Semantic Differentials for Wikipedia using the POLAR Framework

Jan Engler*
jan.engler@rwth-aachen.de
RWTH-Aachen University
Aachen, Germany

Ruixiang Wang*
ruixiang.wang@rwth-aachen.de
RWTH-Aachen University
Aachen, Germany

ABSTRACT

The goal of our project is to produce word embeddings from a Wikipedia dataset, and to apply the POLAR framework to different categories of words (e.g. Countries, Food, Music, etc.) in order to analyse the semantic associations behind concepts on Wikipedia. The POLAR framework is a method that enables interpretability for pre-trained word embeddings by making use of semantic differentials. The semantics of a word can be measured using semantic differentials by analysing its position on a scale between a word and its antonym.

In this paper we generate pairs of antonyms, using different methods, which can be used as the semantic differentials for the POLAR framework. To evaluate the interpretability of the pairs of antonyms generated, we performed a survey to collect the opinions of people from different countries towards the selected polar opposites. We compare the results with the conclusion from the POLAR framework and evaluate the quality of the semantic differentials produced.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning approaches**.

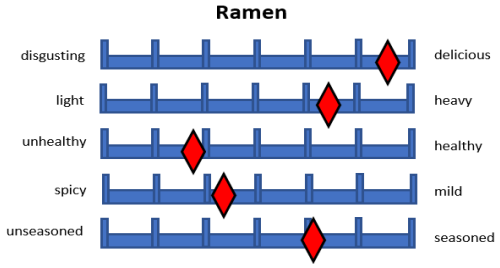
KEYWORDS

Word2Vec, word embeddings, POLAR framework, semantic differential

1 INTRODUCTION

Wikipedia is a free and open content online encyclopedia created through the collaborative effort of a large community of volunteers. According to Wikipedia [8], the English Wikipedia consists of over 6 million articles and on average around 600 articles are added every day. Currently Wikipedia contains over 3.6 billion words and, only considering the article text, the English Wikipedia has a size of way over 12 GB compressed [3, 12]. This huge size of text makes Wikipedia very attractive as a corpus for many different areas of machine learning and especially Natural Language Processing (NLP). The vocabulary size of the English Wikipedia is around 1.9 million [3], which would make it nearly impossible to work with the words on-hot-bit encoded. Therefore, word embedding techniques such as Word2Vec[5] and Glove[9] can be used to map the initial word vectors from a 1.9 million dimensional space into, for example, dimension 300. These smaller vectors can then be used for many NLP tasks.

Figure 1: Polar ratings for the word "Ramen"



However, such NLP methods usually are considered as a black box. The result can not be explained easily and the dimensions used do not have an understandable meaning.

In order to add interpretability into word embedding, Mathew et al. [4] introduced the POLAR framework that transforms the original word vectors into a POLAR space. In this POLAR space, the dimensions are associated with a scale between two polar opposites (e.g. "west" and "east" or "good" and "bad"). The semantics of words are then expressed by values on these antonym scales, which can be easily understood. However, they assume that the polar opposites are given by an oracle.

In out project we start by creating polar opposites ourselves in form of pairs of antonyms that are specialized for certain categories. For example, the rating of the word "ramen" by the polar framework with our pairs of antonyms can be seen in Figure 1.

Problem. We want to enable the analysis of words by semantic differentials from different categories in Wikipedia.

Approach. We first use the English Wikipedia as a corpus to train a Word2Vec model. Then we determine different topics like country, food, music or politics etc., for which we create different pairs of antonyms. Then we apply the POLAR framework to each topic individually, transforming the trained Word2Vec model into the corresponding POLAR space, by using the generated pairs of antonyms as polar opposites. Since the set of polar opposites could be potentially very large, we also select the best fitted antonyms as our POLAR dimensions.

Results and Evaluations. We created a Google survey and used it to collect the opinions of people from different countries towards polar opposites for different topics. Then we analyzed the quality of the produced pairs of antonyms and we compared the ratings of the participants and the ratings produced by POLAR.

The results from the survey indicate that most of the produced pairs of antonyms are relevant and are suitable to describe words

*Both authors contributed equally to this research.

from the chosen categories. The actual ratings of a word on the scale between the pair of antonyms of the POLAR framework is also often similar to the human ratings. The cases where they are different can be used to gain interesting insight of how the Word2Vec model on Wikipedia views the semantics of words.

2 BACKGROUND

In this section, we shortly introduce the POLAR framework and Semantic differential which are the basic concepts of our project.

2.1 POLAR framework

As stated already in the introduction, there are many different methods to embed the meanings of words into a vector, such as Word2Vec [5] and Glove [9], which have greatly improved the performance and accuracy of natural language processing tasks including machine translation [14], sentiment analysis [13], document classification etc. However, it is difficult for humans to make any sense of these word-vectors.

Since word embeddings are often the critical point for downstream machine learning tasks [4] but the results usually can not be explained easily, in many scenarios the interpretability of the word embedding becomes more and more important. The POLAR framework [4] is a good approach to add interpretability to these word vectors. The framework takes pre-trained word embeddings as an input, where the embedding is denoted by $\vec{\mathbb{W}} \in \mathbb{R}^{\mathbf{V}\times\mathbf{d}}$ ($\mathbf{V}$ is the size of the vocabulary and $\mathbf{d}$ is the output dimension). The resulting word vector for word i is denoted by $\vec{\mathbb{W}}_i \in \mathbb{R}^{1\times d}$ with $||\vec{\mathbb{W}}_i|| = 1$. Notice that the dimension of the word vector is d instead of V and usually $d \ll V$.

Then the POLAR framework generates word embeddings with interpretable (polar) dimensions as an output, by transforming the initial word embeddings into the polar space. The POLAR embedding is denoted by $\vec{\mathbb{E}} \in \mathbb{R}^{\mathbf{V}\times\mathbf{N}}$ ($\mathbf{N}$ is the new word embedding dimensions after POLAR transformation).

The framework first utilizes $\mathbf{N}$ word embeddings of antonym pairs (each of them is $(\mathbf{p}^i_z, \mathbf{p}^i_{-z})$) and calculates the director vector of each pair by:

$$\vec{\mathbf{dir}}_i = \vec{\mathbb{W}}_{\mathbf{p}^i_z} - \vec{\mathbb{W}}_{\mathbf{p}^i_{-z}} \tag{1}$$

Then the direction vectors are obtained as $\vec{\mathbf{dir}} \in \mathbb{R}^{\mathbf{N}\times\mathbf{d}}$, and the $\mathbf{n}$ most fitted antonym pairs are selected as the POLAR space (e.g. $\mathbf{n}$=300 using variance maximization and orthogonality maximization, see the following section or the original paper). Then the pre-trained embeddings of words are projected into POLAR space by using the following linear transformation:

$$\begin{aligned} \vec{\mathbf{dir}}^{\,T}\vec{\mathbb{E}} &= \vec{\mathbb{W}} \\ \vec{\mathbb{E}} &= (\vec{\mathbf{dir}}^{\,T})^{-1}\vec{\mathbb{W}} \end{aligned} \tag{2}$$

Finally, we get the new word embeddings $\vec{\mathbb{E}} \in \mathbb{R}^{\mathbf{V}\times\mathbf{n}}$. For a more detailed description the reader is referred to the original paper [4].

2.2 Semantic differential

The concept of semantic differentials developed by Charles E. Osgood [6] originates from the field of psychology and can be used to describe the associated meanings of words and concepts. It uses pairs of descriptive words (e.g. "good"-"bad") and asks the participant with which word she associates the target concept more. For example, in order to describe the connotative meaning of the word *sea*, the participant could rate whether she associates it more with "big" or with "small" and to which degree. Initially, a 7-step scale is given between the two descriptive words where the participant can vote on the degree of association [7]. Polar opposites or antonyms can be used for the pairs of descriptive words.

3 METHODOLOGY

In this section we explain the methods followed in this project. Initially we train a Word2Vec model on Wikipedia and design and generate pairs of antonyms for different categories. Finally, we apply the POLAR framework from Ref. [4] with the generated pairs of antonyms to the word embeddings, in order to add interpretability into the dimensions.

3.1 Train Word2Vec model on Wikipedia and Preprocessing

The goal of the present project is to explain semantic associations behind concepts on Wikipedia. Therefore, we first select a very large corpus containing English Wikipedia articles from the *Wikimedia dump service* [1]. Then, we apply some data preprocessing on this corpus like stopword and punctuation removal, removing all non-English words and lemmatization. On this corpus we train a Word2Vec model [5] using the `gensim`[2] library in python. Using skip-grams to train the model, the word vectors get reduced from the size of the vocabulary, which can be up to 1.9 million dimensions in Wikipedia [3], to 300 in our case. To visualize the result of Word2Vec, we plot the word embeddings of all countries in Figure 2. By using the T-SNE method [11], the similarities of the countries are shown in 2-D. As can be seen, mostly countries that lay close together in reality, are seen as similar to each other by the model. Interestingly, the Word2Vec trained on the Wikipedia corpus rates that "Canada" and "Australia" are very similar. In order to get a closer understanding for those interesting observations, we apply the POLAR framework in the following.

3.2 Create antonym pairs for specific topics

In the POLAR framework the word vectors get transformed into a new polar space. In this space, we can assign meaning to the different dimensions. For our approach, we use for each dimension the scale between a word and its antonym, e.g. good and bad. The chosen antonym pairs are the crucial part of this project. In this section we present multiple ways to create such pairs.

3.2.1 ***Create antonym pairs by hand.*** Initially, we manually created 50 pairs of antonyms (100 words) for each category. Although 50 dimensions are very few and probably not sufficient for most downstream tasks, they were used as a basis for further pair generation. However, coming up with 50 suitable dimensions for each different categories was surprisingly complicated already. For the category countries, for example, we wanted to integrate locational information or information about the language spoken

[1] https://dumps.wikimedia.org/enwiki/20200401/
[2] https://radimrehurek.com/gensim/models/word2vec.html

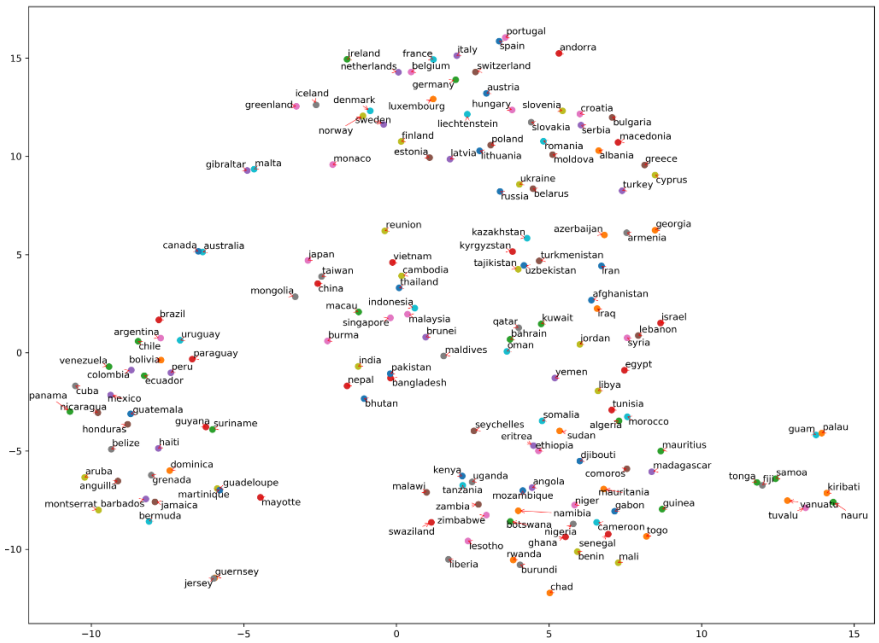


Figure 2: *All Countries Visualization.* We use the T-SNE [11] method to plot the word embeddings of country names into 2-dimensions. This figure shows that different countries form clusters according to their geographical locations *(Asia, Europe, Africa, South America).*

Table 1: Hand-crafted Antonyms

Country		Food		Music	
east	west	delicious	disgusting	classic	modern
democratic	monarchic	dry	juicy	dance	stand
tropical	polar	sour	sweet	monophonic	polyphonic
urban	rural	chewy	tender	pop	classical
ocean	land	baked	raw	soul	body

Table 2: Antonyms generated from scratch

Country		Food		Music	
north	south	hot	cold	popular	unpopular
known	unknown	equal	differ	like	dislike
nuclear	conventional	fire	hire	sound	silent
new	worn	import	export	record	erase
central	peripheral	expensive	cheap	more	fewer

into the dimensions. However, since words like "Europa", "Asia", "English" or "Chinese" do not have antonyms, we could not use these words. One extension to the approach of only using antonym pairs would be to allow pairs that are not strictly antonyms but may be seen as opposites in some situation, e.g. "USA" and "Russia" or "Rock" and "Pop".

For choosing antonyms we used websites like inspirassion.com to find descriptive words for categories and we used different online dictionaries to find antonyms. Examples for pairs of antonyms can be seen in Table 1. As can be seen in the column "Music", we made use of multiple meanings of words. Firstly, we allowed that words (e.g. "classical" or "sweet") can be used multiple times since they could be the only antonym to other important words (e.g. "Pop" or "sour"). Secondly, we used general antonyms like "soul" and "body" because the words could have a meaning specific to some categories, e.g. soul music.

3.2.2 ***Create antonym pairs from scratch.*** Since it is not practical to hand-craft up to 300 antonym pairs for every single category, we designed an approach to generate an antonym set from scratch. In this approach, since we were already using Wikipedia as a corpus, we also use Wikipedia pages from selected words as an input. Those words are examples for a category. We assume that the words used on these pages could be suitable for describing the category.

The steps are as following:

- Retrieve the Wikipedia page for entries from specific category (e.g. *Germany* from category *Country*).
- Preprocess the text of the page.
- Look up antonyms for the words used in the articles.
- Add antonym pairs to the generated antonym pair set.

From only 11 Wikipedia pages of one category, we could usually generate over 1500 antonym pairs, which is oviously much more efficient than hand-crafting them. Some generated antonyms are shown in Table 2. As can be seen, the generated antonyms are very general and less specific to the categories. However, also some category specific pairs were generated, yet overall very few. Since Wikipedia pages do not only describe the words but also their history and many other information related to the words, very general words are used which leads to these general antonyms. Hence, we conclude that either those pairs would have to be filtered to select the interesting dimensions or that we need to use some kind of initial antonym set, as suggested in the previous section, which we extend into a bigger set of category related pairs.

3.2.3 ***Extend hand-crafted antonym pairs.*** In this section we explain how we can extend a small set of hand-picked antonym pairs in order to get a bigger set of category related pairs.

Selection of initial set. Initially, a set of antonym pairs related to the selected category has to be selected. We identified two approaches to do this:

- Create 50 antonym pairs for a specific concept by hand, see Section 3.2.1.
- Select the most related 50 antonym pairs from an external source like [10] for a specific concept. This can either be done by hand or using the following method: For example, we first get the word embedding for the word *country* $\vec{\mathbb{W}}_{country} \in \mathbb{R}^{1\times d}$. Then, for each antonym pair in the given set $(\mathbf{p}^i_z, \mathbf{p}^i_{-z})$ calculate its mean vector value

$$\vec{\mathbb{W}}_{\mathbf{p}^i} = \frac{\vec{\mathbb{W}}_{\mathbf{p}^i_z} + \vec{\mathbb{W}}_{\mathbf{p}^i_{-z}}}{2} \tag{3}$$

  Finally compute the *cosine similarity* between each $\vec{\mathbb{W}}_{\mathbf{p}^i}$ and $\vec{\mathbb{W}}_{country}$ and select the most similar 50 antonym pairs.

Extend initial antonym set. For extending the initial antonym set, different methods can be applied. Our approach consists of two steps.

Firstly, generate similar or related words from the initial set. Here the goal is to select other words that can be used to describe words from the current category. For our initial set we firstly searched *Wordnet*[3] for synonyms which we add to our working list. Alternatively, we make use of the trained Word2Vec model and compute the n most similar words to the words from the initial set and add these to our working list.

Secondly, using all words from the working list, we look up their antonyms to build new pairs. Here we again used *Wordnet*, however, only very few antonyms were found this way and, therefore, we also extended our search by using the online dictionaries *thesaurus*[4] and *synonyms*[5] to crawl antonyms. Therewith, for each category we were able to find more than 300 pairs.

Finally, we perform post processing for the generated antonym pairs, like deleting duplicate pairs and sorting the antonym pairs by similarity score.

In contrast to the pairs generated from scratch in Section 3.2.2, less pairs were found, however, the pairs found were of higher quality for our approach and are still enough for most applications.

While analyzing the newly generated pairs we noticed that the definition of an antonym was sometimes very specific to topics. For example, "narrow" and "liberal" or "wet" and "rainless" are generally only seen as antonyms in specific fields, politics and weather. For this part we are depending on our sources to provide us with quality antonyms. Furthermore, we realized that using synonyms in the first step did result in many antonym pairs that nearly expressed the same, e.g. (salt, sweet) and (saltiness, sweetness). Therefore we mostly used the similarity approach.

We used the extended set of antonyms in the POLAR framework to describe the word "ramen". The result can be seen in Figure 1.

[3] *Wordnet*: https://www.nltk.org/howto/wordnet.html
[4] *thesaurus*: http://www.thesaurus.com
[5] *synonyms*: http://www.synonyms.com

Algorithm 1: generate antonym pairs from root

Input : R - set of root antonyms
Output: G - generated antonym pairs
$G \leftarrow \emptyset$;
$M \leftarrow \emptyset$ (synonym set of one word);
$k \leftarrow$ the number of synonyms we want;
for $p_i \in R$ **do**
  for $i \leftarrow 1$ **to** k **do**
    find a synonym s_i of the p_i;
    $M \leftarrow M \cup \{s_i\}$;
  end
  for $s_i \in M$ **do**
    find an antonym a_i of word s_i;
    $G \leftarrow G \cup \{(a_i, s_i)\}$;
  end
  $M \leftarrow \emptyset$;
end

This figure shows the five most important dimensions to describe the word and the values of the word embedding of "ramen" on these dimensions. For example, the most expressive dimension was the "disgusting" - "delicious" dimension, where the value of "ramen" is very far on the "delicious" side.

3.3 POLAR transformation

3.3.1 ***Select subset of antonym pairs.*** We use the *Orthogonality maximization* method from the POLAR original paper[4]. The idea is that we first select the direction vector $\vec{\mathbf{dir}}_i$ of an antonym pair which has the maximal variance when all the vocabulary $\mathbf{V}$ from Word2Vec model projected on it. Then we select $\mathbf{N}-1$ direction vectors from $\vec{\mathbf{dir}}$ that can maximize the orthogonality. Finally, we get a subset which contains $\mathbf{N}$ antonym pairs. The direction vectors of this subset are also called POLAR dimensions.

3.3.2 ***Linear transformation.*** We use these $\mathbf{N}$ antonym pairs and equation 2 to perform a linear transformation of Word2Vec model $\vec{\mathbb{W}}$. The inverse of the matrix $\vec{\mathbf{dir}}$ is accomplished through Moore-Penrose generalized inverse[1]. Then we apply standard normalization for $\vec{\mathbb{E}}$. This final $\vec{\mathbb{E}}$ is also called POLAR space model.

4 EVALUATION OF INTERPRETABILITY

In order to measure the interpretability of the produced semantic differentials we conducted a survey asking human participants from different countries to rate the semantics of words on a scale between two antonyms. These antonyms came from two sources, the hand-crafted set and the extended set of antonyms as described in the previous section.

The survey pursues multiple goals. The first goal is to evaluate the interpretability and meaningfulness of antonym pairs for different topics and the second goal is to compare the ratings of the human judgment and the results from the POLAR framework. Finally, we wanted to compare the quality of the hand-crafted and the generated set of antonyms.

Table 3: Overview of the participants in our survey.

(a)

Citizenship	Amount
German	16
Chinese	7
French	4
Belgium	1
Croatia	1
Egypt	1

(b)

Average age	23.5
(Est.) Female rate	40%

4.1 Overview.

The survey was conducted among 30 participants from six different countries. The countries are shown in Table 3 a). Nearly 75% of the participants are from Europe and almost 25% come from China. One participant grew up in Egypt and is now studying in Europe. According to surveys conducted in 2010 [2] and 2011 [8], the average age of readers and contributors on Wikipedia lies between 20 and 30. We believe that since the average age of participants in our survey is 23.5, this helps us to compare the results from these two groups. Furthermore, according to [8] the majority of contributors are of higher education, which is similar to the participants in our survey.

4.2 Survey.

The survey was conducted between 1/7/2020 and 3/7/2020, using Google Docs. The task was introduced at the beginning of the survey, where a brief explanation and examples of antonyms were also provided. We asked the participants to rate words and their semantics on a scale between two antonyms and provided the following example question: "How would you rate "steak" on a scale between very salty=0 and very sweet=10?".

Overall, we asked the participants to answer 10 questions about two target words each from three categories. The categories were countries, food and music, and the selected target words were USA, Germany, rice, salad, Pop and Rammstein (a popular German band). Thus, in total, 60 questions were asked to 30 participants yielding 1800 answers.

For each word, five of the 10 pairs of antonyms were taken from our self-designed pairs of antonyms. The other five pairs were taken from the extended pairs, generated by using word similarity and online dictionaries, see Section 3.2.3. The actual pairs of antonyms used in the survey were chosen according to the top absolute values across the dimensions in the POLAR word embedding for each target word separately, as suggested in the original paper [4]. This gave us the top-rated dimensions for the words, which we considered to be more realistic to measure interpretability than by hand-picking the dimensions.

An example question can be seen in Figure 3 and Figure 4. As can be seen, we intentionally kept these questions very generic. We did not specify explicitly that the participants should rate the *food* rice or the *country* USA, but we asked them to rate the word itself. In trained word embeddings, no distinction is usually made between the possible different meanings of a word. For instance, only one embedding is trained for apple as a fruit and Apple as a company.

Using the results from the survey, we can measure the agreement on one of the two antonyms of the survey and the rating of the POLAR framework, which we analyze in the Section 4.4.

We also stated that a rating of 5 should be selected if the participant thinks that the semantic differentials are not fitting to describe the target word. We take a closer look at the results of the survey and analyze the interpretability of the pairs in Section 4.3.

4.3 Interpretability.

In this section, we evaluate the interpretability of the pairs of antonyms that we generated. As mentioned before, we tried to generate antonym pairs that were fitting for the selected category without being too general. The participants could express that they found an antonym to be descriptive for a target word by voting in the direction of that antonym. Thus, a rating close to 0 or 10 means that one of the two antonyms is considered very fitting to describe the target word. On the other hand, a rating of 5 usually means that the given antonyms are not considered fitting for the word. Hence, we decided to measure the interpretability on the basis of the average distance to the rating of 5.

Firstly, we measure the overall mean and standard deviation over all 1800 answers. As expected, this overall mean is close to 5. Secondly, we compute the average answer and the standard deviation for every question individually. By averaging over all these standard deviations, we get an average of 2.03. This result indicates that the answers of the participants are on average 2 steps apart from the mean of the question. Considering that there are only 11 steps on the scale, a 2.03 standard deviation average implies that the answers vary significantly depending on the participants. The average standard deviation is 1.978 for the hand-crafted pairs of antonyms and 2.06 for the computer-generated (see Section 3.2.3) antonyms, which could indicate that the answers vary slightly more for the latter.

There is no "right answer" to the questions in our survey, and the rating is highly subjective. Furthermore, the survey was conducted during the "AVOID 19 pandemic" in July 2020, which obviously affected some of the answers. For instance, when asked to rate "USA" between "liberty" and "captivity". Many participants rated it more on the "captivity" side. This question, its sample mean and standard deviation are shown in Figure 4. The answers are very scattered and on both sides of the antonyms. While the average rating is around 4, the standard deviation is 2.6, indicating that on average the ratings are around 1, 2 or 6, 7 on the scale.

It is also important to distinguish the two types of results with a mean close to 5: If we get a sample mean of around 5 and a small standard deviation, we believe that the pair of antonyms are not suited for describing the target word. For example, for the question to rate "Rice" on a scale between "green" and "red", most answers were around 5 and the standard deviation is only 1.6. This indicates that this pair of antonyms is not suited for describing the word "rice". However, if the mean is close to 5 but the standard deviation is very large, we assume that the answer is very subjective and that the antonyms could still be fitting to describe the target word.

Figure 3: Survey question: Rate "Germany" between "hot" and "cold"

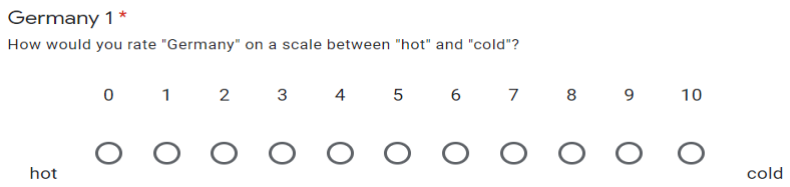
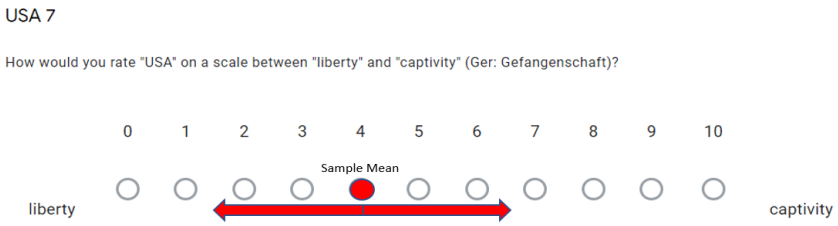


Figure 4: Survey question: Rate "USA" between "liberty" and "captivity"



A clear indicator that a pair of antonyms is relevant for describing a target word is a mean close to 0 or 10. For the question to rate "Salad" on a scale between "plant" and "animal" the average rating was 0.9 with a standard deviation of 1.01, indicating that this dimension is very good to describe the target word and that the participants mostly agreed on the answer.

Thus, we evaluated the average distance from rating 5 to determine which pairs of antonyms are fitting for the categories. The results are summarized in Table 4. The average distance is 2.27 from rating 5, which shows that all-in-all the pairs of antonyms could be used to describe those words. The average distance is smaller for the extended pairs, which could indicate that they are less descriptive than the hand-made ones, which was to be expected. However, this also depends on the category of words. For the country category, the average distance of extended pairs is even higher than for the hand-made pairs, while for the music category it is significantly lower.

Taking a closer look at the generated antonym pairs for music, it appears that some are indeed not very suited to describe the music related target words. For example: "stupid" and "streetwise" or "foggy" and "distinct". It thus appears necessary to be more careful in the generation of pairs of antonyms.

In the end of the survey we also asked the participants if they thought that the antonyms were suited for describing the target words. This question was optional and was answered by 17 of the 30 participants. 9 out of them expressed that most of the antonyms were fitting, e.g. " not always, for example red & green do not relate to rice" or "Most were fitting, at least so that you could have an association.". One participant said "no", and two participants said "yes". The rest wrote something which we would subsume as "sometimes fitting".

For our approach we had to find and generate antonyms. However, the definition of an antonym is rather vague and we had to rely on online dictionaries or *Wordnet* for providing us with real antonyms. In order to confirm that the used pairs were seen as antonyms, we asked the participants for their opinion. Again, the majority of the participants expressed that most of the pairs were "real" antonyms. One participant gave two examples of pairs that he/she thought are not real antonyms: "Most of them, but for example stupid and streetwise and superior and punk aren't real opposites in my opinion.". For both pairs, we confirmed that those are indeed antonyms in online dictionaries.

To summarize this section, we found that the answers to many questions are very subjective and the ratings can vary a lot. Furthermore, the survey indicated that the participants could largely describe the target words with the pairs of antonyms and that the antonyms used are mostly real antonyms and relevant for the categories. However, especially for the music category, the extended pairs of antonyms were seen as less descriptive than the hand-crafted ones.

Table 4: Distance from rating 5

Category	Source	Avg. Distance
Overall	Both	2.27
	Regular	2.45
	Extended	2.08
Country	Regular	2.1
	Extended	2.14
Food	Regular	2.61
	Extended	2.32
Music	Regular	2.63
	Extended	1.79

Table 5: Agreement on the antonym

Category	Source	Avg. Agreement	Avg. Distance
Overall	Both	66.67%	0.75
	Regular	76.67%	0.63
	Extended	56.67%	0.87
Country	Regular	60%	0.85
	Extended	40%	1.1
Food	Regular	90%	0.47
	Extended	70%	0.7
Music	Regular	80%	0.58
	Extended	60%	0.82

4.4 Agreement.

In this section we want to evaluate how the human judgment and the results from the POLAR word embedding agree or disagree for different categories.

First, we want to measure in how many cases the polar relation and the human judgment coincide. Here, we simply compare the direction of the ratings with each other, to see which antonym is more fitting for the target word.

For the hard-crafted pairs, the average agreement was 76.67% so in more than 3 out of 4 questions the average opinion of the participants agreed with the polar rating.

On the other hand, the average agreement on the extended antonym pairs is only 56.67%, which was lower than we expected.

As can be seen in Table 5, the overall agreement for questions related to countries is by far the lowest. For the extended antonym pairs this is only 40% while it is 70% for food and 60% for music. We believe that one reason, for this low agreement is due to the category itself. Opinions amongst participants can vary significantly, especially when it comes to countries. Here, we also want to note that in our survey we did not have any people from the USA, which is the country from which the largest amount of Wikipedia contributors come from [8] and which was one of the target words in this category.

Instead of simply comparing the direction, we also measured the difference in ratings. At this part we would like to emphasize that the POLAR framework relates the closeness of the target word to the antonyms on a scale where the values range mostly from −10 to 10. For the survey, due to limitations in Google Docs, we had to choose a scale between 0 and 10. We also noticed that some participants rated strongly around 0 and 10 while other participants used the values around 5 more. In order to make these results comparable, we normalized the results with the standard score.

With all ratings normalized, we compared the average rating of the participants to the rating of the POLAR framework. The results can be seen in Table 5 and confirm the results observed above. The average distance is higher for the extended antonyms and especially higher for the category "country".

In our approach we did not explicitly seek to get a high agreement. However, the goal was to analyze the semantic associations we get from performing word embeddings on Wikipedia and we believe that especially the dimensions where human judgment differs from the POLAR results could be of special interest.

Summing up this section, we have seen that the human judgment agrees with the ratings of the POLAR framework in many cases. However, some questions, especially for describing countries, are very subjective and therefore can result in differences in ratings.

5 CONCLUSIONS AND FUTURE DIRECTIONS

We have presented methods for generating category specific semantic differentials that can be used in the POLAR framework [4] in order to add interpretability into word embeddings. For evaluation we trained a Word2Vec model [5] on a Wikipedia dataset and then applied the POLAR framework on the resulting model. In this paper, we propose different methods for generating antonym pairs that are tailored to describing words from chosen categories, e.g. countries. The highest quality pairs of antonyms can be generated by extending a small hand-crafted set by searching similar words, using the word embeddings and then looking up their antonyms on *Wordnet* or an online dictionary. We analyzed the interpretability of the produced antonym pairs by conducting a survey. The survey results confirmed that most of the generated pairs can be used to describe words from the selected categories. Overall, the hand-crafted antonyms were seen as a better descriptive in most categories than the extended antonyms. However, by carefully designing the hand-crafted set and by filtering some of the pairs, pairs of high descriptive quality can be generated.

Future direction. As stated before, it is crucial to select well fitting antonym pairs for the categories. A next step would be to create some kind of improved filtering of antonyms. The goal is to filter out too general pairs that do not have any context-related meaning (e.g. 'general', 'particular'). On the other hand, too specific antonyms (e.g. 'staccato', 'legato') could also be excluded since they can not be used or understood easily to explain semantic associations. We suggest to use some kind of stoplist to filter out too general antonyms and to filter out words with a very low word frequency. Another step could be to perform a similar survey with more questions and categories, reaching more participants from more countries, in order to generalize the results. Furthermore, using the created POLAR word embeddings, downstream tasks like classification could be applied and their result and its derivation could be explained using the interpretable dimensions.

REFERENCES

[1] Adi Ben-Israel and Thomas NE Greville. 2003. *Generalized inverses: theory and applications*. Vol. 15. Springer Science & Business Media.

[2] Ruediger Glott, Philipp Schmidt, and Rishab Ghosh. 2010. Wikipedia survey–overview of results. *United Nations University: Collaborative Creativity Group* 8 (2010), 1158–1178.

[3] Chin-Yew Lin, Nianwen Xue, Dongyan Zhao, Xuanjing Huang, and Yansong Feng. 2016. *Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings*. Vol. 10102. Springer.

[4] Binny Mathew, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. 2020. The POLAR Framework: Polar Opposites Enable Interpretability of Pre-Trained Word Embeddings. In *Proceedings of The Web Conference 2020*. 1548–1558.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[6] C.E. Osgood, G.J. Suci, and P.H. Tannenbaum. 1957. *The Measurement of Meaning*. University of Illinois Press. https://books.google.de/books?id=qk5qAAAAMAAJ
[7] Charles E Osgood. 1952. The nature and measurement of meaning. *Psychological bulletin* 49, 3 (1952), 197.
[8] Mani Pande. 2011. Wikipedia editors do it for fun: First results of our 2011 editor survey. *blog.wikimedia.org* (jun 2011). https://blog.wikimedia.org/2011/06/10/wikipedia-editors-do-it-for-fun-first-results-of-our-2011-editor-survey/.
[9] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162
[10] Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2016. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. *arXiv preprint arXiv:1612.04460* (2016).
[11] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605. http://www.jmlr.org/papers/v9/vandermaaten08a.html
[12] Wikipedia contributors. 2020. Wikipedia:Size of Wikipedia. https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia [Online; accessed 07-July-2020].
[13] Lei Johnny Zhang and Bing Liu. 2017. Sentiment Analysis and Opinion Mining. In *Encyclopedia of Machine Learning and Data Mining*.
[14] Will Y. Zou, Richard Socher, Daniel Matthew Cer, and Christopher D. Manning. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *EMNLP*.

6 CONTRIBUTION

Jan Engler. Helped to train the Word2Vec model on Wikipedia and to convert it into POLAR space. Implemented the approaches to generate a set of antonyms from scratch (see Section 3.2.2), designed the final sets of hand-crafted antonyms for different categories (see Section 3.2.1) and implemented the approach to extend the hand-crafted set by using Word2Vec similarity and then Wordnet and www.synonyms.com (see Section 3.2.3). This author also performed the survey and evaluated the results (see Section 4). Finally, he composed and edited most of the final paper, except Section 3.3.

Ruixiang Wang. Use Wikipedia corpus to train Word2Vec model and then plot all countries embedding via T-SNE (see section 3.1). Create root antonym pairs by hand, and design a algorithm to create root antonym pairs automatically (see section3.2.3). Design a algorithm (*see also* sction3.2.3), which uses WordNet and tesaurus to generate new antonym pairs from root antonym paris. Select subset of generated antonym pairs by using *Orthogonality maximization* and do POLAR transformation(section 3.3). Help to collect answers of the survey and write final paper.