [34] use sparse coding on conventional dense word embeddings. However, these methods learn the projection vectors that are used for the transformation from the word embeddings without supervision. For this reason, labels describing the corresponding semantic categories cannot be provided. An alternative approach was proposed in [36], where orthogonal transformations were utilized to increase interpretability while preserving the performance of the underlying embedding. However, [36] has also shown that total interpretability of an embedding is kept constant under any orthogonal transformation and that it can only be redistributed across the dimensions. Rotation algorithms based on exploratory factor analysis (EFA) to preserve the performance

of the original word embeddings while improving their interpretability were proposed in [37]. [35] suggested to deploy a sparse auto-encoder using pre-trained dense word embeddings to improve interpretability.

Previous works on interpretability as mentioned above, with the exception of [32] and our proposed method, do not need external resources, utilization of which has both advantages and disadvantages. Methods that do not use external resources require fewer resources but they also lack the benefits of information extracted from these resources.

3.2 Problem Description

For the task of unsupervised word embedding extraction, we operate on a discrete collection of lexical units (words) $u_i \in \mathcal{V}$ that is part of an input corpus $\mathcal{C} = \{u_i\}_{i \geq 1}$, with number of tokens $|\mathcal{C}|$, sourced from a vocabulary $\mathcal{V} = \{w_1, \ldots, w_V\}$ of size V.[1] In the setting of distributional semantics, the objective of a word embedding algorithm is to maximize some aggregate utility over the entire corpus so that some measure of "closeness" is maximized for pairs of vector representations $(\boldsymbol{w}_i, \boldsymbol{w}_j)$ for words which, on the average, appear in proximity to one another. In the GloVe algorithm [16], upon which we base our improvements, the following objective function is considered:

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(\boldsymbol{w}_i^T \tilde{\boldsymbol{w}}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2 \tag{3.1}$$

In (3.1), $\boldsymbol{w}_i \in \mathbb{R}^D$ and $\tilde{\boldsymbol{w}}_j \in \mathbb{R}^D$ stand for word and context vector representations, respectively, for words w_i and w_j, while X_{ij} represents the (possibly weighted) co-occurrence count for the word pair (w_i, w_j). Intuitively, (3.1) represents the requirement that if some word w_i occurs often enough in the context

[1] We represent vectors (matrices) by bold lower (upper) case letters. For a vector $\boldsymbol{a}$ (a matrix $\boldsymbol{A}$), $\boldsymbol{a}^T$ ($\boldsymbol{A}^T$) is the transpose. $\|\boldsymbol{a}\|$ stands for the Euclidean norm. For a set S, $|S|$ denotes the cardinality, $\mathbb{1}_{x \in S}$ is the indicator variable for the inclusion $x \in S$, evaluating to 1 if satisfied, 0 otherwise.

(or vicinity) of another word w_j, then the corresponding word representations should have a large enough inner product in keeping with their large X_{ij} value, up to some bias terms $b_i, \tilde{b}_j$ (and vice versa). $f(\cdot)$ in (3.1) is used as a discounting factor that prohibits rare co-occurrences from disproportionately influencing the resulting embeddings.

The objective (3.1) is minimized using stochastic gradient descent by iterating over the matrix of co-occurrence records $[X_{ij}]$. In the GloVe algorithm, for a given word w_i, the final word representation is taken to be the average of the two intermediate vector representations obtained from (3.1), i.e. $(\boldsymbol{w}_i + \tilde{\boldsymbol{w}}_j)/2$. In the next section, we detail the enhancements made to (3.1) for the purpose of enhanced interpretability, using the aforementioned framework as our basis.

3.3 Imparting Method

Our approach falls into a joint-learning framework where the distributional information extracted from the corpus is allowed to fuse with the external lexicon-based information. *Word-groups* extracted from Roget's Thesaurus are directly mapped to individual dimensions of word embeddings. Specifically, the vector representations of words that belong to a particular group are encouraged to have deliberately increased values in a particular dimension that corresponds to the word-group under consideration. This can be achieved by modifying the objective function of the embedding algorithm to partially influence vector representation distributions across their dimensions over an input vocabulary. To do this, we propose the following modification to (3.1):

$$\begin{aligned} J = \sum_{i,j=1}^{V} f(X_{ij}) \Bigg[& \left(\boldsymbol{w}_i^T \tilde{\boldsymbol{w}}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2 \\ & + \; k \left(\sum_{l=1}^{L} \mathbb{1}_{i \in F_l} \, g(\boldsymbol{w}_{i,l}) + \sum_{l=1}^{L} \mathbb{1}_{j \in F_l} \, g(\tilde{\boldsymbol{w}}_{j,l}) \right) \Bigg] \end{aligned} \tag{3.2}$$

In (3.2), L is the number of word-groups and F_l denotes the indices for the elements of the lth concept word-group which we wish to assign in the vector

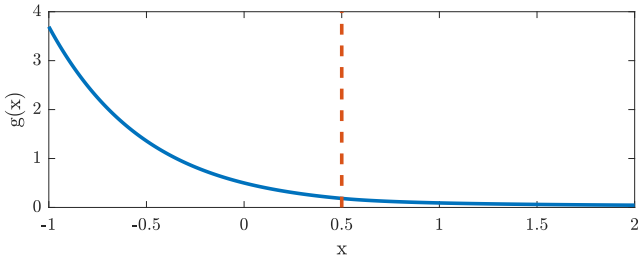

Figure 3.1: Function g in the additional cost term.

dimension $l = 1, \ldots, L$ ($L \leq D$, where D is the number of dimensions). The objective (3.2) is designed as a mixture of two individual cost terms: the original GloVe cost term along with a second term that encourages embedding vectors of a given concept word-group to achieve deliberately increased values along an associated dimension l. The relative weight of the second term is controlled by the parameter k. The simultaneous minimization of both objectives ensures that words that are similar to, but not included in, one of these concept word-groups are also "nudged" towards the associated dimension l. The trained word vectors are thus encouraged to form a distribution where the individual vector dimensions align with certain semantic concepts represented by a collection of concept word-groups, one assigned to each vector dimension. To facilitate this behaviour, (3.2) introduces a monotone decreasing function

$$g(x) = \begin{cases} \dfrac{1}{2}\exp\left(-2x\right) & \text{for } x < 0.5 \\ \dfrac{1}{4ex} & \text{otherwise} \end{cases} \tag{3.3}$$

which serves to increase the total cost incurred if the value of the lth dimension for the two vector representations $\boldsymbol{w}_{i,l}$ and $\tilde{\boldsymbol{w}}_{j,l}$ for a concept word $\boldsymbol{w}_i$ with $i \in F_l$ fails to be large enough. $g(x)$ is plotted in Fig. 3.1.

The objective (3.2) is minimized using stochastic gradient descent over the co-occurrence records $\{X_{ij}\}_{i,j=1}^{V}$. Intuitively, the additional terms in (3.2) introduce the effect of selectively applying a positive step-type input to the original descent updates of (3.1) for concept words along their respective vector dimensions, which influences the dimension value in the positive direction. The parameter k in (3.2)

allows for the adjustment of the magnitude of this influence as needed.

In the next section, we demonstrate the feasibility of this approach experimentally with an example collection of concept word-groups extracted from Roget's Thesaurus.

3.4 Experiments and Results

We first identified 300 concepts ($L = 300$), one for each dimension of the 300-dimensional vector representation, by employing Roget's Thesaurus. This resource follows a tree structure which starts with a *Root* node that contains all the words and phrases in the thesaurus. The root node is successively split into *Classes* and *Sections*, which are then (optionally) split into *Subsections* of various depths, finally ending in *Categories*, which constitute the smallest unit of word/phrase collections in the structure. The actual words and phrases descend from these *Categories*, and make up the leaves of the tree structure. We note that a given word typically appears in multiple categories corresponding to the different senses of the word. To construct concept word-groups from Roget's Thesaurus, we first filtered out the multi-word phrases and the relatively obscure terms from the thesaurus. The obscure terms were identified by checking them against a vocabulary extracted from Wikipedia. We then obtained 300 word-groups as the result of a *partitioning* operation applied to the subtree that ends with *categories* as its leaves. The partition boundaries, hence the resulting word-groups, can be chosen in many different ways. In our proposed approach, we have decided to determine this partitioning by traversing the tree structure in breadth-first order, and by employing a parameter λ for the maximum *size* of a node. Here, the size of a node is defined as the number of unique words that ever-descend from that node. During the traversal, if the size of a given node is less than this threshold, we designate the words that ultimately descend from that node as a concept word-group. Otherwise, if the node has children, we discard the node, and queue up all its children for further consideration. If this node does not have any children, on the other hand, the node is truncated to λ elements with

Table 3.1: Sample concepts and their associated word-groups from Roget's Thesaurus

MANKIND	BUSINESS	SIMPLE QUANTITY	CONDUCT	ARRIVAL
one	living	size	government	land
population	work	way	life	home
people	line	point	game	light
world	place	force	role	airport
state	service	station	race	return
family	role	range	record	come
national	race	standard	process	complete
public	office	rate	business	port
party	act	stage	career	hit
million	case	mass	campaign	meeting
⋮	⋮	⋮	⋮	⋮

the highest frequency-ranks, and the resulting words are designated as a concept word-group. We note that the choice of λ significantly affects the resulting collection of word-groups: Excessively large values result in few word-groups that greatly overlap with one another, while overly small values result in numerous tiny word-groups that fail to adequately represent a concept. We experimentally determined that a λ value of 452 results in the healtiest number of relatively large word-groups (113 groups with size ≥ 100), while yielding a preferably small overlap amongst the resulting word-groups (with average overlap size not exceeding three words). A total of 566 word-groups were thus obtained. 259 smallest word-groups (with size < 38) were discarded to bring down the number of word-groups to 307. Out of these, seven groups with the lowest median frequency-rank were further discarded, giving the final 300 concept word-groups used in the experiments. We present some of the resulting word-groups in Table 3.1.[2]

By using the concept word-groups, we have trained the GloVe algorithm with the proposed modification given in Section 3.3 on a snapshot of Wikipedia measuring 8GB in size, with the stop-words filtered out. Using the parameters given in Table 3.2, this resulted in a vocabulary of size 287,847. For the weighting

[2]All the vocabulary lists, concept word-groups and other material necessary to reproduce this procedure will be made available online.

Table 3.2: GloVe Parameters

VOCAB_MIN_COUNT	65
ALPHA	0.75
WINDOW_SIZE	15
VECTOR_SIZE	300
X_MAX	75

parameter in (3.2), we used a value of $k = 0.1$ whose effect is analysed in detail in Section 3.4.4. The algorithm was trained over 20 iterations. The GloVe algorithm without any modifications was also trained as a baseline with the same parameters. In addition to the original GloVe algorithm, we compare our proposed method with previous studies that aim to obtain interpretable word vectors. We train the *improved projected gradient* model proposed in [31] to obtain word vectors (called OIWE-IPG) using the same corpus we use to train GloVe and our proposed method. Using the methods proposed in [34, 35, 37] on our baseline GloVe embeddings, we obtain SOV, SPINE and Parsimax (orthogonal) word representations, respectively. We train all the models with the proposed parameters. However, in [37], the authors provide results for a relatively small vocabulary of 15,000 words. When we trained their model on our baseline GloVe embeddings with a large vocabulary of size 287,847, the resulting vectors performed significantly worse on word similarity tasks compared to the results presented in their paper. We evaluate the interpretability of the resulting embeddings qualitatively and quantitatively. We also test the performance of the embeddings on word similarity and word analogy tests.

In our experiments, vocabulary size is close to 300,000 while only 16,242 unique words of the vocabulary are present in the concept groups. Furthermore, only dimensions that correspond to the concept group of the word will be updated due to the additional cost term. Given that these concept words can belong to multiple concept groups (two on average), only 33,319 parameters are updated. There are 90 million individual parameters present for the 300,000 word vectors of size 300. Of these parameters, only approximately 33,000 are updated by the additional cost term.

3.4.1 Qualitative Evaluation for Interpretability

In Fig. 3.2, we demonstrate the particular way in which the proposed algorithm (3.2) influences the vector representation distributions. Specifically, we consider, for illustration, the 32^{nd} dimension values for the original GloVe algorithm and our modified version, restricting the plots to the top-1000 words with respect to their frequency ranks for clarity of presentation. In Fig. 3.2, the words in the horizontal axis are sorted in descending order with respect to the values at the 32^{nd} dimension of their word embedding vectors coming from the original GloVe algorithm. The dimension values are denoted with blue and red/green markers for the original and the proposed algorithms, respectively. Additionally, the top-50 words that achieve the greatest 32^{nd} dimension values among the considered 1000 words are emphasized with enlarged markers, along with text annotations. In the presented simulation of the proposed algorithm, the 32^{nd} dimension values are encoded with the concept `JUDGMENT`, which is reflected as an increase in the dimension values for words such as `committee`, `academy`, and `article`. We note that these words (*red*) are *not* part of the pre-determined word-group for the concept `JUDGMENT`, in contrast to words such as `award`, `review` and `account` (*green*) which are. This implies that the increase in the corresponding dimension values seen for these words is attributable to the joint effect of the first term in (3.2) which is inherited from the original GloVe algorithm, in conjunction with the remaining terms in the proposed objective expression (3.2). This experiment illustrates that the proposed algorithm is able to impart the concept of `JUDGMENT` on its designated vector dimension above and beyond the supplied list of words belonging to the concept word-group for that dimension.

We also present the list of words with the greatest dimension value for the dimensions 11, 13, 16, 31, 36, 39, 41, 43 and 79 in Table 3.3. These dimensions are aligned/imparted with the concepts that are given in the column headers. In Table 3.3, the words that are highlighted with green denote the words that exist in the corresponding word-group obtained from Roget's Thesaurus (and are thus explicitly forced to achieve increased dimension values), while the red words

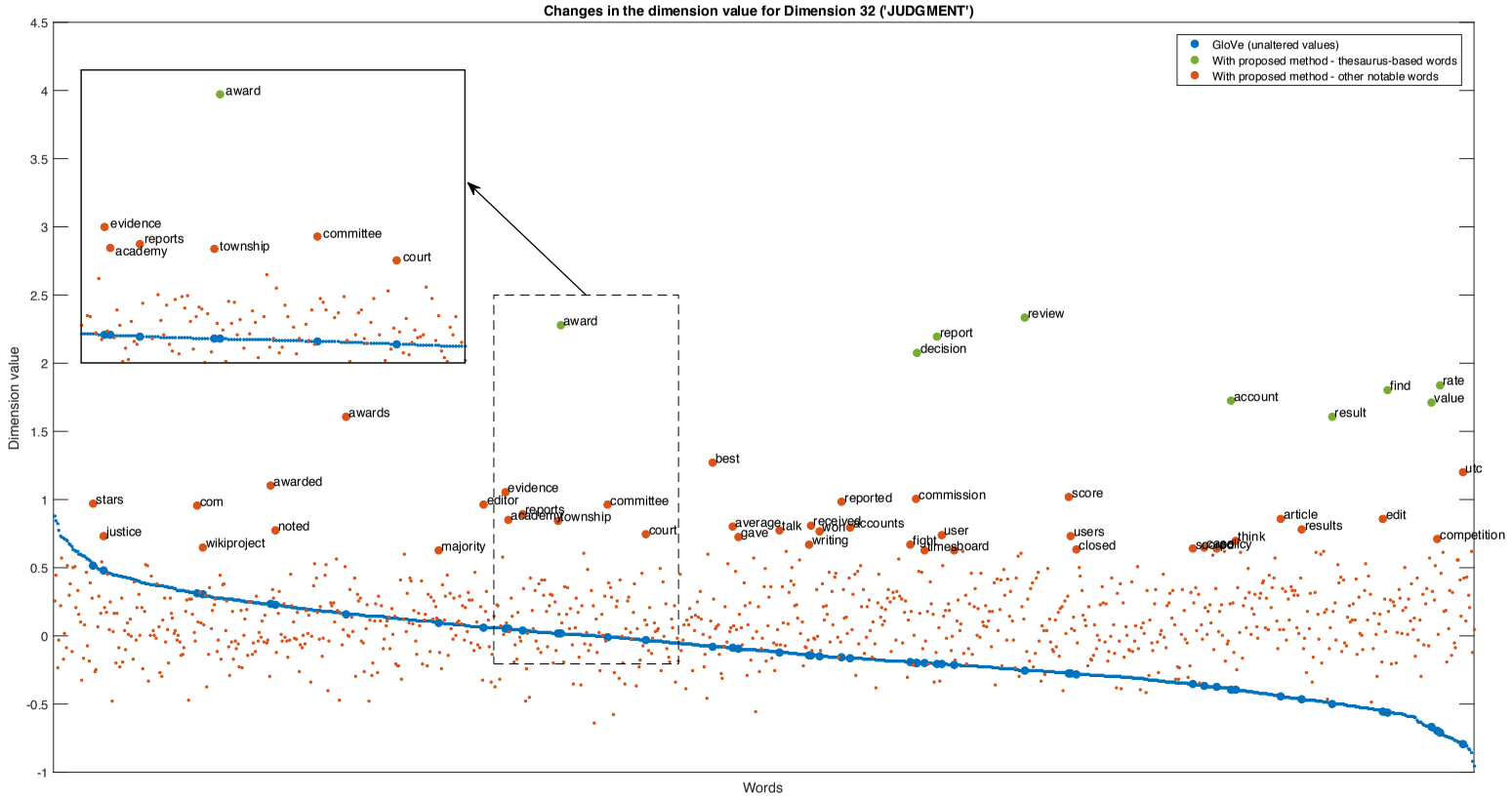

Figure 3.2: Most frequent 1000 words sorted according to their values in the 32^{nd} dimension of the original GloVe embedding are shown with blue markers. Red and green markers show the values of the same words for the 32^{nd} dimension of the embedding obtained with the proposed method where the dimension is *aligned* with the concept `JUDGMENT`. Words with green markers are contained in the concept `JUDGMENT`, while words with red markers are not.

Table 3.3: Words with largest dimension values for the proposed algorithm

GOVERNMENT	CHOICE	BOOK	NEWS	PROPERTY IN GENERAL
republic	poll	editor	radio	lands
province	shortlist	publisher	news	land
provinces	vote	magazine	tv	ownership
government	selection	writer	broadcasting	possession
administration	televoting	author	broadcast	assets
prefecture	preference	hardcover	broadcasts	acquired
governor	choosing	paperback	simulcast	property
county	choose	books	channel	acres
monarchy	choice	page	television	estate
region	chosen	press	cnn	lease
territory	elect	publishing	jazeera	inheritance
autonomous	list	edited	fm	manor
administrative	election	volume	programming	holdings
minister	select	encyclopedia	bbc	ploughs
senate	preferential	published	newscast	estates
districts	option	publications	simulcasts	owner
democratic	voters	bibliography	syndicated	feudal
legislature	ballots	periodical	media	heirs
abolished	votes	publication	reporter	freehold
presidency	sssis	essayist	cbs	holding
⋮	⋮	⋮	⋮	⋮

TEACHING	NUMBERS IN THE ABSTRACT	PATERNITY	WARFARE	FOOD
curriculum	integers	family	battle	meal
exam	polynomial	paternal	war	dishes
training	integer	maternal	battles	bread
school	polynomials	father	combat	eaten
students	logarithm	grandfather	military	dessert
toefl	modulo	grandmother	warfare	cooked
exams	formula	mother	fighting	foods
teaching	coefficients	ancestry	battlefield	dish
schools	multiplication	son	guerrilla	food
education	finite	hemings	fought	meat
teach	logarithms	ancestor	campaign	eating
karate	algebra	patrilineal	fight	cuisine
taught	integrals	daughter	insurgency	beverage
courses	primes	grandson	armed	soup
civics	divisor	descent	tactics	snack
instruction	compute	house	operations	pork
syllabus	arithmetic	parents	army	eat
test	algorithm	descendant	mujahideen	wine
examinations	theorem	grandparents	armies	beef
instructor	quadratic	line	soldiers	fried
⋮	⋮	⋮	⋮	⋮

Table 3.4: Words with largest dimension values for the proposed algorithm - Less Satisfactory Examples

MOTION	TASTE	REDUNDANCY	FEAR
nektonic	polish	eusebian	horror
rate	classical	margin	fear
mobile	taste	drug	dread
movement	culture	arra	trembling
motion	corinthian	overflow	scare
evolution	przeworsk	overdose	terror
gait	artistic	extra	panic
velocity	judge	excess	anxiety
novokubansk	aesthetic	bonus	$\phi\grave{o}\beta o\varsigma$
brownian	amateur	synaxarion	phobia
port	critic	load	fright
flow	kraków	padding	terrible
gang	elegance	crowd	frighten
roll	aesthetics	redundancy	pale
stride	plaquemine	overrun	vacui
run	judgment	boilerplate	haunt
kinematics	connoisseur	excessive	afraid
stream	katarzyna	$\tau\iota\tau\lambda o\iota$	fearful
walk	cucuteni	lavish	frightened
drift	warsaw	gorge	shaky
⋮	⋮	⋮	⋮

denote the words that achieve increased dimension values by virtue of their co-occurrence statistics with the thesaurus-based words (indirectly, without being explicitly forced). This again illustrates that a semantic concept can indeed be coded to a vector dimension provided that a sensible lexical resource is used to guide semantically related words to the desired vector dimension via the proposed objective function in (3.2). Even the words that do not appear in, but are semantically related to the word-groups that we formed using Roget's Thesaurus, are indirectly affected by the proposed algorithm. They also reflect the associated concepts at their respective dimensions, even though the objective functions for their particular vectors are not modified. This point cannot be overemphasized. Although the word-groups extracted from Roget's Thesaurus impose a degree of supervision to the process, the fact that the remaining words in the entire vocabulary are also indirectly affected makes the proposed method a semi-supervised approach that can handle words that are not in these chosen word-groups. A qualitative example of this result can be seen in the last column of Table 3.3. It is interesting to note the appearance of words such as `guerilla`, `insurgency`, `mujahideen`, `Wehrmacht` and `Luftwaffe` in addition to the more obvious and straightforward `army`, `soldiers` and `troops`, all of which are not present in the associated word-group `WARFARE`.

Most of the dimensions we investigated exhibit similar behaviour to the ones presented in Table 3.3. Thus, generally speaking, we can say that the entries in Table 3.3 are representative of the great majority. However, we have also specifically looked for dimensions that make less sense and determined a few such dimensions which are relatively less satisfactory (Table 3.4). These examples are also interesting in that they shed light into the limitations posed by polysemy and existence of very rare, outlier words.

3.4.2 Quantitative Evaluation for Interpretability

One of the main goals of this study is to improve the interpretability of dense word embeddings by aligning the dimensions with predefined concepts from a suitable

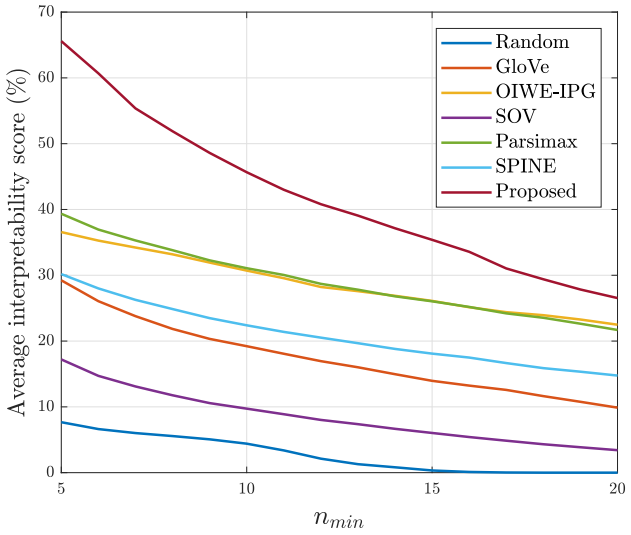

Figure 3.3: Interpretability scores averaged over 300 dimensions for the original GloVe method, the proposed method, and four alternative methods along with a randomly generated baseline embedding for $\lambda = 5$. Embedding generated by the proposed method is significantly more interpretable than the alternatives.

lexicon. A quantitative measure is required to evaluate the achieved improvement reliably. One of the methods proposed to measure the interpretability is the word intrusion test [39]. But, this method is expensive to apply since it requires evaluations from multiple human evaluators for each embedding dimension. In this study, we use a semantic category-based approach based on the method and category dataset (SEMCAT) introduced in Section 2.3 to quantify interpretability. Specifically, we apply the modified version of the approach presented in (2.7) in order to consider possible sub-groupings within the categories[3] in SEMCAT.

Fig. 3.3 presents the measured average interpretability scores across dimensions for original GloVe embeddings, for the proposed method and for the other four methods we compare, along with a randomly generated embedding. Results are calculated for the parameters $\lambda = 5$ and $n_{min} \in \{5, 6, \ldots, 20\}$. Our proposed method significantly improves the interpretability for all n_{min} compared to the original GloVe approach and the alternatives.

[3]Note that the usage of "category" here in the setting of SEMCAT should not be confused with the "categories" of Roget's Thesaurus.

The proposed method and interpretability measurements are both based on using concepts represented by word-groups. Therefore, it is expected that there will be higher interpretability scores for some of the dimensions for which the imparted concepts are also contained in SEMCAT. However, by design, word groups that they use are formed by using different sources and are independent. Interpretability measurements use SEMCAT, while our proposed method uses Roget's Thesaurus.

3.4.3 Intrinsic Evaluation of the Embeddings

It is necessary to show that the semantic structure of the original embedding has not been damaged or distorted as a result of aligning the dimensions with given concepts, and that there is no substantial sacrifice involved from the performance that can be obtained with the original GloVe. To check this, we evaluate performances of the proposed embeddings on word similarity [55] and word analogy [14] tests. We compare the results with the original embeddings and the three alternatives excluding Parsimax [37] since orthogonal transformations will not affect the performance of the original embeddings on these tests.

Word similarity test measures the correlation between word similarity scores obtained from human evaluation (i.e. true similarities) and from word embeddings (usually using cosine similarity). In other words, this test quantifies how well the embedding space reflects human judgements in terms of similarities between different words. The correlation scores for 13 different similarity test sets are reported in Table 3.5. We observe that, let alone a reduction in performance, the obtained scores indicate an almost uniform improvement in the correlation values for the proposed algorithm, outperforming all the alternatives in nearly all test sets. Categories from Roget's thesaurus are groupings of words that are similar in some sense which the original embedding algorithm may fail to capture. These test results signify that the semantic information injected into the algorithm by the additional cost term is significant enough to result in a measurable

Table 3.5: Correlations for Word Similarity Tests

Dataset (EN-)	GloVe	OIWE-IPG	SOV	SPINE	Proposed
WS-353-ALL	0.612	0.634	0.622	0.173	**0.657**
SIMLEX-999	0.359	0.295	0.355	0.090	**0.381**
VERB-143	0.326	0.255	0.271	0.293	**0.348**
SimVerb-3500	0.193	0.184	0.197	0.035	**0.245**
WS-353-REL	0.578	0.595	0.578	0.134	**0.619**
RW-STANFORD	0.378	0.316	0.373	0.122	**0.382**
YP-130	0.524	0.353	0.482	0.169	**0.589**
MEN-TR-3k	0.710	0.684	0.696	0.298	**0.725**
RG-65	0.768	0.736	0.732	0.338	**0.774**
MTurk-771	0.650	0.593	0.623	0.199	**0.671**
WS-353-SIM	0.682	0.713	0.702	0.220	**0.720**
MC-30	0.749	**0.799**	0.726	0.330	0.776
MTurk-287	**0.649**	0.591	0.631	0.295	0.634

improvement. It should also be noted that scores obtained by SPINE is unacceptably low on almost all tests, indicating that it has achieved its interpretability performance at the cost of losing its semantic functions.

Word analogy test is introduced in [8] and looks for the answers to questions in the form "X is to Y, what Z is to ?" by applying simple arithmetic operations to vectors of words X, Y and Z. We present precision scores for the word analogy tests in Table 3.6. It can be seen that the alternative approaches that aim to improve interpretability have poor performance on the word analogy tests. However, our proposed method has comparable performance with the original GloVe embeddings. Our method outperforms GloVe in semantic analogy test set and in overall results, while GloVe performs slightly better in syntactic test set. This comparable performance is mainly due to the cost function of our proposed method that includes the original objective of the GloVe.

To investigate the effect of the additional cost term on the performance improvement in the semantic analogy test, we present Table 3.7. In particular, we present results for the cases where i) all questions in the dataset are considered, ii) only the questions that contain at least one concept word are considered, iii) only the questions that consist entirely of concept words are considered. We note specifically that for the last case, only a subset of the questions under the semantic

Table 3.6: Precision scores for the Analogy Test

Methods	# dims	Analg. (sem)	Analg. (syn)	Total
GloVe	300	78.94	**64.12**	70.99
OIWE-IPG	300	19.99	23.44	21.84
SOV	3000	64.09	46.26	54.53
SPINE	1000	17.07	8.68	12.57
Proposed	300	**79.96**	63.52	**71.15**

Table 3.7: Precision scores for the Semantic Analogy Test

Questions Subset	# of Questions Seen	GloVe	Proposed
All	8783	78.94	**79.96**
At least one concept word	1635	67.58	**67.89**
All concept words	110	77.27	**83.64**

category `family.txt` ended up being included. We observe that for all three scenarios, our proposed algorithm results in an improvement in the precision scores. However, the greatest performance increase is seen for the last scenario, which underscores the extent to which the semantic features captured by embeddings can be improved with a reasonable selection of the lexical resource from which the concept word-groups were derived.

3.4.4 Effect of Weighting Parameter k

The results presented in the above subsections are obtained by setting the model weighting parameter k to 0.1. However, we have also experimented with different k values to find the optimal value for the evaluation tests and to determine the effects of our model parameter k to the performance. Fig. 3.4 presents the results of these tests for $k \in [0.02,\ 0.4]$. Since k adjusts the magnitude of the influence for the concept words (i.e, our additional term), average interpretability of the embeddings increases when k is increased. However, the increase in the interpretability saturates and we almost hit the diminishing returns beyond $k = 0.1$. It can also be observed that by further increasing k beyond 0.3 no additional increase in the interpretability can be obtained. This is because interpretability measurements are based on the ranking of words in the embedding dimensions.

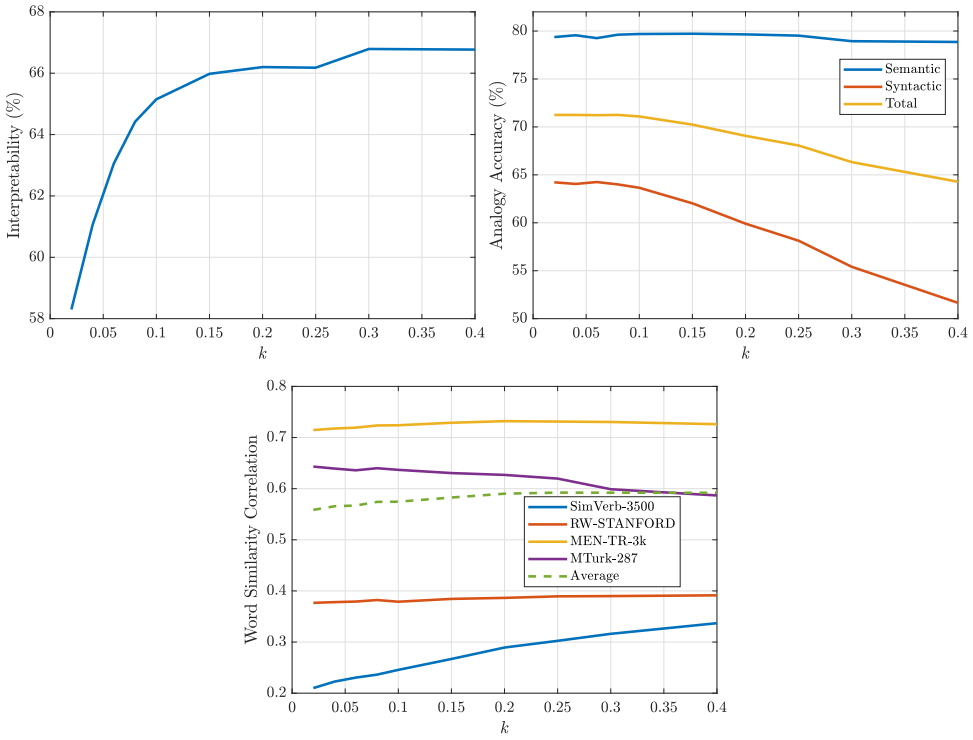

Figure 3.4: Effect of the weighting parameter k is tested using interpretability (top left, n_{min}=5, λ=5), word analogy (top right) and word similarity (bottom) tests for $k \in [0.02,\ 0.4]$.

With increasing k, concept words (from Roget's Thesarus) are strongly forced to have larger values in their corresponding dimensions. However, their ranks will not further increase significantly after they all reach to the top. In other words, a value of k between 0.1 and 0.3 is sufficient in terms of interpretability.

Will high k values harm the underlying semantic structure? To test this, our standard analogy and word similarity tests that are given in the previous subsections are deployed for a range of k values. Analogy test results show that larger k values reduce the performance of the resulting embeddings on syntactic analogy tests, while semantic analogy performance is not significantly affected. For the word similarity evaluations, we have used 13 different datasets. In Fig. 3.4 we present four of them as representatives along with the average of all 13 test sets to simplify the plot. Word similarity performance slightly increases for most of the datasets with increasing k, while performance slightly decreases (or does not

change) for the others. On average, word similarity performance increases slowly with increasing k and is less sensitive to variations in k than the interpretability and analogy tests.

Combining all these experiments and observations, empirically setting k to 0.1 is a reasonable trade-off since it significantly improves interpretability without sacrificing analogy/similarity performances.

3.5 Discussion

In this chapter, we presented a novel approach to impart interpretability into word embeddings. We achieved this by encouraging different dimensions of the vector representation to align with predefined concepts, through the addition of an additional cost term in the optimization objective of the GloVe algorithm that favors a selective increase for a pre-specified input of concept words along each dimension.

We demonstrated the efficacy of this approach by applying qualitative and quantitative evaluations for interpretability. We also showed via standard word-analogy and word-similarity tests that the semantic coherence of the original vector space is preserved, even slightly improved. We have also performed and reported quantitative comparisons with several other methods for both interpretability increase and preservation of semantic coherence. Upon inspection of Fig. 3.3 and Tables 3.5, 3.6, and 3.7 together, it should be noted that our proposed method achieves both objectives simultaneously: increased interpretability and preservation of the intrinsic semantic structure.

An important point was that, while it is expected for words that are already included in the concept word-groups to be aligned together since their dimensions are directly updated with the proposed cost term, it was also observed that words not in these groups also aligned in a meaningful manner without any direct modification to their cost function. This indicates that the cost term we added

works productively with the original cost function of GloVe to handle words that are not included in the original concept word-groups, but are semantically related to those word-groups. The underlying mechanism can be explained as follows. While the outside lexical resource we introduce contains a relatively small number of words compared to the total number of words, these words and the categories they represent have been carefully chosen and in a sense, "densely span" all the words in the language. By saying "span", we mean they cover most of the concepts and ideas in the language without leaving too many uncovered areas. With "densely" we mean all areas are covered with sufficient strength. In other words, this subset of words is able to constitute a sufficiently strong skeleton, or scaffold. Now remember that GloVe works to align or bring closer related groups of words, which will include words from the lexical source. So the joint action of aligning the words with the predefined categories (introduced by us) and aligning related words (handled by GloVe) allows words not in the lexical groups to also be aligned meaningfully. We may say that the non-included words are "pulled along" with the included words by virtue of the "strings" or "glue" that is provided by GloVe. In numbers, the desired effect is achieved by manipulating less than only 0.05% of parameters of the entire word vectors. Thus, while there is a degree of supervision coming from the external lexical resource, the rest of the vocabulary is also aligned indirectly in an unsupervised way. This may be the reason why, unlike earlier proposed approaches, our method is able to achieve increasing interpretability without destroying underlying semantic structure, and consequently without sacrificing performance in benchmark tests.

Upon inspecting the 2nd column of Table 3.4, where qualitative results for concept `TASTE` are presented, another insight regarding the learning mechanism of our proposed approach can be made. Here it seems understandable that our proposed approach, along with GloVe, brought together the words `taste` and `polish`, and then the words `Polish` and, for instance, `Warsaw` are brought together by GloVe. These examples are interesting in that they shed insight into how GloVe works and the limitations posed by polysemy. It should be underlined that the presented approach is not totally incapable of handling polysemy, but cannot do so perfectly. Since related words are being clustered, sufficiently

well-connected words that do not meaningfully belong along with others will be appropriately "pulled away" from that group by several words, against the less effective, inappropriate pull of a particular word. Even though `polish` with lowercase "p" belongs where it is, it is attracting `Warsaw` to itself through polysemy and this is not good. Perhaps because `Warsaw` is not a sufficiently well-connected word, it ends up being dragged along, although words with greater connectedness to a concept group might have better resisted such inappropriate attractions.

In this study, we used the GloVe algorithm as the underlying dense word embedding scheme to demonstrate our approach. It is possible for our approach to be extended to other word embedding algorithms which have a learning routine consisting of iterations over co-occurrence records, by making suitable adjustments in the objective function. Since word2vec model is also based on the co-occurrences of words in a sliding window through a large corpus, we expect that our approach can also be applied to word2vec after making suitable adjustments, which can be considered as an immediate future work. Although the semantic concepts are encoded in only one direction (positive) within the embedding dimensions, it might be beneficial to pursue future work that also encodes opposite concepts, such as good and bad, in opposite directions of the same dimension.

The proposed methodology can also be helpful in computational cross-lingual studies, where the similarities are explored across the vector spaces of different languages [56, 23].

Chapter 4

Advantages of Imparting Meaning

In this chapter, we investigate the imparting method for its advantages other than interpretability. We propose new word groups to combine with those extracted from Roget's Thesaurus and show that imparting meaning to embedding dimensions can greatly improve the intrinsic performance of the embedding. We also demonstrate that this approach can even be more advantageous for low resource languages. Furthermore, we show that one can concentrate gender information in a few dimensions and, then, reduce the gender bias significantly by removing them. Finally, we show that resulting embedding space from imparting method can directly be used to decompose words in terms of the categories corresponding to embedding dimensions. These decompositions can reveal the gender bias present in the word embeddings and may be useful for developing new debiasing algorithms.

The organization of the chapter is as follows. In Section 4.1, we introduce new concept groups to impart to embedding dimensions. In that section we also discuss the gender bias in the word embeddings. We present the experimental results in Section 4.2 and conclude the chapter in Section 4.3.

4.1 Methods

4.1.1 Word Groups

In Chapter 3, we extracted 300 word groups from Roget's Thesaurus [53, 54] by partitioning the tree structure starting from the root node and assigned these groups to word embedding dimensions using (3.2). It is straightforward to use the partitioning process to construct any number of word groups up to 1000.

One interesting result presented in Section 3.4.3 is the precision scores for the semantic analogy test. In particular, the significant improvement in the precision score for the questions that solely consist of words in the word groups constructed from the thesaurus is intriguing. This result suggests that analogy performance can be significantly improved by simply boosting the words from the analogy word groups in word embedding dimensions. In order to test this assertion, we combine the word groups from thesaurus with the analogy word groups. The analogy test set introduced in [8] contains 5 semantic and 9 syntactic analogy groups. Each analogy question in these groups consists of 4 different words such as "*man*", "*woman*", "*king*" and "*queen*". The test checks whether the vector of the last word, "*queen*", can be obtained by "*king*" − "*man*" + "*woman*" operation on the corresponding word vectors. For each analogy group, first and third words share a common property while second and fourth words share a different common property (i.e, male − female). Based on this observation, we divide each analogy group in the test set into two word groups resulting in 28 different word groups. Then, we construct 272 word groups from Roget's Thesaurus and combine them with these 28 analogy word groups. By combining Roget with analogy word groups, we also address a shortcoming of the Roget word groups: Roget does not contain any syntactic categories and categories of proper nouns such as countries or languages.

4.1.2 Gender Bias

It is first shown in [57] that unsupervised word embeddings often contain gender bias that is noticeable especially for occupations. In [57] it is shown that other than the gender appropriate *she-he* analogies such as *queen-king*, word embeddings also encode some gender stereotype *she-he* analogies such as *nurse-surgeon* and *volleyball-football* due to the inevitable and unwanted bias of the source corpora. In [57] a method is proposed to quantify the extent of gender bias (direct bias) present in a word embedding. Direct bias is calculated as[1]

$$\text{DirectBias}_c = \frac{1}{|N|} \sum_{w \in N} |\cos(\boldsymbol{w}, g)|^c \tag{4.1}$$

where N is the set of gender neutral words, g is the gender vector calculated as $g = w_{she} - w_{he}$ and c is the parameter that determines how strict is the bias measure. [57] suggests to take gender neutral words to be the complement of the gender specific words S such that $N = W \backslash S$ where W is set of all words. They also present a list of 218 gender specific words (S) that is obtained using the word definitions in Wordnet [7]. In our study, we take N as $P \backslash S$ where P is the set of 291 professions[2] provided in [57] instead of all vocabulary for computational simplicity.

In addition to the word groups from Roget and analogy tests, we also construct two word groups, of size 44 each, corresponding to *male* and *female* concepts by manually selecting words from S. Since the main focus with gender word groups is not increasing intrinsic performance but reducing the gender bias in the word embeddings, we do not combine these word groups with those from Roget and analogy set. Instead, we impart only two dimensions of the embedding space. By doing so, we collect the gender information in these two dimensions. Then, we remove these dimensions from the embedding space as suggested in

[1] In Equation (4.1), $|\cdot|$ corresponds to cardinality operation for N and absolute value operation for $cos()$.

[2] https://github.com/tolga-b/debiaswe/blob/master/data/professions.json
Multiword professions are filtered out from the list.

	Wikipedia	text8
VOCAB_MIN_COUNT	65	5
ALPHA	0.75	0.75
ETA	0.05	0.05
WINDOW_SIZE	15	15
VECTOR_SIZE	300	300
X_MAX	75	10

Table 4.1: Training Parameters

[58]. Moreover, we apply hard debiasing [57] to the resulting embeddings. We measure the bias levels before and after applying the debiasing operation.

4.2 Experiments and Results

We train the modified version of the GloVe algorithm given in (3.2) on English Wikipedia (measuring 11.6 GB in size which resulted in a vocabulary of size $V = 288,188$). One significant difference of these experiments from the ones presented in Chapter 3 is that we do not remove the stop words from the source corpus since some of them (*he, she ,him, her* etc.) are necessary for gender bias measurements. We also train the modified GloVe algorithm on the first 100 MB of the English Wikipedia, called text8[3], in order to investigate the performance benefits of the imparting in low resource settings. The parameters we use for training the large and small corpora are given in Table 4.1. To examine the effect of the weighting parameter, we train the embedding algorithm for various values of $k \in [0.05, 10]$. Specifically, for the experiments with 300 Roget word groups we use $k \in [0.05, 0.8]$, for the experiments with 272 Roget and 28 analogy word groups we use $k \in [0.05, 0.5]$ and finally for the experiments with two gender word groups we use $k \in [2, 10]$. For the gender bias experiments, we train the modified GloVe algorithm three times independently and use a smaller initial learning rate (ETA) of 0.004 in order to prevent exploding gradient problem we encounter due to the large weighting parameter k.

[3]http://mattmahoney.net/dc/text8.zip

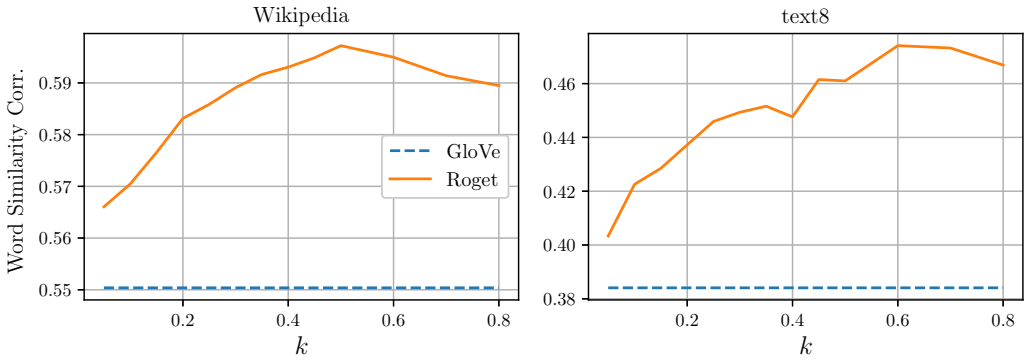

Figure 4.1: Average word similarity performance of Roget imparted GloVe for different k along with a baseline GloVe trained on Wikipedia (left) and text8 (right).

4.2.1 Intrinsic Evaluation

We evaluate the performance of the Roget imparted embeddings for different selections of the weighting parameter k on word similarity tests that measures the correlation between word similarities from human judgment and from word embeddings. There are many different datasets of different sizes for word similarity evaluations. To provide more reliable results, we use a word similarity evaluation tool from [55] that evaluates the word similarity performance on 13 different similarity datasets. In Figure 4.1 we present the average similarity scores from 13 datasets for $k \in [0.05, 0.8]$ for Wikipedia and text8 corpora, along with a baseline that is the performance of the original GloVe algorithm. For embeddings trained on the large corpus, average word similarity correlation increases with increasing k and approaches 0.60 for $k = 0.5$ (9% relative increase) and then starts to decrease. For the low resource settings, again performance increases with increasing k and achieves an impressive 23% relative improvement at $k = 0.6$.

Next, we investigate the effect of combining Roget and analogy word groups on word analogy tests [8]. As discussed in Section 4.1.1, a word analogy test checks whether the word vectors can answer questions in the form of "X is to Y as Z is to ?" by simple arithmetic operations on vectors of words X, Y and Z. Figure 4.2 presents the syntactic, semantic and overall analogy performances of

embeddings that are imparted using Roget and Roget + analogy word groups for $k \in [0.05, 0.5]$ along with a baseline from original GloVe trained on Wikipedia and text8. For the trainings on Wikipedia, imparting Roget and analogy word groups together significantly outperforms the GloVe baseline and imparting only Roget word groups in both semantic and syntactic analogy tests. It can be seen that performance of the Roget imparted GloVe decreases with increasing k while performance for the GloVe with proposed categories increases. Improvement in the accuracy with respect to the baseline is around 10% for syntactic and around 3% for semantic analogy tests for $k = 0.5$. For the trainings on text8 corpus, we can see that imparting analogy word groups along with Roget drastically improves the performance on both semantic and syntactic analogy tests. In fact, the accuracy of the Roget + Analogy imparted embeddings doubles that of the baseline on the syntactic analogy test. These results, along with the word similarity results for text corpus, show that low resource languages that do not have massive corpora to train embedding algorithms can greatly benefit from the imparting method. One only needs to construct suitable semantic and syntactic word groups for the low resource language in order to significantly boost the performance of the resulting embeddings.

One may argue that it is 'cheating' to use the words in the analogy tests in order to improve the performance on that task. This argument is understandable since we do use the words in the test set during the training process. However, the imparting method does not include any special process for the analogy task. It uses these word groups merely to impart meaning for the corresponding embedding dimensions, and indeed analogy word groups are reasonable groups corresponding to real semantic or syntactic concepts such as `countries` and `past tense`. It is noticeable that resulting embeddings' performance improve greatly on the analogy task by only forcing these word groups to have large values in different embedding dimensions.

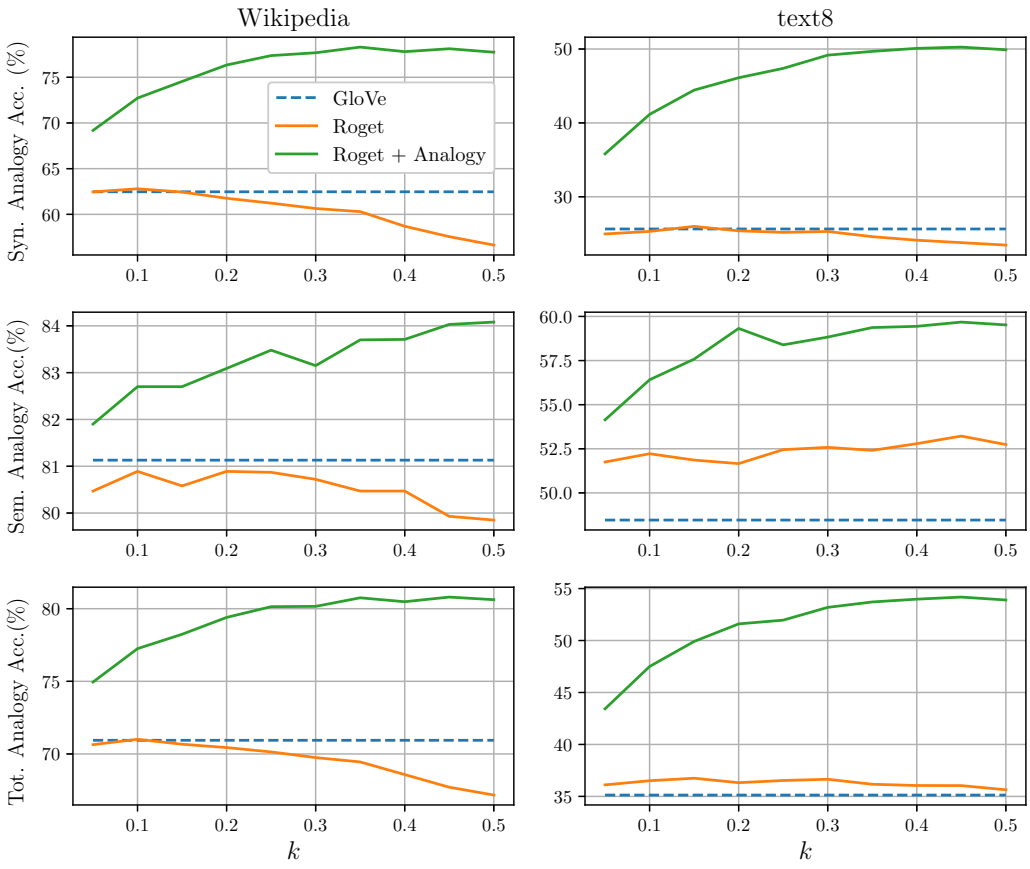

Figure 4.2: Performances of original, Roget imparted and Roget + Analogy imparted GloVe embeddings trained on English Wikipedia (left column) and trained on text8 (right column) on syntactic (top) and semantic (middle) analogy test along with the total performance (bottom).

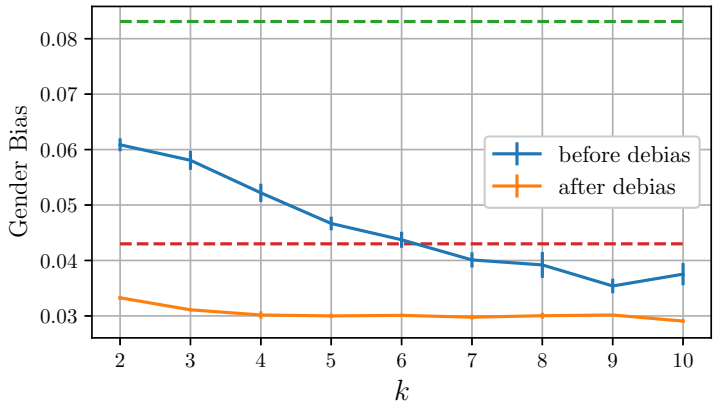

Figure 4.3: Average gender bias in the reduced embedding spaces for $k \in [2, 10]$ before and after applying hard debiasing. Error bars represent the standard deviation of the results from three independent training of the algorithm. Green and red dashed lines correspond to the gender bias levels of the embeddings from original GloVe algorithm before and after debiasing, respectively.

4.2.2 Gender Bias

After training the embeddings with two gender imparted dimensions, we have removed these dimensions from the embedding space to obtain a reduced, gender-free embedding space and measured the gender bias of the resulting space using Equation (4.1) with $c = 1$. Then, we applied hard debaising to the reduced space and measured gender bias again. Figure 4.3 displays average gender bias in the reduced embedding spaces for $k \in [2, 10]$ before and after applying the hard debiasing method from [57]. Green and red dashed lines in Figure 4.3 correspond to the gender bias levels of the embeddings from original GloVe algorithm before and after debiasing, respectively. It can be seen that removing the gender imparted dimensions from the embedding space significantly reduces gender bias. For $k > 6$, we can see that solely removing gender dimensions reduces the gender bias to a lower level than applying hard debiasing to original embeddings (red dashed line). Applying hard debiasing to the reduced embedding space decreases the gender bias to even a lower level. These results indicate that imparting method can be used to obtain a more gender-neutral embedding space.

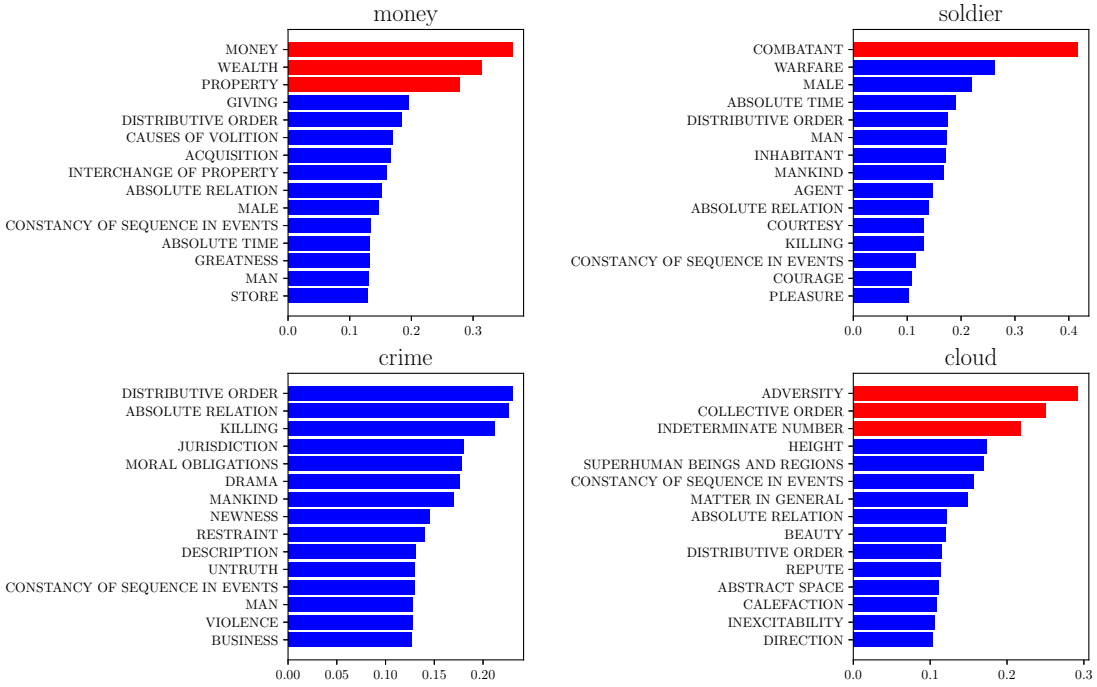

Figure 4.4: Decompositions of words *money*, *soldier*, *crime* and *cloud* in terms of Roget categories. Red bars correspond to categories that contain the decomposed word, while blue bars correspond to categories that do not contain it.

4.2.3 Semantic Decomposition of Words

In addition to the discussed above advantages of imparting meaning to the word embedding dimensions, we also discuss a simple but useful result of having meaning-imparted dimensions or having an interpretable embedding space in general. As presented in Section 2.2.2.2, interpretable embedding spaces provide meaningful decompositions for the words in terms of the concepts represented by the dimensions. Figure 4.4 displays 15 categories from Roget ($k = 0.1$) corresponding to dimensions with the highest values for the words *money*, *soldier*, *crime* and *cloud*. The red bars represent the categories that contain the decomposed word, while blue bars represent the categories that do not contain it.

One interesting result in Figure 4.4 is that `man` category is among the top ranks

in the decompositions of words *money*, *crime* and *soldier* and `male` category is also among the top ranks for words *money* and *soldier*. (Here, `male` category from Roget should not be confused with the "male" gender category we manually constructed for imparting gender.) In other words, these word decompositions also reveal the gender bias present in the word embeddings and may be instrumental in developing better debiasing algorithms.

Another observation we make regarding word decompositions is that the `Distributive Order` category is among the top ranks for most of the words we investigated. This is because of the fact that `Distributive Order` category contains significantly more frequent words than the other categories and most of the words we investigate are also considerably frequent with respect to most of the words in the vocabulary. Therefore, this category focuses more on the frequency of the word instead of its actual meaning.

4.3 Discussion

In this chapter, we have demonstrated several advantages of imparting meaning to word embedding dimensions, a method we proposed primarily for increasing interpretability. First, we showed that imparting meaning to embedding dimensions has potential to significantly improve the intrinsic performance of the embeddings by proper selection of the weighting parameter and word groups. We introduced 28 new word groups from word analogy test sets and combined them with Roget groups and demonstrated that by simply boosting the words from the analogy groups, performance of the word embedding can be drastically increased for both semantic and syntactic analogy tasks. Furthermore, we performed experiments with a considerably small corpus and showed that performance improvements are more dramatic under low resource settings indicating that imparting method can be even more beneficial for languages that lack large corpora to train the embedding algorithm.

Second, we focused on the gender bias problem of the word embeddings. We

introduced two new word groups for the two genders and showed that by boosting these groups in embedding dimensions, we can capture most of the gender information in only two dimensions, which is beneficial for debiasing. As an alternative, we demonstrate that simply removing these gender dimensions from the embedding space also reduces the gender bias and creates a genderless space.

Finally, we demonstrated that the embedding space trained with the imparting method can be used to decompose words in terms of categories that are imparted to embedding dimensions.

Chapter 5

Conclusion

In this thesis, we focused on interpretability of word embeddings, which are unsupervised methods to represent words as vectors in dense continuous vector spaces. Word embeddings have become increasingly popular among the NLP researchers in recent years and have been extensively employed to map the semantic properties of words to vectors thanks to the state of the art performance they provide in many NLP tasks. It is repeatedly shown that word embeddings are considerably successful in capturing semantic relations between words. Therefore, a meaningful semantic structure must be present in the respective vector spaces. However, in many cases, this semantic structure is broadly and heterogeneously distributed across the embedding dimensions. In other words, vectors corresponding to the words are only meaningful relative to each other. Neither the vector nor its dimensions have any absolute meaning, making interpretation of dimensions a big challenge.

We proposed a statistical method to uncover the underlying latent semantic structure of dense word embeddings. Our proposed method leverages the category theory and is based on a new dataset (SEMCAT) we introduced. SEMCAT contains more than 6,500 words semantically grouped under 110 categories. Utilizing the semantic categories in SEMCAT, we provided a semantic decomposition of the word embedding dimensions and verified our findings using qualitative and

quantitative tests.

We also addressed the problem of measuring the interpretability of word embeddings. In previous studies, interpretability is commonly measured using word intrusion test which requires significant human effort. In word intrusion test each embedding dimension is evaluated by several human evaluators independently making the test laborious and difficult to reproduce due to its subjective nature. We proposed a new automated evaluation method based on the categories in SEMCAT. We also proposed another alternative method which takes subcategories into account during evaluations. Our methods are automated (hence, effort-free), yet they reflect human judgement since categories are manually constructed. Therefore, they have potential to replace the word intrusion test for interpretability evaluations.

Although our detailed analyses showed that SEMCAT can provide decent evaluations, extended datasets with improved coverage and expert labeling by multiple observers would further improve the reliability of the proposed approach. To do this, a synergistic merge with existing lexical databases such as WordNet might prove useful. It is straightforward to extend our method to other languages by constructing category datasets. In this thesis, in addition to English, we applied our method to Turkish. For this purpose, we introduced a new category dataset (ANKAT) and measured interpretability of Turkish word embeddings.

Then, we focused on improving the interpretability of the word embeddings without reducing their expressive performance. We proposed a novel approach to impart interpretability into word embeddings by encouraging different dimensions of the vector representations to align with predefined concepts. We combined the cost function of the GloVe embedding algorithm with a new cost function that favors a selective increase for a pre-specified input of concept words along each dimension. We used Roget's Thesaurus to extract 300 concept groups where each group is aligned with one of the 300 dimensions of the embedding.

We demonstrated the effectiveness of our approach by applying qualitative and

quantitative evaluations for interpretability. We also tested the expressive performance of the resulting embeddings via standard word-analogy and word-similarity tests. We performed quantitative comparisons with several other methods from the literature that aim to improve embedding interpretability using different approaches. Evaluation results showed that our proposed method is significantly more interpretable than the alternatives. Moreover, our method preserves the semantic structure and performance of the original embedding space, while the alternative methods (except parsimax) perform poorly on these tests.

An important observation we made by manually inspecting the dimensions of our proposed embedding was that, while it is expected for words that are already included in the concept word-groups to be aligned together due to the direct effect of the additional cost term, words that are not in these groups also aligned in a meaningful manner without any direct modification to their cost function. This implies that the additional cost term works productively with the original cost function of GloVe to handle words that are not included in the original concept word-groups, but are semantically related to those word-groups.

In this thesis, GloVe is used as the underlying dense word embedding algorithm to demonstrate our approach. However, our imparting approach can be extended to other word embedding algorithms that iterates over cooccurrence records during training, by making suitable adjustments in the objective function. One such model is the popular word2vec model and we expect that our approach can also be applied to word2vec, which can be considered as an immediate future work for our approach. Another possible extension of our proposed method is encoding concepts in the negative direction as well. Using both directions of each dimension for imparting will provide better utilization of the embedding space and allow for encoding of more semantic concepts. Moreover, it might be beneficial to encode opposite concepts, such as good and bad, in opposite directions of the same dimension. This way a dimension can represent a single concept such as goodness in a continuity.

In addition to improving interpretability, we also demonstrated several advantages of imparting meaning to word embedding dimensions. We showed that

imparting meaning to embedding dimensions has potential to significantly improve the performance of the embeddings on different tests by proper selection of the word groups. We introduced 28 new word groups from word analogy test sets and combined them with Roget groups. We demonstrated that by simply boosting the words from the analogy groups, performance of the word embedding can be drastically increased for both semantic and syntactic analogy tasks. By conducting experiments on a considerably small corpus, we showed that performance improvements are even more dramatic under low resource settings which indicates that imparting method can be even more beneficial for low-resource languages.

We also focused on the gender bias problem of the word embeddings which has drawn increasing attention in the last years. We introduced two new word groups for the two genders and showed that by boosting these groups in embedding dimensions, we can capture most of the gender information in only two dimensions. Our experimental results demonstrated that combining gender information in a few dimensions is useful for debiasing. As an alternative, we simply removed these gender dimensions from the embedding space and showed that the reduced embedding space contains less gender bias. Following a similar approach, one can concentrate any concept or information in specific embedding dimensions using our method and then remove these dimensions to eliminate the unwanted information.

Bibliography

[1] B. MacWhinney, "Language evolution and human development," in *Origins of the social mind: Evolutionary psychology and child development*, pp. 383–410, New York: Guilford Press, 2005.

[2] A. M. Turing, "Computing machinery and intelligence," in *Parsing the Turing Test*, pp. 23–65, Springer, 2009.

[3] F. De Saussure, *Course in general linguistics*. Columbia University Press, 2011.

[4] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of physiology*, vol. 117, no. 4, pp. 500–544, 1952.

[5] T. Winograd, "Understanding natural language," *Cognitive psychology*, vol. 3, no. 1, pp. 1–191, 1972.

[6] W. A. Woods, "Transition network grammars for natural language analysis," *Communications of the ACM*, vol. 13, no. 10, pp. 591–606, 1970.

[7] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[9] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "Joint learning of words and meaning representations for open-text semantic parsing," in *Artificial Intelligence and Statistics*, pp. 127–135, 2012.

[10] Z. S. Harris, "Distributional structure," *WORD*, vol. 10, no. 2-3, pp. 146–162, 1954.

[11] J. Firth, *Papers in linguistics, 1934-1951*. Oxford University Press, 1957.

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J Am Soc Inf Sci*, vol. 41, no. 6, p. 391, 1990.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J Mach Learn Res*, vol. 3, no. Jan, pp. 993–1022, 2003.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), pp. 3111–3119, Curran Associates, Inc., 2013.

[15] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[16] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[17] C.-C. Lin, W. Ammar, C. Dyer, and L. Levin, "Unsupervised pos induction with word embeddings," *arXiv preprint arXiv:1503.06760*, 2015.

[18] D. Chen and C. Manning, "A fast and accurate dependency parser using neural networks," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 740–750, Association for Computational Linguistics, October 2014.

[19] S. K. Sienčnik, "Adapting word2vec to named entity recognition," in *NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, no. 109, pp. 239–243, Linköping University Electronic Press, 2015.

[20] I. Iacobacci, M. T. Pilehvar, and R. Navigli, "Embeddings for word sense disambiguation: An evaluation study.," in *ACL (1)*, 2016.

[21] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-supervised recursive autoencoders for predicting sentiment distributions," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 151–161, Association for Computational Linguistics, 2011.

[22] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings for sentiment analysis," in *EMNLP*, pp. 545–550, 2017.

[23] L. K. Senel, V. Yücesoy, A. Koç, and T. Çukur, "Measuring cross-lingual semantic similarity across european languages," in *TSP*, 2017.

[24] Y. Goldberg and G. Hirst, *Neural Network Methods in Natural Language Processing*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2017.

[25] L. D. Vine, M. Kholghi, G. Zuccon, L. Sitbon, and A. Nguyen, "Analysis of word embeddings and sequence features for clinical information extraction," in *Proceedings of the Australasian Language Technology Association Workshop 2015*, (Parramatta, Australia), pp. 21–30, December 2015.

[26] A. Joshi, V. Tripathi, K. Patel, P. Bhattacharyya, and M. Carman, "Are word embedding-based features useful for sarcasm detection?," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1006–1011, Association for Computational Linguistics, 2016.

[27] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in neural information processing systems*, pp. 2172–2180, 2016.

[28] O. Levy and Y. Goldberg, "Dependency-based word embeddings.," in *ACL (2)*, pp. 302–308, 2014.

[29] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[30] B. Murphy, P. P. Talukdar, and T. M. Mitchell, "Learning effective and interpretable semantic models using non-negative sparse embedding," in *COLING*, pp. 1933–1950, Indian Institute of Technology Bombay, 2012.

[31] H. Luo, Z. Liu, H.-B. Luan, and M. Sun, "Online learning of interpretable word embeddings.," in *EMNLP*, pp. 1687–1692, 2015.

[32] A. Fyshe, P. P. Talukdar, B. Murphy, and T. M. Mitchell, "Interpretable semantic vectors from a joint model of brain-and text-based meaning," in *ACL*, vol. 2014, p. 489, NIH Public Access, 2014.

[33] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski, "Linear algebraic structure of word senses, with applications to polysemy," *Transactions of the Association of Computational Linguistics*, vol. 6, pp. 483–495, 2018.

[34] M. Faruqui, Y. Tsvetkov, D. Yogatama, C. Dyer, and N. A. Smith, "Sparse overcomplete word vector representations," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1491–1500, Association for Computational Linguistics, 2015.

[35] A. Subramanian, D. Pruthi, H. Jhamtani, T. Berg-Kirkpatrick, and E. Hovy, "Spine: Sparse interpretable neural embeddings," *Proceedings of the Thirty Second AAAI Conference on Artificial Intelligence (AAAI)*.

[36] A. Zobnin, "Rotations and interpretability of word embeddings: the case of the russian language," in *International Conference on Analysis of Images, Social Networks and Texts*, pp. 116–128, Springer, 2017.

[37] S. Park, J. Bak, and A. Oh, "Rotated word vector representations and their interpretability," in *EMNLP*, pp. 401–411, 2017.

[38] G. Murphy, *The big book of concepts*. MIT press, 2004.

[39] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *NIPS*, pp. 288–296, 2009.

[40] K.-R. Jang and S.-H. Myaeng, "Elucidating conceptual properties from word embeddings," *SENSE 2017*, p. 91, 2017.

[41] I. Vulić, D. Gerz, D. Kiela, F. Hill, and A. Korhonen, "Hyperlex: A large-scale evaluation of graded lexical entailment," *arXiv preprint arXiv:1608.02117*, 2016.

[42] A. Gladkova, A. Drozd, and C. Center, "Intrinsic evaluations of word embeddings: What can we do better?," *ACL 2016*, p. 36, 2016.

[43] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

[44] M. Yu and M. Dredze, "Improving lexical embeddings with semantic knowledge," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (Baltimore, Maryland), pp. 545–550, Association for Computational Linguistics, June 2014.

[45] C. Xu, Y. Bai, J. Bian, B. Gao, G. Wang, X. Liu, and T.-Y. Liu, "Rc-net: A general framework for incorporating knowledge into word representations," November 2014.

[46] Q. Liu, H. Jiang, S. Wei, Z.-H. Ling, and Y. Hu, "Learning semantic word embeddings based on ordinal knowledge constraints," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Beijing, China), pp. 1501–1511, Association for Computational Linguistics, July 2015.

[47] D. Bollegala, M. Alsuhaibani, T. Maehara, and K. Kawarabayashi, "Joint word representation learning using a corpus and a semantic lexicon," *CoRR*, vol. abs/1511.06438, 2015.

[48] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith, "Retrofitting word vectors to semantic lexicons," in *Proc. of NAACL*, 2015.

[49] S. K. Jauhar, C. Dyer, and E. Hovy, "Ontologically grounded multi-sense representation learning for semantic vector space models," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Denver, Colorado), pp. 683–693, Association for Computational Linguistics, May–June 2015.

[50] R. Johansson and L. Nieto Piña, "Embedding a semantic network in a word space," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Denver, Colorado), pp. 1428–1433, Association for Computational Linguistics, May–June 2015.

[51] N. Mrkšić, D. Ó Séaghdha, B. Thomson, M. Gašić, L. M. Rojas-Barahona, P.-H. Su, D. Vandyke, T.-H. Wen, and S. Young, "Counter-fitting word vectors to linguistic constraints," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 142–148, Association for Computational Linguistics, June 2016.

[52] N. Mrkšić, I. Vulić, D. Ó Séaghdha, I. Leviant, R. Reichart, M. Gašić, A. Korhonen, and S. Young, "Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 309–324, 2017.

[53] P. M. Roget, *Roget's Thesaurus of English Words and Phrases*. TY Crowell Company, 1911.

[54] P. M. Roget, *Roget'S International Thesaurus, 3/E*. Oxford and IBH Publishing, 2008.

[55] M. Faruqui and C. Dyer, "Community evaluation and exchange of word vectors at wordvectors.org," in *Proceedings of ACL: System Demonstrations*, 2014.

[56] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *arXiv preprint arXiv:1309.4168*, 2013.

[57] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, (USA), pp. 4356–4364, Curran Associates Inc., 2016.

[58] P. Dufter and H. Schütze, "Analytical methods for interpretable ultradense word embeddings," *arXiv preprint arXiv:1904.08654*, 2019.