Incorporating Background Knowledge into Text Classification

Reihane Boghrati, Justin Garten, Aleksandra Litvinova, Morteza Dehghani
boghrati, jgarten, alitvino, mdehghan@usc.edu
University of Southern California
Los Angeles, CA 90089 USA

Abstract

It has been shown that prior knowledge and information are organized according to categories, and that also background knowledge plays an important role in classification. The purpose of this study is first, to investigate the relationship between background knowledge and text classification, and second, to incorporate this relationship in a computational model. Our behavioral results demonstrate that participants with access to background knowledge (experts), overall performed significantly better than those without access to this knowledge (novices). More importantly, we show that experts rely more on relational features than surface features, an aspect that bag-of-words methods fail to capture. We then propose a computational model for text classification which incorporates background knowledge. This model is built upon vector-based representation methods and achieves significantly more accurate results over other models that were tested.

Keywords: text classification; background knowledge; distributed representation; similarity

Introduction

The fcat taht radeers are slitl albe to uansdrnetd tihs txet, aoughlth it is is far form benig galarlmticmy cecorrt, iullstartes how peoicerptn, cetagotizroain and unedranstding is ienfueldcd by piror kowlndgee. Previous research on the organization of knowledge in the human mind has proposed that knowledge is saved in form of concepts and organized according to categories (Smith, 1995). However, how certain categories are formed, and according to which criteria humans place objects into categories, has been a challenge for cognitive science. The first intuitive approach that comes to mind, categorizing objects according to their superordinate definition, leads to a huge bag of miscellaneous words, since it is for example not so easy to define what makes a bird a bird. Some birds can fly, others have wings, but cannot fly (e.g. penguins) and some animals can fly but are not birds (e.g. bats). The underlying difficulty is that not all members of a category share the same features. Although, not all members share the same features, Wittgenstein (1953) noted that members of a category still resemble each other in some way, which led to the emergence of the prototype approach to categorization. Rosch (1973) proposed that membership of a category is defined by the comparison of the object to the prototype of the category, where the prototype represents a blend of the most common category members. According to Rosch (1973), an object that closely resembles the prototype image of a category will be more likely to be classified according to that category than an object that has only little resemblance. However, this theory cannot explain why a Pomeranian dog, that actually has more resemblance with a cat than a dog, is nevertheless categorized as a dog. Furthermore, a typical representation of a category strongly depends on context. For humans living in warmer areas, a robin might be a typical member of the category 'birds', whereas for Eskimos a penguin might be a more typical member of the same category. Besides studying the way certain objects are assigned to categories, investigating how categories are organized yields relevant information about the structural organization of knowledge in the human mind. An experiment conducted by Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976), asked subjects to list features common to most objects from the categories 'furniture', 'table' and 'kitchen table', in order to investigate whether a certain level of a category is more prevalent than another level. On average participants named 3 features from the global level 'furniture', 9 features from the basic level 'table' and 10.3 features from the specific level 'kitchen table'. Based on those results Rosch et al. (1976) argued that the basic category level is, from a psychological perspective, the most informative level, since the global level provides relatively less information (3 vs. 9) and the specific level only marginally more information than the basic level. More recent approaches to knowledge categorization focus on the relationships between concepts and categories. Rottman, Gentner and Goldwater (2012) examined the classification differences between novices and experts in the physical sciences. In their experiment, students were asked to sort descriptions of real-world phenomena varying in causal structures (e.g. common cause vs. causal chain) and in content domain (engineering vs. biology). Their results showed that novices in physical sciences sorted descriptions based on the content domain, whereas experts sorted those descriptions according to their causal structure, thereby emphasizing the importance of causal relationships in knowledge organization. Moreover, in a series of studies, Bang, Medin and Atran (2007) demonstrated the role of culture and experience in categorization-based reasoning, essentially arguing that "what people think about can affect how they think".

Given the vast amount of available data and increasing computational power, our study aims to further investigate the principles of human categorization of text, in order to inform machine learning methods in the domain of natural language processing. More precisely we are interested in the different text classification patterns between novices and experts. Based on previous research by Rottman et al. (2012), we hypothesize that novices categorize similar text according to surface features, whereas experts classify similar text according to deeper relational features. In other words, we assume that the background knowledge of experts allows them to take into account more relational features for the classification of similar texts, whereas novices are forced to rely on surface

features of the content.

We begin by discussing our behavioral experiment which investigates the role of background knowledge in text classification. Next, we summarize recent developments in distributed representations of text. Then we describe our computational model and our second experiment. Finally, we discuss the shortcomings of the model and future work.

Experiment 1

The goal of the first study is to explore the role of background knowledge in text classification with a behavioral experiment. The result of this experiment will guide our computational modeling work in the next section. In both experiments, we are interested in the classification of movie reviews.

In this study, we examine how a group of participants with no background knowledge about a set of movies (novices) differ from those who have access to more relational knowledge about the movies (experts). Based on previous findings (Rottman et al., 2012), our assumption is that the existence of background knowledge would allow experts to perform more accurate classifications, as they base their classification more on relational features compared to novices, who might only be able to categorize according to surface differences. In other words, we expect that access to background knowledge would result in classification based on relational features, and incorporating this finding into computational models would increase classification performance.

Method We designed a simple task in which participants were asked to decide whether a set of movie reviews belonged to the same movie or not. In order to make sure none of the participants had seen the movies they were being tested on, we chose a set of foreign language movies and reviews not belonging to mainstream blockbusters. Participants were divided into two groups: novices and experts. Prior to performing the classification task, participants in the expert condition, read full-length articles containing the storyline, plot and highlights of each movie. Further, after reading each article, they were asked a few questions about the article to make sure that they had actually read the article. An example of the articles we used is given in figure 1.

Next, for each question, participants were asked to select all of the reviews that they thought belong to the same movie. Overall, participants were tested on reviews from four movies. Each question, had exactly two out of three options that matched the same movie. A sample of the classification tasks is shown in figure 2.

We systematically varied the type of similarity between the movie reviews, where the reviews either matched in surface similarity, in structural similarity, or in both. Apparent features regarding a movie such as the cast members or filming locations were considered as surface features, while details such as the relation between the actors or inferences made from the plot were categorized as relational features.

In the group where both surface and relational features were different, although the reviews might describe the same movie but because they point to different aspects of the movie, and use different words to describe it, the features tend to be different. The second category was the opposite: both the surface features and relational features of the reviews were the same. Although the reviews might have described different movies, but they used a number of similar and shared words and relations. The other two categories had either surface features in common or relational features.

> An army colonel and his new wife are coming to visit their relatives, who live in a small apartment complex. Mom wants to make sure that they get the best treatment possible, and arranges for a big feast. The mother is stressing because she is poorer and wants to impress the colonel. The father is a cinema projectionist and tries to create fun for the guests. They have little food for a banquet, which stresses mom even more. Her little son goes to steal food but the shop owner finds out and kick he and his friend out, later on he feels sympathetic and brings the food for mom. They live around a bunch of neighbors including an old lady with chickens, a pharmacology student studying for an exam the next day, a couple that argues, and some other people. There's a lot of yelling and chaos, but everybody will come together and help mom to cook for dinner, and somehow, everything works out. At the end colonel and his new wife want to leave, and mom secretly wants them to leave, but every time they try to leave, everybody asks them to stay, just because it is the polite thing to do. Accidentally mom doesnt feel good, and they take her to hospital. When they come back, they each go to their room, turn off the lights and sleep.

Figure 1: Example of one of the training articles read by participants in the expert condition

Participants 152 participants located in the US were recruited from Amazon Mechanical Turk. 76 participants were randomly assigned in the novice condition and 76 in the expert condition. After making sure that none of the subjects are familiar with the movies, participants in the expert condition first had to read a summary about each movie before completing the survey, whereas participants in the novice condition were immediately directed to the survey, without receiving background knowledge on the movies.

Results Overall, participants in the expert condition were able to make significantly more correct classifications than participants in the novice condition $t(142)=3.44$, $p=0.0008$. In other words, experts made fewer errors in classifying movie reviews.

Specifically, experts answered those questions which had similar surface features for all three reviews significantly better than novices $t(133)=3.13$, $p=0.002$. This observation suggests that novices, who rely on surface features can easily be distracted by common words shared among the reviews. On the other hand, experts who look for deeper features and do not rely only on surface features, were more successful in picking the reviews which belonged to the same movie.

Experts, however, did not essentially do better on the questions where relational features were shared among reviews $t(148)=0.83$, $p=0.4$. This result shows that when reviews have

Question: From the reviews below, please select all the reviews which you think are about the same movie. (It can be two, three, or none of them) (Different relational features/Similar surface features)
1- (LEILA) The sound and the visuals aren't groundbreaking, but it gets the job done. There are occasional funny parts stuck in there (especially with the main role's uncle). The movie gives one a good glimpse of upper middle-class society in Iran.
2- (MUM's GUEST) In this movie, the director has shown an Iranian little society with its all humors. You could find in this movie, one social stratum of Iranian people, all have their own problems, and how they live together.
3- (MUM's GUEST) This movie is both a social comedy and a love letter to cinema. Mum's husband is a cinema projectionist in Iran who, together with his colleagues in a memorable scene, recite music and dialog from classic films.

Figure 2: Example of one of the questions answered by participants in the two groups

common relational features, experts, who are looking for relational features, are distracted with the similarity of the features and cannot predict accurately.

Discussion Comparing the two groups, our results indicate that overall people who had read articles about the movies, performed significantly better than our novice group. More importantly, analyzing based on type of similarity revealed that this higher performance was due to the ability to categorize based on relational features, and not due common words and shared surface features. In the questions in which all three reviews had similar relational features, there was no significant difference between experts and novices.

Our finding demonstrates how access to some textual knowledge can affect classification in subsequent tasks. Moreover, it shows that simply relying on surface features cannot help us distinguish between items which have relational commonalities, but do not share the same words. In other words, this experiment provides an explanation why simple bag-of-words approaches to text classification may not only fail to capture human approaches to simple text classification tasks, but also how poorly they would perform when obvious relational features exist between the groups.

Distributed Representations

Representation of conceptual knowledge has been a key challenge for the development of cognitive models. One major approach to this issue has come in the form of distributed representations, where words or concepts are represented in the form of n-dimensional vectors. This approach has been used extensively in connectionist models starting with Parallel Distributed Processing (McClelland, Rumelhart, Group, et al., 1986) where distributed representations fit naturally as corresponding to the weights of nodes in the neural networks (whether as inputs, outputs, or in hidden layers).

In part driven by the resurgence of neural networks in recent years, distributed representations have seen widespread adoption with applications across the fields of natural language processing (Bengio, Courville, & Vincent, 2013; Mikolov, 2012; Socher, Bauer, Manning, & Ng, 2013) and cognitive modeling (Serre, Oliva, & Poggio, 2007).

In this process, a number of approaches, new and old, have been explored for the generation of these representations. On the neural network side, modern algorithmic improvements (Krizhevsky, Sutskever, & Hinton, 2012) have been combined with a range of training approaches in systems such as Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013). Approaches based on building and then reducing the dimensionality of large co-occurrence matrices such as Latent Semantic Analysis (LSA) (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990) have received renewed attention. And techniques from topic modeling such as Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) have been explored for the creation of distributed representations.

As techniques for the generation of individual word representations have matured, focus has increasingly shifted towards composing these representations to capture the meaning of larger pieces of text. This has proved particularly important in application areas such as sentiment analysis (Pang & Lee, 2008) where handling issues like negation is critical. A number of approaches have been explored to compositionality including additive compositionality (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) recursive deep networks (Socher, Perelygin, et al., 2013), and matrix-vector representations (Socher, Huval, Manning, & Ng, 2012).

We focus here on a particular line of work which combines word and context information through the usage of distributional representations for context. In particular, we look at the Paragraph Vector (Le & Mikolov, 2014) method which simultaneously learns representations for words and larger textual contexts (generically: "paragraphs"). Words are represented as columns in a matrix W and paragraphs as columns in a matrix D. Given a sequence of words, the model either averages or concatenates the previous window of words with the local paragraph vector (Figure 3a). The resulting vector is used as the input to a hierarchical softmax classifier (Morin & Bengio, 2005) which predicts the next word in the sequence. The paragraph and word vectors are trained with stochastic gradient descent using backpropagation (Rumelhart, Hinton, & Williams, 1986).

A variant of this model was released (Mikolov, 2014) which combined the use of context vectors with the base code for the Word2Vec program. This allowed the usage of the Skip-gram model for learning the word representations and the replacement of the hierarchical softmax with negative sampling (Goldberg & Levy, 2014). These changes led to slight overall improvements in system performance.

Experiment 2

Computational text classification has been widely studied from both semantical and syntactical aspects. In some classification settings, instead of using a training dataset, the models rely on predefined set of features or words (Sagi & De-

hghani, 2014). Even though, this might be considered a first step towards incorporating background knowledge into text classification, here we take a step further. As demonstrated in the first experiment, background knowledge can have a major role in classification. The goal of our modeling effort is to investigate how background knowledge can get incorporated into vector-based models of words representation, and to investigate whether or not incorporating such knowledge can result in more accurate classifications.

Our Model As discussed previously, Word2Vec and paragraph vector algorithms both use neural networks to train vector representation of words and documents, treating all inputs (words and documents) similarly. In order to examine our hypothesis, we designed a modified version of these systems, allowing them to integrate background knowledge, and as a result demonstrate improvements on the classification task. We extend the Mikolov (2014) variation of the Paragraph Vector approach by adding an additional input vector representing background knowledge. Background knowledge vectors are stored as columns in a new matrix B (Figure 3c), similar to the matrices D and W for document and word representations. Matrix B and D are similar in the way they are represented, with the difference being that, unlike matrix D, matrix B is static and does not change throughout the training process. In other words, matrix B is present in the training of document and words vectors and it influences their vector representation, without being affected itself in this stage. Matrix B can be thought of as a filter (or biasing lens), through which new information gets interpreted based on background knowledge.

Method To show the effectiveness of this method, and investigate whether it can model our behavioral results, we examined the performance of two different baseline word2vec models and compared their results to our modified version discussed above. We used 929 movie reviews from 30 different movies which were collected from Stanford Treebank corpus (Maas et al., 2011) as data for this experiment.

The experiment was designed to be similar to our behavioral study. Each question had three reviews and the task for the model was to predict which of the three reviews are about the same movie. For each question, two of the reviews were selected from one movie, and one review was selected from another movie. Similar to the behavioral experiment, the model had the option to pick two of them, all or none of them.

For the baseline models we used paragraph vector method to generate the vector representation of the reviews. In the first baseline setting (Figure 3a), only the reviews themselves were used for training. In the second setting (Figure 3b), we used the reviews in addition to full-length articles about each movie for training the vectors. For this purpose, the text of the articles were concatenated to the reviews, and fed to paragraph vector method. In both of these settings, the models had to predict which of three reviews were related to the same movie. For the first setting, this was achieved by only calculating the cosine similarity of the reviews against one another, and if the similarity score was above a threshold, then the model categorized them as belonging to the same movie. In the second setting, it also needed to determine the most similar article to each of the movie reviews (based on the cosine similarity of their vector representations). This setting predicted whether two reviews are about the same movie, based on both their individual cosine similarity to one another and the movies they were mapped to.

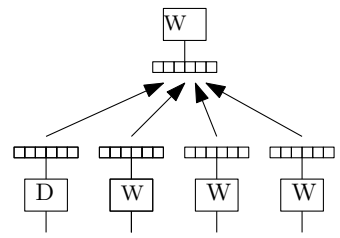
For our model, we generated the vector representation of the articles separately using paragraph vector method. These article vectors were used as matrix B in our model to provide background knowledge. We then ran our model by using the corpus of reviews along with the fixed article vectors. Similar to the second setting of the baseline model, the most similar movie to a review was determined by calculating the cosine similarity of the review vectors with the article vectors. The model made a decision about whether two reviews belong to the same movie based on the cosine similarity of the review vectors and the movies they were mapped to.

Result Table 1 shows the results of our experiment. Accuracy was measured by calculating how many times the model made the correct classification, i.e. it correctly predicted that the two first reviews were about the same movie and were different from the third review. As shown in this table, classifying text with no background knowledge (baseline, first setting) reached to an accuracy of 32%. Adding the text of the articles to the reviews (baseline, second setting) increased the performance to 35%. Our model, achieved an accuracy of 41%, which is significantly higher than the second model($X^2 = 399, p < 0.001$) and also the first baseline ($X^2 = 343, p < 0.001$).

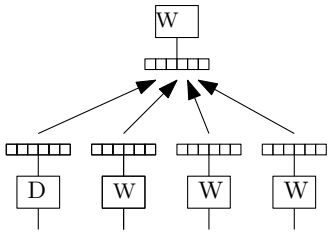
Method	Chance	Baseline1	Baselin2	Proposed Model
Accuracy	20%	31%	35%	41%

Table 1: Computational Results

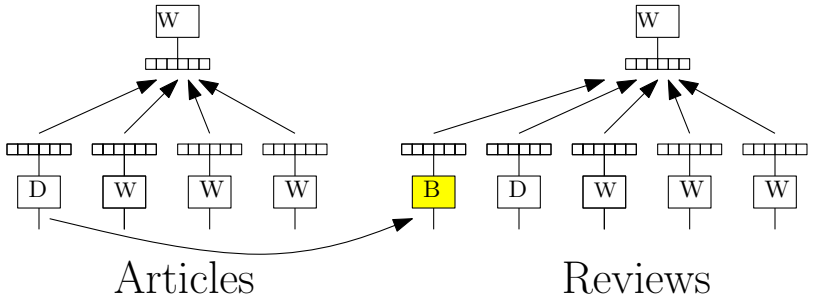
Discussion The implications of this experiment are two fold: 1. Our results demonstrate that background knowledge can significantly improve text classification 2. Simple bag-of-words techniques for incorporating knowledge may not work as well, and may lack cognitive plausibility. Even though our model has access to the same amount of information (text) as the baseline model in the second setting, it significantly outperformed it. Specifically, we argue that background knowledge is treated differently than regular words used in a document, and should be used as an interpretive lens, rather than similar to other documents. In our model the effect of background knowledge was fixed and present during the whole document vector training.



(a) Baseline, first setting framework.



(b) Baseline, second setting framework.



(c) Proposed model.

Figure 3: The three different evaluated vector models. 3.a: Paragraph vector model which receives the reviews as input and trains the vector representation of each document. 3.b: A revised version of 3.a where the input to the neural network are the reviews which are concatenated with their related movie articles. 3.c: Our proposed model, where the vector representation of each article is calculated (left side of 3.c) and it is then provided to the right-side framework. Matrix B, which is a representative of background knowledge, is fixed during the training process and is used to a biasing factor for the vector representation of the reviews.

Conclusion

Using two experiments, we demonstrated a significant improvement in text classification as a result of introducing background knowledge. Specifically, we demonstrated: (1) improvement in text classification accuracy of human participants that were trained with some background knowledge compared to novices, (2) the effect of incorporating background knowledge to a vector-based representation model.

In the first experiment, we asked participants to answer four text classification questions, in which experts, who were trained to have some background knowledge about the movies, performed significantly better on classifying movie reviews compared to novices. This indicates that when people have textual prior information in a particular domain, they can perform more accurate classifications. Furthermore, analyzing the results based on the similarity of relational and surface features throughout the reviews, we demonstrated that when reviews shared common words and surface features, experts were able to select the reviews which belonged to the same movie based on their relational features. This proves the hypothesis that experts are able to identify deeper layers of similarity among reviews, while novices focus on surface features.

In the second experiment, we examined if incorporating background knowledge to a vector-based representation model would improve text classification accuracy. The task was to predict which of the three reviews were about the same movie based on the cosine similarity of the document vectors. The results indicated that providing textual background knowledge to the computational model improves the accuracy of text classification. We built our model by adding the background knowledge as a fixed vector to the neural network which was present during the document vector training process. Our results indicate that our model achieved significantly higher accuracy compared to the two first settings. This observation demonstrates that background knowledge should not be treated as simple bag-of-words, but it rather should be used as an interpretive lens through which other texts get trained.

A particular application of our model could be culturally-specific text classification. Our model could potentially be used to investigate the role of cultural knowledge in text comprehension and classification. Our prior work demonstrates that some cultural differences are evident even in how children's story books are written by different authors (Dehghani et al., 2013). Our proposed model can be used to further investigate such differences.

One limitation of this model is that it lacks a mechanism to form background knowledge or to update existing background knowledge. If this knowledge needs to be changed, or

if it is context dependent, we would need to manually feed the system with the newly fixed vectors representing background knowledge. One way to address is to use delayed-updateting rather than fixed vectors.

References

Bang, M., Medin, D. L., & Atran, S. (2007). Cultural mosaics and mental models of nature. *Proceedings of the National Academy of Sciences*, *104*(35), 13868–13874.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *35*(8), 1798–1828.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JAsIs*, *41*(6), 391–407.

Dehghani, M., Bang, M., Medin, D., Marin, A., Leddon, E., & Waxman, S. (2013). Epistemologies in the text of children's books: Native-and non-native-authored books. *International Journal of Science Education*, *35*(13), 2133–2151.

Goldberg, Y., & Levy, O. (2014). word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1* (pp. 142–150).

McClelland, J. L., Rumelhart, D. E., Group, P. R., et al. (1986). Parallel distributed processing. *Explorations in the microstructure of cognition*, *2*.

Mikolov, T. (2012). Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*.

Mikolov, T. (2014). *Word2vec toolkit discussion board.* https://groups.google.com/d/topic/word2vec-toolkit/Q49FIrNOQRo/discussion. (Last Accessed: 2015-01-29)

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Morin, F., & Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In *Aistats* (Vol. 5, pp. 246–252).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1–135.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Rosch, E. H. (1973). Natural categories. *Cognitive psychology*, *4*(3), 328–350.

Rottman, B. M., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive science*, *36*(5), 919–932.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Sagi, E., & Dehghani, M. (2014). Measuring moral rhetoric in text. *Social Science Computer Review*, *32*(2), 132–144.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15), 6424–6429.

Smith, E. E. (1995). Concepts and categorization.

Socher, R., Bauer, J., Manning, C. D., & Ng, A. Y. (2013). Parsing with compositional vector grammars. In *In proceedings of the acl conference.*

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1201–1211).

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (emnlp)* (Vol. 1631, p. 1642).

Wittgenstein, L. (1953). Philosophical investigations. basil & blackwell. *OxfordWittgensteinPhilosophical investigations Basil1953*.