Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4754–4763
Marseille, 11–16 May 2020

Towards a Gold Standard for Evaluating Danish Word Embeddings

Nina Skovgaard Schneidermann[1], Rasmus Hvingelby[2], Bolette Sandford Pedersen[1]
[1]Center for Language Technology, University of Copenhagen,
Njalsgade 138, [2]Alexandra Institute, Njalsgade 76
ninasc91@gmail.com, rasmus.hvingelby@alexandra.dk, bspedersen@hum.ku.dk

Abstract

This paper presents the process of compiling a model-agnostic similarity gold standard for evaluating Danish word embeddings based on human judgments made by 42 native speakers of Danish. Word embeddings resemble semantic similarity solely by distribution (meaning that word vectors do not reflect relatedness as differing from similarity), and we argue that this generalisation poses a problem in most intrinsic evaluation scenarios. In order to be able to evaluate on both dimensions, our human-generated dataset is therefore designed to reflect the distinction between relatedness and similarity. The goal standard is applied for evaluating the "goodness" of six existing word embedding models for Danish, and it is discussed how a relatively low correlation can be explained by the fact that semantic similarity is substantially more challenging to model than relatedness, and that there seems to be a need for future human judgements to measure similarity in full context and along more than a single spectrum.

Keywords: word embeddings, semantic similarity, Danish

1. Introduction

'You shall know a word by the company it keeps' is a common phrase used to express the linguistic hypothesis that the meaning of a given word is, at least to some extend, a function of its surrounding context (Lenci, 2008). This assumption, better known as the distributional hypothesis, was earliest expressed by Firth (1957) and has, in spite of its widely debated psycholinguistic validity, been a basis for the linguistic analysis of meaning in certain methodological paradigms; namely, it has become a cornerstone for methodology in corpus linguistics, as well as the basis for generating computational models of semantic analysis using distributed vector representations of words (Goldberg and Levy, 2014; Mikolov et al., 2013a; Mikolov et al., 2013b).

One of the dominant trends in natural language processing (NLP) is the use of neural word embeddings, a non-linear model in which the linguistic input is represented as a vector in a dense vector space. Intuitively, feature vectors, which can be conceived as coordinate points in a vector space, represent words in a vocabulary as points in a distributional space, where each feature encodes a statistical association between the word and its surrounding context as defined by the model. As such, the assumption that semantic distances between two words are a function of their distributional similarity is encoded in the vector space as distances between points, typically measured using cosine similarity (Goldberg, 2016). Traditionally, techniques in natural language processing have relied on sparse vector inputs to a linear machine learning model such as logistic regression or support vector machines (SVM). In contrast, word embeddings differ by representing features as dense vector inputs trained on a neural network architecture using popular approaches such as CBOW or Skip-gram (Mikolov et al., 2013a; Mikolov et al., 2013b). Word embeddings trained over large portions of unannotated text, so-called *unsupervised word embeddings*, have major advantages compared to linearly trained models in their generalization power, i.e. they enable features to more efficiently encode statistical associations such that similar words have similar feature representations. As a consequence, word representations occupying similar positions in the vector space should occur in similar semantic contexts.

This paper presents a monolingually based similarity dataset for Danish word embeddings compiled on the basis of set of informants and applied to evaluate six pretrained word embedding models. It is our hope that, in an attempt to contribute to the optimisation of Danish word embedding models, the dataset assists in narrowing the increasing gap that exists between resources and research in language technology for Danish and for major European languages such as English (Pedersen et al., 2012; Kirchmeier et al., 2019). The dataset is made publicly available on GitHub[1] and through the DK-CLARIN platform.

One criterion for a good word embedding model is that the computational relationship between two vectors should mirror the linguistic relationship between the words they represent. In practice, this means that the distance between two vectors should reflect the more abstract notion of semantic similarity between two words, but semantic similarity is in principle not a well-defined concept. In terms of a word embedding model, similarity is defined solely by distribution; if two words occur in similar contexts, the words are taken to be similar. However, distributional similarity covers a wide range of semantic relations such as synonymy ('intelligent / smart'), hyperonymy ('bee / insect'), and cohyponymy ('cat / dog'). Furthermore, antonyms ('interesting / boring', 'fast / slow') present its own challenge to distributional models, because they may tend to appear in similar contexts, but have completely opposite meanings, which may also lead to questioning exactly how semantic similarity should be formally defined.

Even more importantly, distributional models do not tend to distinguish between semantic similarity and semantic re-

[1] https://github.com/kuhumcst/Danish-Similarity-Dataset

drink, (‘drink’)	0.6505241394042969
ostemad, (‘cheese sandwich’)	0.6464417576789856
kande , (‘pot’)	0.6422114372253418
tår , (‘sip’)	0.6274476647377014
kaffe , (‘coffee’)	0.6246160268783569
bajer , (‘beer’)	0.6167631745338440
croissant , (‘croissant’)	0.6159899234771729
sjus , (‘drink’)	0.6063054800033569
stempelkande , (‘cafetierre’)	0.6043441295623779
cappuccino , (‘cappucino’)	0.6042838096618652

Table 1: Most similar words to ‘kop’ (‘cup’) based on the `dsl` word embeddings

latedness of two words, which results in word pairs such as ‘coffee / cup’ being rated as distributionally similar due to their frequent co-occurrence, even though they hardly have similar meanings and are therefore only semantically associated. For illustration, consider Table 1 which shows an example of a word embedding query for the Danish word ‘kop’ (‘cup’), indicating that the model rates the pair ‘kop / drink’ (’cup / drink’) to be more similar than ‘kop / kande’ (’cup / pot’) even if cups and pots are solid physical objects formed as containers whereas drink is a liquid[2].

The purpose of our dataset is to be able to evaluate the “goodness” of Danish word embeddings using a constructed test set of 99 similarity judgements and subsequently evaluate existing pretrained Danish word embeddings. Our intend is to present a model-agnostic similarity goal standard for Danish that can be used to evaluate the performance on word embeddings, as well as provide linguistically interesting clues to the role of distribution in relation to meaning.

The paper is organised as follows: Section 2 discusses related work and methodological issues in relation to how to evaluate word embeddings. In Section 3 we present the query and experimental design behind our evaluation dataset, and in Section 4 the results are presented in terms of the achieved inter-annotator agreement as well as the correlation with the six word embedding models. In Section 5 we discuss and conclude.

2. Related work and methodological issues

Despite the popularity of word embeddings in NLP, there is as of yet no scientific consensus for the most adequate method of evaluating word embeddings (Bakarov, 2018). The most basic distinction between evaluation metrics is that of extrinsic and intrinsic evaluation: Extrinsic evaluation focuses on the application power of the model by assessing the word vector representation based on its performance on downstream tasks; i.e. it evaluates the ability of word embeddings to be used as feature vectors in a supervised machine learning task. Those tasks are usually computationally expensive, and it is widely agreed upon that those methods don’t transfer more generally; how well a word embedding does at one machine learning task doesn’t predict how well it will do at another task of a completely different nature (Bakarov, 2018; Schnabel et al., 2015). In contrast, a much more varied set of methods is those of intrinsic evaluation, which use experiments from cognitive sciences and psycholinguistics to directly explore syntactic or semantic relations between words (Baroni et al., 2014; Hill et al., 2015). Typically, such experiments involve a pre-selected query inventory consisting of word pairs that are then judged based on some criteria of semantic quality, yielding an aggregate score that functions as an absolute gold standard for evaluating the quality of semantic models. Such experiments usually involve crowd sourcing, although automatic extraction of linguistic information through annotated corpora or wordnets have recently become more common (Tsvetkov et al., 2015).

Within the group of intrinsic methods of evaluation, the use of word similarity judgements is by far the oldest and most represented evaluation metric in the literature (Bakarov, 2018; Faruqui et al., 2016). The word similarity method is based on the idea that distances between two word vectors in some embedding space can be assessed based on human judgements on the semantic distances between two words, usually normalized to a continuous scale in the interval 0-1. In the most common evaluation tasks, participants are given a set of manually selected word pairs and asked to assess the degree of similarity of each pair, which will then comprise a dataset of word pairs and their average similarity. Each of those pairs is then compared to the cosine distance between word vectors for a given model, yielding a single measure that reflects how well the model replicates similarity as defined by the dataset (Bakarov, 2018).

Several datasets consisting of similarity judgements have been conducted for the English language, most of which were not exclusively designed towards word embedding evaluation; see for instance the RG dataset, which was created by (Rubenstein and Goodenough, 1965) in order to empirically test the distributional claim that words common to the context of two words is a function of their degree of synonymy.

One of the most popular gold standards used for measuring the quality of word embeddings is Wordsim353, which consists of 353 word pairs rated on a 0-10 point scale by 13-16 participants on average for each pair, where “(...) 0 = words are totally unrelated, 10 = words are VERY closely related” (Finkelstein et al., 2002). In spite of its frequent usage, the dataset has been criticized on a number of methodological issues, namely that it is not explicitly clear whether the dataset measures semantic similarity or semantic relatedness, and that the dataset is arbitrary with respect to the query selection; e.g. it consists of pairs with mixed parts of speech (‘white / rabbit’, ‘run / marathon’, which is arguably counter-intuitive to how humans think about word similarity (Hill et al., 2015).

More recent English similarity datasets have attempted to correct for those errors, the largest of which is Simlex999 (Hill et al., 2015). Simlex999 consists of 999 word pairs judged by 500 participants, where each participant rated 119 pairs. Along with being the largest dataset, it is also the most rigid with respect to attempting to cover a wide range of linguistic concepts; the experiment attempted to prevent

[2] The word embedding query is generated using the `dsl` word embedding model (Sørensen and Nimb, 2018) with the Gensim library(Řehůřek and Sojka, 2010)

the arbitrariness of the query selection by distributing the words over the three major open word classes, in different ranges of frequency and concreteness of the words, and used a free association corpus to determine relatedness between words (Hill et al., 2015). It is the most explicit dataset with respect to ensuring that participants measure similarity under the same definition and emphasizing that related words are not necessarily similar.

For the time being, no similarity gold standard has as of yet been constructed for Danish on a monolingual, human-generated basis. Currently, the only available resource is a direct translation in Danish of Wordsim353[3] which has currently been used for evaluation and comparison for lack of better, but which undoubtedly introduces a language bias where a representative sample of the particular Danish vocabulary is not achieved, and where differences in ambiguity across languages introduces undesirable noise to the data. It is our hope that a Danish similarity dataset based on monolingual grounds can shed better light on the nature of word embeddings and their performance, as well as contribute to the pool of semantic resources for Danish.

2.1. Distinction between similarity and relatedness

Despite the frequent usage of the similarity method and its strong psycholinguistic background, a number of practical and theoretical problems with the reliability and validity of summary scores based on similarity have been identified.

One important methodological issue concerns the subjectivity of the notion of similarity in the different evaluation tasks. Specifically, many existing datasets for English do not distinguish explicitly between similarity and relatedness, which makes comparisons of gold standards challenging (Faruqui et al., 2016). The notion of *similarity* refers to the idea that two words belong to the same or a similar category, implying that they represent the same or a similar type of thing and can fulfill similar syntactic and/or semantic function in a sentence, whereas *relatedness*, also sometimes termed as *association* or *topical similarity*, merely requires two words to frequently occur in similar contexts. As an example, ‘coffee’ and ‘cup’ are related, but dissimilar, in that they describe completely different types of things; ‘cup’ refers to a human-made object used for ingesting liquids, while ‘coffee’ refers to a plant or a hot drink (Faruqui et al., 2016). Conversely, items such as ‘car’ and ‘train’ share numerous common properties, namely being vehicles and consisting of similar parts, and are thus functionally similar. To put it in more formal terms, the semantic relations that best represents similarity defined in this way is that of near synonymy (‘smart / intelligent’, ‘happiness / joy’ etc), and to a lesser extend hypernym/hyponym and co-hyponym pairs (‘bee / insect’, ‘cat / dog’), while related but dissimilar pairs are best described by the relation of meronymy (‘knife / blade’) or the concept of association, also sometimes termed topical similarity (Batchkarov et al., 2016). In this paper, semantic similarity will be defined as the extend to which two words both occur in similar contexts and express similar meanings. As a consequence of this definition, antonym pairs (‘short / long’, ‘interesting / boring’) should also be considered dissimilar and given a low similarity rating, challenging the model’s tendency to give high scores to antonym pairs. The assumption that antonym pairs are semantically dissimilar is henceforth taken for granted, because this lets us compare our dataset directly to Simlex999 by Hill et al. (2015), which employ a similar definition of semantic similarity.

As a default, word embedding models do not appear to distinguish between relatedness and similarity: For instance, the `dsl` model judges ‘cup / coffee’ to have a score of 0.624 on a 0-1 interval, which is much higher than some genuinely similar pairs; e.g. ‘car / train’ receives a score of 0.100. This is partially a feature of the distributional approach itself, in that this approach only requires two words to be part of the same context, not that the words have similar meanings. Even though a major appeal of word embeddings is that they can be applied to a wide variety of tasks without modification, research indicates that it may be beneficial for particular downstream tasks to specialize the model for either similarity or relatedness depending on the downstream task; namely, for applications such as topic modelling or document classification, it might be more interesting to know that ‘seat’ is associated with ‘car’ rather than knowing that ‘car’ is a hyponym of ‘vehicle’, whereas if machine translation, POS tagging, or synonymy detection is the application, relations of similarity are more relevant to achieving an accurate output. (Kiela et al., 2015) demonstrated this by using additional semantic resources to specialize word embeddings for either similarity or relatedness and subsequently comparing the retrofitted models with the unspecified learning approach on a range of extrinsic evaluation tasks, which resulted in a significant improvement on document classification and synonym detection with the tweaked models than with the unspecified approach.

For this reason, it is useful for datasets that function as evaluation benchmarks of word embeddings to be explicit about which of these components they measure. However, as hinted earlier, this is not always the case; for instance, the instructions for Wordsim353 are ambiguous with respect to the distinction and furthermore specify that antonym pairs should be considered “similar (i.e., belonging to the same domain or representing features of the same concept)” (Finkelstein et al., 2002). As a consequence, many dissimilar pairs receive high ratings, namely, the pair ‘coffee / cup’ is rated to be more similar than ‘car / train’, receiving a normalized mean similarity rating of 0.658 and 0.631 in the interval 0-1, which penalizes the model for displaying a preference for similar pairs over related, dissimilar ones, as well as for attributing low scores to antonym pairs (Batchkarov et al., 2016).

In an attempt to draw the distinction of similarity and relatedness on Wordsim353, (Agirre et al., 2009) separated the 353 evaluated pairs in Wordsim353 into two mutually exclusive datasets: WS-sim contained the set of pairs that were identified to contain either similar or unrelated concepts, while WS-rel contained the set of pairs identified to contain no similar concepts; i.e. the ‘union of related and unrelated pairs’ (Agirre et al., 2009). However, while this split allows for the related word pairs to be excluded, this

[3] `https://github.com/fnielsen/dasem/tree/master/dasem/data/wordsim353-da`

method still does not test a model's ability to attribute low scores to related, but dissimilar concepts.

The Simlex999 dataset (Hill et al., 2015) is the only well-known evaluation dataset to explicitly measure similarity, asking the participants to give items a high score if they had similar meanings and supplying examples of near synonym pairs to illustrate the point (Hill et al., 2015). This is also the approach taken for constructing a similarity dataset in this project due to the inability to reliably measure both similarity and relatedness with the resources available. Henceforth, we will use the term 'similarity' with no specified modifier to refer to the condition of words being semantically or functionally similar. The term 'relatedness' will generally denote distributional/topical similarity, although we may refer to 'distributional' or 'model similarity' when referring to similarity outputted by the model.

3. The evaluation dataset

3.1. Query design

Our similarity dataset consists of 99 word pairs of relatively frequent words selected from a sample of the 10,000 most frequent Danish lemmas as provided by ordnet.dk at The Society for Danish Language and Literature (DSL).

During the sampling, we attempted to select concept pairs covering a broad spectrum of semantic relations associated with semantic similarity, as well as associated but similar concept pairs. In general, the word pairs were selected from the sample by the following process: First, one query term was drawn randomly from the sample of 10,000 lemmas. Next, a target word was selected from the same sample if it met one or more of the following criteria:

- The target word had an existing relation with the query word in DanNet, the Danish wordnet (Pedersen et al., 2009), that could be associated with semantic similarity; e.g. namely near synonymy, antonymy, co-hyponymy, and hyper/hyponymy. This concerns relatively few of the final word pairs in the data, since the dataset does not aim to reflect semantic relations systematically. However, many of the word pairs exist on a spectrum of synonymy, which were drawn with inspiration from Simlex999.

- The word pair had an easily translatable equivalent word pair in Simlex999. Since both words were required to be within the 10,000 most frequent lemmas, this would mitigate the language bias that translation would provide. Most of the related (associated) concept pairs were selected based on their low scores in Simlex999, and other pairs were selected for their degree of synonymy, antonymy, or hyperonymy they reflected.

The final sample of query words are distributed over the 3 open word classes, nouns, verbs, and adjectives, the frequency distribution of which corresponds to their relative frequencies in KorpusDK. The choice to include multiple parts of speech as opposed to only nouns is motivated by the observation that different parts of speech exhibit distinct semantic properties which may influence the way they are rated; namely, adjectives can be considered to be more abstract items than verbs, which in turn are more abstract than nouns (Hill et al., 2015). Including multiple word classes therefore allows analyses of tendencies in the dataset. The sample consists of 55 nouns, 25 verbs, and 19 adjectives, with approximately half of the words falling within the 1000 most frequent lemmas in the corpus, whereas the remaining words are fairly evenly scattered across the remaining 9 intervals.

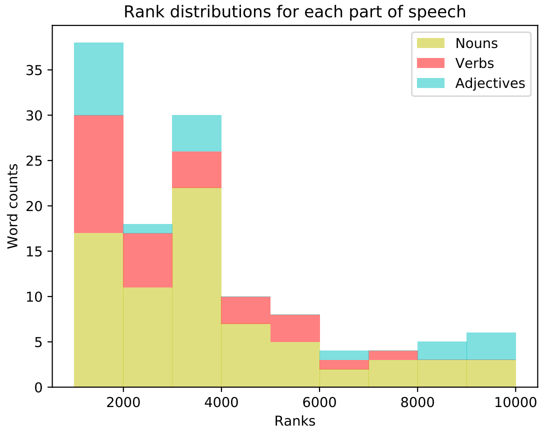

Figure 1: Frequency rank distribution of words over each part of speech in similarity dataset

3.1.1. Problem cases

An important methodical short-coming in our dataset is the ability to account for polysemy: Some words are either polysemous or homographs with common etymology; namely *svær*, which can mean either 'difficult' or 'physically straining', 'serious' (e.g. a disease), or 'heavy'. In those cases, due to the instructions being specific about considering word similarity as a function of synonymy, it is expected that the other word disambiguates the word sense to a certain extend such that participants select the score that results in the highest possible similarity between the words. Participants inquired into this and were asked to follow this procedure.

A further issue concerns the fact that our dataset does not evenly distribute between frequency bins (see Figure 1 and 2), and as such does not completely mitigate frequency effects. A more thorough query dataset sampling multiple frequency bins and containing rare words will therefore be required when the data set is upscaled.

3.2. Experimental design

3.2.1. Questionnaire structure

The questionnaire design of our dataset is inpired by that of Simlex999 (Hill et al., 2015). The questionnaire comprised of 112 questions spread over 16 groups in order to ease the burden of the annotators. The sample was not randomized, so all participant would answer the questions in the same order; first the 19 adjective pairs, then the 25 verb pairs, and last the 55 noun pairs. Within the same part of speech, each group except the last would contain 7 questions, of which for each consecutive group from the 2nd group to the 2nd last group, the last question of the previous group would be repeated as the first question in the subsequent group. This design was intended to ensure that participants recalibrated their ratings relative to the other 6 word pairs within the group, since inevitably, participants would be rating the similarity of each pair in comparison to each of the other pairs in the group.

3.2.2. Instructions

Participants were asked to enter an integer between 0 and 6 for each item reflecting how similar they assessed the pair to be, where 0 represented complete dissimilarity and 6 complete similarity. Two words were specified to be considered similar if they had similar meanings, which was exemplified with near synonym pairs as very similar (5-6) and antonym pairs as dissimilar (0-1), and participants were instructed to consider the examples of synonymy and assess word pairs according to the degree that they could replace one another in the same context without any change of meaning. Participants were also introduced to the distinction between similarity and relatedness and the idea that two words could be associated, i.e. belong to the same domain, without being similar. Last, the following illustrative example pairs with scores were presented as a guideline to ensure that participants understood the essence of the task:

- Klog / intelligent ('clever / intelligent'): 6
- Misundelse / jalousi ('jealousy / envy'): 5
- Kage / brød ('cake / bread'): 2
- Bil / motorvej ('car / highway'): 1
- Stor / lille ('big / small'): 1
- Højtaler / blomst ('loudspeaker / flower'): 0

3.2.3. Participants

Participants were mainly recruited from linguistics and data science student boards on social media and mailing lists. The respondents were required to be fluent speakers of Danish and use it throughout their daily lives in multiple contexts, although this was not formally tested for. The response time for each participant varied significantly, although the average response time was 15 minutes, which was also the estimated time given to the participants in the instructions. No data other than their response and time spent was preserved for each participant. 94 participants filled out the survey, but more than half of the participants did not complete the survey. Of those, 52 participants left 50 questions or more unanswered and were automatically excluded on this condition, since it is crucial that each item is rated by approximately the same number of annotators in order for further analysis to be possible. This left 42 participants who responded to 108 or more of the 112 questions in the survey, of which some were excluded based on outlier criteria.

3.3. Post processing

The post-processing of the collected data consisted of dealing with missing values in the data, calibration, and normalization of the mean similarity scores. Subsequently, since repeated questions were meant to recalibrate scores, the first of the repeated pairs were removed from the dataset before any further post-processing steps were applied. Next, in order to control for systematic biases between raters (Hill et al., 2015), we computed the absolute difference between the total mean similarity score for the dataset and the mean similarity score for each annotator. For three raters, this value exceeded 1; which means that 3 annotators had a tendency to rate items as either more or less similar than the general rater population. In those cases, all the scores for those raters were either decreased or increased by 1, except in cases where the annotator gave items either the minimum (0) or maximum (6) rating score. This calibration resulted in a small increase to the inter-annotator agreement.

After correcting for this systematic rater bias, we excluded participants whose average pairwise Spearman rank correlation with each of the other participants was less than 1 standard deviation below the mean Spearman rank correlation for the whole dataset. Four outliers were excluded based on this condition, leaving 38 participants in the final dataset.

Finally, the mean similarity scores were computed for each pair, and the mean similarity scores were linearly transformed from the range 0-6 to the range 0-1 (Hill et al., 2015) (Finkelstein et al., 2002). This lifts the inter-annotator agreement for the whole dataset from 0.634 to 0.687 (see section 5 for further details). This still leaves 7 missing values in the dataset in total. In general, during analyses, these values are left out of the calculation.

4. Results

4.1. The data

4.1.1. Inter-annotator agreement

Researchers typically report inter-annotator agreement as the mean Spearman rank correlation coefficient over all pairwise comparisons; either by calculating the correlation of each participant with every other participant or by comparing each participant to the overall gold standard, i.e. the mean similarity score over all items (Hill et al., 2015) (Dror et al., 2018). This captures the fact that similarity is measured on a continuous scale, which contrasts with many other NLP tasks where variables are categorical, in which cases Cohen's Kappa is used instead (Batchkarov et al., 2016).

Figure 2 shows the pairwise correlations between all annotators compared with the correlations between each pair and the (almost) gold standard for the dataset. In general, all annotators rank fairly highly measured against the similarity gold standards,, the values ranging from a minimum of 0.62 to a maximum of 0.92, with a mean score of 0.82. The pairwise correlations are slightly lower, ranging from a minimum of 0.29 to a maximim of 0.87, with a mean inter-annotator agreement of 0.68. Judged by the scores on Wordsim353 and Simlex999, which have an average pairwise correlation of 0.61 and 0.67 respectively, this inter-annotator agreement seems to lie within the expected range for similarity datasets (Hill et al., 2015).

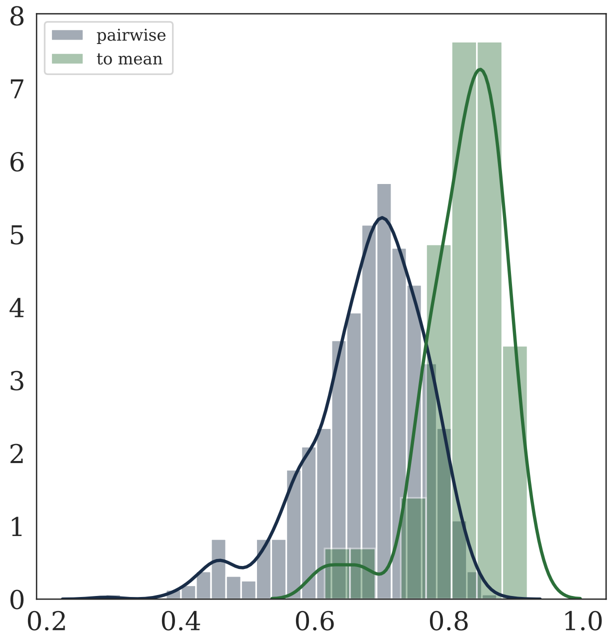

Figure 2: Pairwise and mean Spearman's ρ correlations for dataset

4.1.2. Distribution of scores

Figure 3 shows the distribution of similarity scores between all annotators, as well as the difference between the mean similarity gold standards and all raters. The 3 peaks in the graphs indicate values with the lowest spread in inter-annotator agreement.

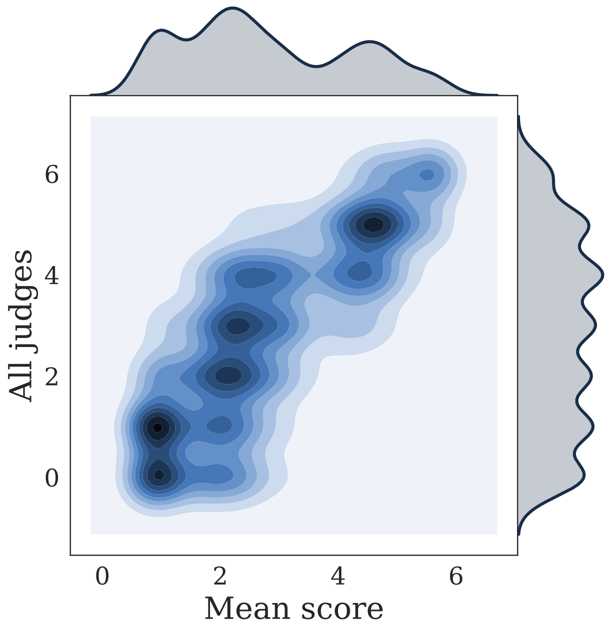

Figure 3: Distribution of similarity scores between annotators

From the graph displaying the scores for all judges, it's clear that the similarity scores are fairly evenly distributed, with an almost uniform distribution between values 0-5, but with a slightly lower number of maximum similarity scores (6). As such, there is a slight bias towards the lower similarity scores, which can also be confirmed by the fact that the non-normalized mean for the dataset is 2.805, and that 75 % of the mean scores lie beyond 4.5[4]. This may just be a consequence of the word pairs chosen in the data rather than a reflection on biases in the participants, since the Wordsim353 dataset has been reported to be biased towards higher scores (Batchkarov et al., 2016). However, the relatively uniform distribution of scores indicates that the dataset manages to capture word pairs on the entire spectrum of similarity, and that the 7 point similarity scale is a fairly reasonable choice of range for participants to manage.

A slightly more undesirable characteristic of the dataset is that the variance between annotators is quite high in most cases when compared to the mean similarity scores. From the 2nd graph in Figure 3, we can observe that participants are more in agreement about items that receive lower scores (1-2) and higher scores (4-5), and more uncertain about pairs receiving scores in the middle. Furthermore, by comparing the difference between the two graphs, one can also observe that the minimum mean score in the data is closer to 1 than 0, even though participants give 0 approximately the same amount of times as they give 1. This indicates that participants differ significantly with respect to the items that get the lowest score. This observation is interesting, because it could be interpreted as a confirmation of the methodological problem with rating pairs of different semantic relations on the same scale; for instance, it might be difficult to determine whether antonym pairs such as 'borring / interesting' should be judged as more or less dissimilar to each other than associated pairs such as 'car / seat'. Thus, it is necessary to further test inter-annotator agreement over different concept types in order to observe more consistent patterns in the data.

4.1.3. Differences across concept types

For the purpose of the analysis, the dataset was divided into 6 subsets based on semantic relations; synonymy, antonymy, hyperonymy, co-hyponymy, association/other, and similarity. The 5 first subsets cover the entire dataset. The relation 'associated/other' covers the 42 items in the dataset that were not divided into any of the 4 primary relation types. The 6th subset, similarity, is then the union of the subsets of pairs falling into the primary 4 relations, which results in 57 pairs estimated to exist somewhere on the spectrum of similarity[5].

Table 2 shows the inter-annotator agreement, response consistency (given by the standard deviation of the average pairwise correlations (Hill et al., 2015), and the normalized mean, minimum, and maximum similarity score for each subset of the data, including the division between similar and related items.

[4] Note that the minimum and maximum scores of the non normalized means are not 0 and 6 respectively, but 0.184 and 5.947

[5] Wherever possible, relations were extracted from DanNet (Pedersen et al., 2009). In cases where words had multiple synset specifications, the sense of the target word that maximized the similarity with the context word was chosen, in accordance with how participants were assumed to disambiguate word senses

Relation	Count	ρ	Cons.	Mean	Min	Max
synonymy	23	0.316	0.238	0.813	0.641	1.0
antonymy	12	0.180	0.410	0.059	0.0	0.215
hypernymy	14	0.454	0.302	0.482	0.262	0.754
co-hypo	8	0.230	0.398	0.293	0.159	0.410
sim all	57	0.757	0.097	0.499	0.0	1.0
assoc/other	42	0.431	0.169	0.297	0.015	0.656

Table 2: Comparison of measures on semantic relations

Model	DSD-ρ	WS353-ρ	WS353-OOV	Voc	Dims
dsl	**0.342**	0.531	1.13%	1.2M	500
cc	0.313	0.533	1.70%	2M	300
news	0.306	0.541	4.25%	2.4M	300
wiki	0.205	**0.639**	0.85%	0.3M	300
sketcheng	0.197	0.626	0.85%	2.4M	100
conll17	0.150	0.549	1.70%	1.7M	100

Table 3: Evaluation on Danish Similarity Dataset and WS353-da. The highest correlation is in bold

We can observe that participants generally appear to give overall low scores to antonym pairs and overall high scores to synonym pairs. In order of preference of overall mean similarity scores, participants give the highest score to synonyms, slightly lower to hyponyms, relatively low for co-hyponyms, and lowest to antonyms. Furthermore, the maximum normalized score score of 1 is given to the synonym pair 'rollemodel / forbillede' and the minimum normalized score of 0 is given to the antonym pair 'skadelig / harmløs', which also reflect the fact that synonymy and antonymy is on either end of the spectrum of similarity. However, the spread between the annotators appears to be quite high; when inspecting the scores, the pair 'glad / positiv' received 0 from one participant, and the antonym pair 'spare / investere' receives 5 by another. Most interestingly, the scores for co-hyponym pairs like 'kat / hund' and 'dyr / menneske' are much lower than expected given that those are in many cases substitutable and describe items within similar categories, although they clearly do not have similar meanings, which may indicate that inter-annotators don't consider similarity between meanings on one spectrum according to strictly defined semantic categories.

Regarding the distribution of scores over similar and associated items, The general trend is that both inter-annotator agreement and overall similarity scores are higher on the similarity subset than on the association/other subset; particularly, the mean score for the related items is only 0.397, which seems to verify that low scores are generally given to related but dissimilar items, even though antonym pairs have the lowest mean similarity score. The inter-annotator agreement on the similarity subset is 0.757 vs a mere 0.431 on the relatedness subset. From this, we can conclude that despite the high variation in scores, the overall mean scores do seem to reflect similarity rather than relatedness, and the dataset is therefore a reasonable initial attempt at a gold standard for measuring semantic similarity as defined in this paper. Conversely, since the dataset does not measure the ability of the models to score related pairs adequately according to syntagmatic contexts, it would be incorrect to make conclusions on how well the dataset measures relatedness/association.

4.2. Model evaluations

We evaluate six Danish word embeddings all trained with either Word2Vec or fastText (Bojanowski et al., 2017; Mikolov et al., 2013a). The Word2Vec model is a shallow, two-layer neural network that either predicts a current word given a window of surrounding words or predicts the surrounding words given a current word. The former is known as Continuous Bag-of-Words(CBOW) and the latter is known as Skip-Gram. The model can be trained on large corpora of raw text as it requires only valid text and the hidden layers are then used as representation of a word. The Word2Vec model takes as input a whole word, which introduces the risk of out-of-vocabulary words, whereas the fastText model is similar to the Word2Vec model but it used character n-grams as input.

Three of the word embeddings are trained using Word2Vec namely news, conll2017 and dsl. The news word embeddings are 300 dimensional and trained with Skip-Gram on approximately 30 million Danish digitized newspapers pages from 1880 to 2005[6]. The conll2017 word embeddings[7] are 100 dimensional and are trained with Skip-Gram on the Danish part of the CoNLL 2017 Shared Task (Ginter et al., 2017) raw data. The raw data were collected from CommonCrawl and Wikipedia and the language has been identified by a language detection tool. The dsl embeddings from (Sørensen and Nimb, 2018) are trained using CBOW features on a Danish corpus containing roughly 920 million running words at the time of training, spanning over a variety of text types from between 1982 and 2017, namely newswire, extracts from magazines, transcripts from the Danish parlament, and fiction. The model is trained over 500 features with a symmetric context window size of 5 and a minimum word count of 5 for all word form types.

The remaining three word embeddings are trained with fastText (Bojanowski et al., 2017). The wiki embeddings [8] are 300 dimensional and were trained with Skip-Gram on the Danish Wikipedia, the cc embeddings [9] (Grave et al., 2018) are 300 dimensional and were trained with CBOW on the Danish Wikipedia and CommonCrawl, where the language of text was identified with a language detection tool, and the sketchengine word embeddings are 100 dimensional and were trained with Skip-Gram on approximately 2 billion tokens of Danish web text, gathered by SketchEngine[10].

All the word embeddings have been evaluated on Danish Similarity Dataset and the Danish Wordsim353. We report the Spearman's ρ correlation coefficient along with the OOV-rate in Table 3.

[6] https://loar.kb.dk/handle/1902/329

[7] http://nlpl.eu/repository/

[8] https://fasttext.cc/docs/en/pretrained-vectors.html

[9] https://fasttext.cc/docs/en/crawl-vectors.html

[10] https://embeddings.sketchengine.co.uk/static/index.html

5. Discussion and concluding remarks

It is evident from the comparative results in the previous section that all models consistently receive a lower score on the DS dataset than on the translated Wordsim353, with the highest Spearman correlation on the `dsl` model of $\rho = 0.342$. This difference across the board may indicate the difficulty of modelling semantic similarity, which is further supported by the fact that (Hill et al., 2015) also report lower scores on Simlex999, which uses similar instructions and metrics as the DS dataset; namely, a state of the art model trained on English Wikipedia by (Mikolov et al., 2013a) receives merely $\rho = 0.414$ compared to $\rho = 0.655$ on the original Wordsim353 dataset. Similarly, (Batchkarov et al., 2016) report a model score of $\rho = 0.31$ on Simlex999 compared to $\rho = 0.64$ on Wordsim353. Given that this feature of the results are supported by earlier findings in English, this may suggest that semantic similarity is substantially more challenging to model than relatedness, namely in that semantic similarity introduces a meaning component in addition to the distributional analysis, and that semantic similarity comprises more than one semantic relation, suggesting that similarity may not be measured accurately on one scale.

One issue in the dataset occurs due to the inability to account for homography or polysemy, which suggests that certain words may have another meaning in the training corpora of the models than the one measured by the DS dataset; namely, the pair 'yderlig / radikal', in which *radikal* can both denote 'radical' and a large Danish political party. In this case, the word *yderlig* ('extreme') suggests that 'radical' should be selected by the annotators as the prefered meaning. However, the `dsl` embeddings suggests 'konservativ' and 'socialdemokratisk' ('conservative' and 'social democratic' as the 2 most similar words to 'radikal' [11].

Most recent gold standard initiatives address the issue of polysemy by considering words with respect to a context. Namely, Pilehvar and Camacho-Collados (2019) presents a dataset for evaluating context-sensitive word embeddings, in which a target word is evaluated with respect to two contexts represented by text examples. The two text examples then receive a binary label that indicates whether the occurrence of the target word corresponds to the same or a different meaning. Such similarity datasets allow for intrinsic evaluation of the newer contextual word embeddings (Peters et al., 2018; Devlin et al., 2019) as these models rely on the context of a word as a basis for forming word representations. Currently no such pretrained model exists for Danish, however constructing a dataset with words in context would be an interesting research direction to allow for future pretrained Danish contextual word embedding models. Alternatively, in other gold standards, such as the one constructed by Schnabel et al. (2015), participants are asked to rank the similarity of a target word with respect to the query words in a specific word embedding model. While this seems undesirable as a general model-agnostic metric for comparison of different embeddings, Implementations of such evaluation metrics for Danish might be another interesting research step to consider.

A further methodological issue that is evident in our dataset is the high variance in scores and low inter-annotator agreement; Even when participants were instructed to consider similarity on a scale of synonymousness, this only guarantees a relatively well defined notion about pairs with obviously highly synonymous meanings, but it does not necessarily specify whether antonyms should be considered more or less similar than related pairs, or whether cohyponyms can in fact be considered more in the relatedness category than the similarity one, given their low scores on the dataset. This suggests that participants do not consider similarity on a spectrum, and that it might therefore be problematic to rate different semantic relations on the same scale (Avraham and Goldberg, 2016; Faruqui et al., 2016). This speaks to the fact that semantic similarity is an intuitive concept that is difficult to quantify or model. To achieve a less artificial insight into the way humans conceptualize meaning and distribution, supplementing similarity experiments with methods in psycholinguistics, such as semantic priming or neural activation patterns, might give a more accurate insight into the cognitive reality of semantic similarity (Auguste et al., 2017; Bakarov, 2018). It might also be worth considering whether the choice of a relatively fine-grained scale of similarity from 0 to 6 is suitable for computational purposes; considering similarity as a binary classification task or using a 0-3 interval might raise the inter-annotator agreement while still give a sufficient picture of similarity for the purposes of word embedding evaluation.

Finally, a specific issue with our dataset is its small size, which makes it difficult to ensure that the query inventory covers a linguistically representative sample of the Danish language, and therefore, that the results derived are statistically significant. The lack of Danish linguistic resources presented an additional challenge, particularly with regards to making sure that similar and related pairs were relatively evenly distributed across the data. Most of the related pairs were translated from Simlex999 and selected as a consequence of achieving low scores on that dataset, which, although being the least time consuming, is not an optimal methodology. A larger and more systematic query selection would be ideal as a continuation of the similarity experiment; particularly, it would be interesting to cover less frequent words in order to figure out whether the evaluation on the models was biased by frequency effects, since models trained with Skip-gram on character n-grams tend to have an advantage with respect to rare words (Mikolov et al., 2013a).

6. Bibliographical references

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Auguste, J., Rey, A., and Favre, B. (2017). Evaluation of

[11] see `http://wstest.dsl.dk/w2v/most_similar?positive[]=radikal`

word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 21–26.

Avraham, O. and Goldberg, Y. (2016). Improving reliability of word similarity evaluation by redesigning annotation task and performance measure. *arXiv preprint arXiv:1611.03641.*

Bakarov, A. (2018). A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536.*

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247.

Batchkarov, M., Kober, T., Reffin, J., Weeds, J., and Weir, D. (2016). A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 7–12. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276.*

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1):116–131.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis.*

Ginter, F., Hajic, J., Luotolahti, J., Straka, M., and Zeman, D. (2017). Conll 2017 shared task-automatically annotated raw texts and word embeddings. lindat/clarin digital library at the institute of formal and applied linguistics, charles university.

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

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893.*

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Kiela, D., Hill, F., and Clark, S. (2015). Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2048.

Kirchmeier, S., Henrichsen, P. J., Diderichsen, P., and Hansen, N. B. (2019). Dansk sprogteknologi i verdensklasse. *Language Technology Committee under The Danish Language Council.*

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

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

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

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., and Lorentzen, H. (2009). Dannet: the challenge of compiling a wordnet for danish by reusing a monolingual dictionary. *Language resources and evaluation*, 43(3):269–299.

Pedersen, B. S., Wedekind, J., Kirchmeier-Andersen, S., Nimb, S., Rasmussen, J.-E., Larsen, L. B., Bøhm-Andersen, S., Henriksen, P., Kjærum, J. O., Revsbech, P., Thomsen, H. E., Hoffensetz-Andresen, S., and Maegaard, B. (2012). *Det danske sprog i den digitale tidsalder.* Springer.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

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

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. http://is.muni.cz/publication/884893/en.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal, September. Association for Computational Linguistics.

Sørensen, N. H. and Nimb, S. (2018). Word2dict–lemma selection and dictionary editing assisted by word embeddings. In *The XVIII EURALEX International Congress*, page 146.

Tsvetkov, Y., Faruqui, M., Ling, W., Lample, G., and Dyer, C. (2015). Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2049–2054.