Semantics is Actually 82% Distributional, but Neural Networks Aren't.

Anonymous ACL submission

Abstract

Distributional semantics is often proposed as the linguistic theory underpinning many of the most efficient current NLP systems. In the present paper, we question the linguistic well-foundedness of these models, addressing it from the perspective of distributional substitution. To that end, we provide a dataset of human judgments on the distributional hypothesis, and highlight how humans cannot systematically distinguish pairs of words solely from contextual information. We stress that earlier static embedding architectures are competitive with more modern contextual embeddings on the distributional substitution task, and that neither serve as good models of human linguistic behavior.

1 Introduction

One would not argue that the manner by which a pocket calculator estimates the value of 156 837 × 86 942 correctly depicts the mental processes of a human tasked with the same problem. Curiously enough, the same isn't held for language tools.

Recent NLP neural networks boast impressive feats on a wide variety of benchmarks. Crucially, the crux of NLP research has focused on efficiency: how to produce the highest score on some well delineated task. Little interest has been directed to assessing the linguistic value of these models: oftentimes, authors only hearken back to early works in distributional semantics, such as Harris (1954). There is however some criticism directed towards this framework: works such as Searle (1980), Harnad (1990) or Bender and Koller (2020) argue that text alone cannot suffice to derive meaning.

Given these known flaws, it is worth asking what value our recent large pretrained models have to the linguist. Here, we set out to see what quantitative arguments can be made in this debate: to what degree is the distributional hypothesis of Harris (1954) invalid? To what degree does it fit the behavior of models such as BERT (Devlin et al., 2019) or word2vec (Mikolov et al., 2013a)?

The approach we take here consists in testing models of distributional semantics on the distributional substitution task, which we frame in a manner reminiscent of the Cloze Task (Taylor, 1953). Given a target word, a distractor and a set of contexts containing the target but not the distractor, we replace target words with blank tokens and investigate whether models distinguish the target from the distractor. This task has a number of merits. It allows us to test many distributional models through their objective function, rather than rely on external parameters as with probing methodologies. It is also fairly intuitive to explain to annotators; and we therefore can compare networks to humans.

Our findings highlight a number of counterintuitive facts: recent contextualized embeddings are comparable to earlier static embeddings, and noticeably under-perform human annotators. More intriguing is the fact that embeddings do not appear to match human behavior more closely than n-gram baselines, casting doubt on their validity as models of human linguistic behavior.

The rest of this article will be structured as follows. We sketch a description of the theoretical framework of our analyses in Section 2, and detail the empirical data we base our experiments on in Section 3. Sections 4, 5 and 6 describe the experiments we conduct. Lastly, we provide a brief review of existing related works in Section 7 and some perspectives for future work in Section 8.

2 Distributional substitution

In his seminal paper on distributional semantics, Harris (1954) proposed the distributional hypothesis: word meaning should correlate with word distribution. We refer the reader to reviews such as Lenci (2018) or Boleda (2020) for a more thorough introduction. Here, we adopt the following view on distributional semantics: a distributional semantics model (or ‘DSM’) must be able to express which of two words is more appropriate in a given linguistic context. More formally, we expect of a DSM that it provides an estimate for:

$$p(w_1|c) > p(w_2|c) \tag{1}$$

This equation can be seen as implementing the distributional principle of substitutability, which was already sketched out in Harris (1954). In essence, we expect that these models are able to characterize the effect of substituting one word (w_2) for another one (w_1) within a given linguistic context (c). This principle of substitution has been used in other studies (Ferret, 2021, e.g.).

This view is also grounded in the fact that many word embedding and distributional models are able to yield an expression such as the one above. If we adopt the “count” vs. “predict” dichotomy of Baroni et al. (2014), which categorizes DSMs according to whether they are derived by tabulation (“count”) or inference (“predict”), we can see that both “count” and “predict” models are based on estimates of the conditional probability of words given their linguistic contexts. The main difference lies in that “count” models derive this estimate from descriptive statistics, whereas “predict” models learn it using inferential models such as neural networks.

Another argument, first expressed by Sahlgren (2008) and discussed by Gastaldi (2021), stresses the theoretical connection between this conditional probability $p(w|c)$ and the paradigmatic axis in the structuralist framework of linguistics (Saussure, 1916). Both are referring to the ability to model which linguistic expressions fit in a given context.

3 Dataset

To study how models perform on the distributional substitution task, we begin by collecting human judgments, using a crowd-sourced gamified approach in compliance with GDPR laws. All source corpora are made available under CC-BY-SA licenses; we will release our collected data under the same license upon acceptance. A companion paper describes the data collection procedure in depth.

3.1 Dataset construction

We collect data in 5 languages: English, Spanish, French, Italian and Russian. Analyses presented here are derived from a set of 14493 annotations.

	en	es	fr	it	ru
$k=1$	329	110	540	161	113
$k=3$	58	90	136	73	90
$k=5$	2223	2044	3719	816	3991
Total	2610	2244	4395	1050	4194

Table 1: Number of items collected

Annotation items are based on k contexts (with $k \in \{1, 3, 5\}$)[1] and two words: a target w_t and a distractor w_d. All k sentences contain the target w_t, but not the distractor w_d. Sentences are presented at once, with the target w_t replaced by a blank token. Annotators are then asked which of the two words w_t or w_d they believe was originally present in the k sentences. This task corresponds to a variation on the Cloze test (Taylor, 1953) where annotators see more than one context. An overview of the data volume collected thus far is given in Table 1.

The contexts presented to annotators were preselected from four genres: Wikipedia dumps, books corpora (Gutenberg Project, WikiSource, LiberLiber.it), parliamentary debates (EuroParl, Koehn, 2005; UN Corpus, Ziemski et al., 2016) and subtitles (OpenSubtitles, Lison and Tiedemann, 2016), for a total of 4M sentences.

We consider three strategies to construct word pairs.[2] First, we select items which we expect to be difficult *a priori*: ordinal and cardinal numbers, months, days of the week and colors. Second, we select items that maximize distributional similarity, using word2vec models. Lastly, annotators also had the possibility to suggest pairs of words that they expect to be difficult to distinguish.

To construct our dataset, we collect: the target w_t, the distractor w_d, the k context sentences, whether the annotator correctly selected the target w_t, the time taken to provide an answer, and identifiers for the annotator and the creator of the word pair. Table 2 provides an example item.

3.2 Dataset contents

Fig. 1 displays the overall success rate of annotators; i.e., the percentage of annotations where they were able to select the target word over the distractor. Each subfigure presents a different condition: Subfigure 1a shows results over the full dataset,

[1] Annotators can freely set k, by default, $k = 5$.

[2] Throughout the paper, “word pair” refers to *order-insensitive* word pairs.

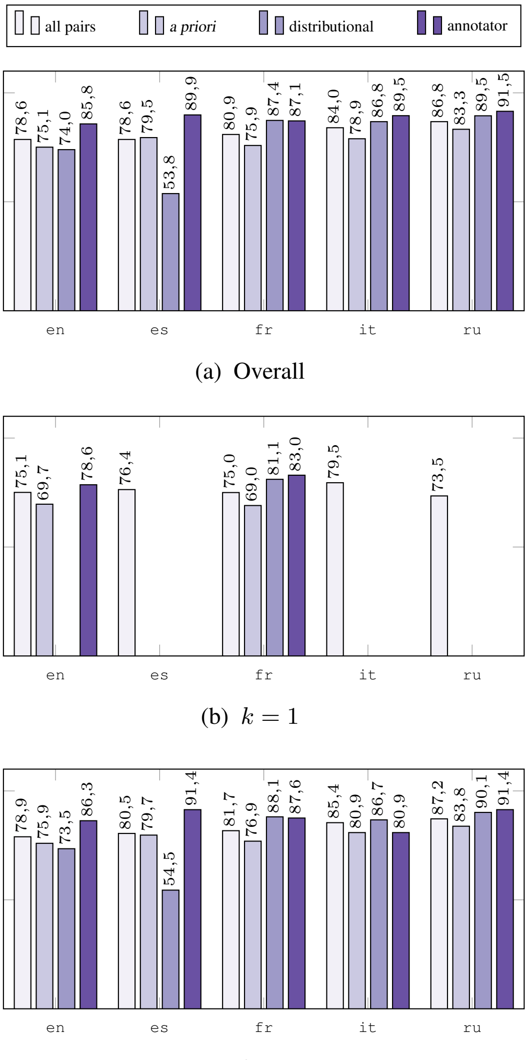
Target:	pleura	**Distractor:**	diaphragm
Correct:	No	**Time:**	35.84 s
Contexts:	best way to dissect the aortic ______. the ______ and pericardium have both been recorded as points of outlet. if the ______ be implicated, greater expansion of the upper and outside portion of the left side of the chest in inspiration takes place.		
Annotator ID:	dYaGLiFsJz8		
Creator ID:	N/A (distributional)		

Table 2: Example annotation item.

whereas Subfigures 1b and 1c display results according to the number of contexts shown to the annotators. We do not include results for $k = 3$, as most groups contained less than 100 items.

If we look at the overall tally (Subfigure 1a), and average across all five languages of our study, we get a success rate of 82%. For all languages, at least 13% of the items considered here have received an incorrect response from human annotators. The overall difficulty can jump to more than 26% if we consider the most challenging setups, where annotators only have access to $k = 1$ sentences (Subfigure 1b). Even in the most informed setup with $k = 5$ (Subfigure 1c), we find that the best language remains below 90% accuracy overall. It is also instructive to compare the strategies used to define word pairs: those suggested by annotators tend to be the easiest of all; whereas *a priori* word pairs tend to be harder than the average case. Lastly, the surprising difficulty for Spanish distributional word pairs comes from the fact that our original Wikipedia sample contain a number of extremely similar sentences all focusing on botany.

Even in the best of cases, annotators select the distractor almost one out of every ten items. This difficulty could be due in part to our methodology: we preprocess the sentences we present to annotators automatically and rely on crowd-sourcing to retrieve human judgments on the distributional substitution task. Nonetheless, it suggests that meaning cannot be entirely retrieved from distribution alone: extralinguistic context is necessary (Searle, 1980; Harnad, 1990; Bender and Koller, 2020). Adding strength to this analysis, we can tentatively identify word pairs that are not reliably distinguished by human annotators. For all languages, roughly 5% of word pairs seen by 5 or more annotators have average success rates at or below chance level. Such pairs are often co-hyponyms: aquarelle 'watercolor' & gouache 'gouache', frambuesa 'raspberry' & fresa 'strawberry', беркут 'golden eagle' & кречет 'gyrfalcon', or baseball & basketball.



(a) Overall

(b) $k = 1$

(c) $k = 5$

Figure 1: Success rates (in %).[3]

4 Success rates

To assess how well DSMs model Eq. (1), we can look at how often they correctly retrieve the target.

4.1 Methodology

We start by considering a 1-gram baseline and a 2-gram baseline. Both are tabulated from corpora comparable to the ones used as basis for our dataset,

[3] Groups with fewer than 100 items not included.

using the same number of sentences from the same four genres in the same proportions. We further ensure that there is no overlap between the corpora we use to compute our n-gram baselines and those used to construct our dataset.

We also include pretrained models based on the BERT architecture of Devlin et al. (2019) or variants thereof. We use BERT (base, uncased) for English, BETO (Cañete et al., 2020) for Spanish, CamemBERT (Martin et al., 2020, base) for French, UmBERTo[4] for Italian and RuRoberta (large)[5] for Russian. We also include word2vec models (Mikolov et al., 2013a) trained on up to 500M sentences from the Oscar dataset (Ortiz Suárez et al., 2019), using `gensim` (Řehůřek and Sojka, 2010) with default hyper-parameters.

As all our models are able to assess the probability of a word in a given context $p(w|c)$, we can extract a prediction by considering whether the probability associated to the target word $p(w_t|c)$ is greater than the probability associated to the distractor in the same context $p(w_d|c)$. In practice, we found it more effective to consider the sum of log probabilities across all contexts $c_1, \ldots, c_k$:

$$\sum_k \log p(w_t|c_k) - \sum_k \log p(w_d|c_k) > 0 \qquad (2)$$

Whenever Eq. (2) holds true, the associated model correctly assigns a higher probability to the target w_t than to the distractor w_d. It should be noted that Eq. (1) and (2) are not strictly equivalent. However, using log-probabilities matches more closely the training objectives of the models we consider: both the MLM objective of BERT-like models and the objective function of word2vec models are implemented as cross-entropy minimization objectives.

BERT models rely on masking word pieces, hence we derive scores in Eq. (2) by masking all the word pieces of the target or distractor, and sum the associated log-probabilities. For word2vec models, we use the explicit probability distribution as derived during training.

4.2 Results

We can now compare the success rate of models to that of humans. To tabulate these scores, we dropped annotations that took too long or too short: we dropped any annotations where the logarithm of the time taken by the annotator was more than one standard deviation apart from the mean, to ensure that we remove the least trustworthy annotations. To avoid likely train/test overlaps, we also remove any sentence originating from Wikipedia. The quantitative impact of this preprocessing is displayed in Table 3.

	`en`	`es`	`fr`	`it`	`ru`
Size	2051	1686	3443	749	3926
Reduct. (%)	78.6	75.1	78.3	71.1	95.0

Table 3: Effects of filtering on dataset size

	`en`	`es`	`fr`	`it`	`ru`
human	83.1	86.9	83.8	89.1	87.8
1-gram	51.9	56.2	53.4	50.8	57.2
2-gram	60.4	71.2	66.0	70.7	60.1
BERTs	75.8	71.6	74.1	76.1	74.4
W2Vs	75.5	77.1	75.5	74.8	72.5

Table 4: Success rates (in %)

Results are described in Table 4. We include the success rates of human annotators on the items we retain for comparison. All models considered yield results above chance level (50%). The various BERT models attain a success rate between 71.6% and 76.1%; the macro-average across all languages reaches 74.4%. This is still below what we see for humans (83.1% to 89.1%, averaging to 86.1%), but systematically above n-gram baselines: the 1-gram average across languages is at 53.9%, the 2-gram average is at 65.7%. The real surprise here is the performance of the word2vec models: despite being designed as purely static embeddings, they achieve a 75.1% average success rate on this contextual task, slightly above what we observe for the BERT models.

4.3 Discussion

This overview of models' success rates highlights that word2vec models can obtain performances comparable to what we observe for BERT-like models. This may be due in part to the size of our training corpora, ranging from 60G (EN) to 90G (RU) of data: this is often (but not always) above what some of the BERT models were trained with.

It is surprising to see that these static embed-

[4] `https://github.com/musixmatchresearch/umberto`

[5] `https://huggingface.co/sberbank-ai/ruRoberta-large`

dings can rival contextual embeddings on a contextual task. This lends depth to previous studies which have found static embeddings to be comparable to contextual embeddings on word-type benchmarks (Vulić et al., 2020; Lenci et al., 2021, a.o.). Nonetheless we still observe a gap between these models and human performance in our dataset.

5 Comparing human and model behaviors

Section 4 has given us a quantitative estimate of the performance of our DSMs. We now turn to assessing whether they can be construed as models of the linguistic behavior of our annotators.

5.1 Binary classification approach

The first approach we consider is to reframe this question as a binary classification problem. Let us assume our models are perfect linguistic models of human capabilities: if so, we would expect them to match human failure with failure. In other words, any incorrect annotation item should correspond to a negative score, as assessed by Eq. (2).

Hence we could consider human behavior as the "gold standard" that a model of human linguistic capabilities would try to match. By assessing how our models perform on this binary classification task, we are able to surmise whether their behavior matches that of human—are they puzzled by sentences humans got wrong? Are they confident with sentences humans got right? To answer this question, we can use standard binary classification tools. More specifically, we turn to Matthews' correlation coefficient (MCC) to see whether model predictions match with human behavior.

	en	es	fr	it	ru
1-gram	0.157	0.158	0.158	0.119	0.177
2-gram	0.156	0.211	0.200	0.193	0.143
BERTs	0.208	0.178	0.150	0.077	0.230
W2Vs	0.135	0.185	0.170	0.122	0.199

Table 5: Matthews' correlation coefficient

Results are shown in Table 5. The difference between n-gram baselines and distributional semantics models that clearly emerged from Table 4 is no longer present. In our three Romance languages, the 2-gram baseline yields a higher correlation coefficient than both word2vec and BERT. In English, the word2vec model is found to yield the lowest MCC; in French and Italian, the CamemBERT and UmBERTo models yield the lowest MCC.

It is hard to argue that the distributional models correlate more with human behavior than the n-gram baselines. We can stress that all the models we tested yielded a positive correlation, which suggests that the behavior of our DSMs is not unrelated to humans. Yet the mistakes and successes of our DSM models overall do not necessarily align with that of human annotators, as expressed in our dataset.

5.2 Ranking approach

There are two obvious caveats that one can think of in the methodology we adopted in Subsection 5.1. First, it pits model efficiency against linguistic validity: a model cannot be both always correct and match human failures with failures of its own. It also relies entirely on treating human annotations as a gold standard—even when annotators have selected the wrong answer.

The simplest way to address both of these concerns is to depart from the binary approach, and see instead whether human uncertainty is matched with lower scores from the models. In principle, a model could always choose the right answer, but lower its score for difficult items—i.e., those annotators struggle with. Considering the uncertainty of our annotators also entails that we factor in how confident they are in their judgments.

This approach requires some sort of measurement of annotator uncertainty, beyond the binary annotations we have exploited thus far. To that end, we focus on the time it takes an annotator to answer a question. We can expect that an annotation item that is easy to judge should take less time than an item requiring careful consideration. Furthermore, as annotators should have no difficulty to correctly guess easier items, we expect that the time taken to answer correctly should be less than the time taken to answer incorrectly. We also consider normalizing the time taken by the number of sentences (i.e., k), the number of words across all sentences, or the number of characters across all sentences. Our reasoning is that the time taken by an annotator also depends on how much text they have to read.

In Table 6, we consider various time indicators: either the raw log seconds taken,[6] or variants normalized by some measure of the length of the an-

[6] The logarithmic transformation shifts data distribution from a power law to an almost normal distribution.

Norm.	en	es	fr	it	ru
none	–	–	–	0.417	0.390
sents.	–	0.458	0.449	0.373	0.376
words	0.447	0.385	0.454	0.421	0.452
chars.	0.462	0.395	0.455	0.417	0.459

Table 6: Effect size from Mann-Whitney U-tests for log time taken when answering correctly vs. incorrectly

	en	es	fr	it	ru
1-gram	0.149	0.115	0.132	0.163	0.147
2-gram	0.119	0.150	0.228	0.267	0.146
BERTs	0.225	0.152	0.204	0.218	0.258
W2Vs	0.145	0.196	0.244	0.165	0.248

Table 7: Spearman correlations of model scores and time-weighted human judgments

notation item. Measurements are done using a Mann-Whitney U-test, to see whether the distributions of time indicators differ between correctly annotated items and incorrectly annotated items: we then compute the common-language effect size, i.e., the U-statistic divided by the maximum value it could assume. Here, a lesser value of ρ entails a greater certainty that the incorrect annotations take longer than the correct annotations. Statistically insignificant effect sizes are not reported.

Table 6 shows that raw time measurement is not always significant, but when factoring in the length of an annotation item we detect that annotators take longer when they answer incorrectly than correctly. This is consistent with time being an indicator of uncertainty. We note that the best length normalization differs across languages; explaining what factors drive this difference is beyond our scope.

Having found a way to quantify uncertainty, we can now include it in our original annotations. We reweight human annotations to factor in time, such that highly confident correct answers lie at one end of the spectrum, and highly confident wrong answers lie at the other end of the spectrum. This also ensures that we match as closely as possible how we derive scores from our models. Technically, we reweight human judgments as follows:

$$(\max s^* - s) \times \begin{cases} +1 & \textit{if correct} \\ -1 & \textit{otherwise} \end{cases} \tag{3}$$

where s is the length-normalized time indicator $\log t/N$, with N either the number of sentences (for FR, IT, RU) or words (for EN and ES), and $\max s^*$ is the maximum value observed for s across all annotations for that language.

As we have two related series of continuous measurements, we can apply a simple correlation metric, such as Spearman's ρ, between time-weighted annotator responses and model scores. This is shown in Table 7. In English, French and Italian, either or both DSMs yield a lower correlation than what we observe for n-grams, while in Spanish the margin between BETO model and the 2-gram baseline is less than 0.002. Only in Russian do we find a sharp distinction between DSMs and n-grams. Overall, allthough correlation scores are always positive, they remain fairly low ($\rho < 0.27$).

5.3 Discussion

In all, while the models do display some degree of performance (as shown in Section 4), neither the sort of mistakes they do (Subsection 5.1) nor the confidence in their answer (Subsection 5.2) matches human behavior closely. In many cases, distinguishing DSMs from n-gram baselines can prove very arduous. In other words, this experiment suggests that efficient models do not necessarily reflect human linguistic behavior.

6 Manipulating the distributional hypothesis

We have focused thus far on whether DSMs model human behavior. We could instead reverse the setup: does a low score from a DSM entail a greater hesitation from the human annotators?

6.1 Methodology

This time, we select sentences that either maximize or minimize Eq. (2), and see how human annotators fare on these contexts and how confident they are in their answers. We start by selecting the most extreme word pairs in terms of average success rate. For each word pair, we select a random sample of up to 10 000 sentences from the original sentence corpora detailed in Section 3, and rank them according to the score a BERT-like model would give them following Eq. (2). We then restrict our random sample to the five sentences with the lowest scores and the five sentences with the highest scores. We ensure that sentences are uniquely associated to word pairs: if some sentence c_p is among the ten items chosen for a pair $\langle w_t^n, w_d^n \rangle$, then it will not be chosen for any other pair $\langle w_t^m, w_d^m \rangle$.

We then hire native speakers to annotate this data. Unlike the main dataset, we present one context at a time, since we are interested in the ability of a DSM to rank specific contexts. We also ask annotators to express themselves using a five-point Likert scale, ranging from high confidence in the target to high confidence in the distractor. All items are doubly annotated. In total, we gathered 500 items for English, 432 in Spanish and 500 in French.

6.2 Results

Fig. 2 pits the scores derived from BERT on the y-axis against the corresponding Likert scale annotations, for each language; the heatmap in the middle of each picture displays how the two distributions coincide. These illustrations clearly show that both annotators and the BERT models behave differently across languages. However there are similarities: in all three languages, annotators match high BERT scores with a strong preference for the target. In French (Subfigure 2c), annotators and BERT seem to closely match in their behavior: a neutral response is elicited when the score is low, whereas a confident preference for the target corresponds to a high score. In the other two languages, low scores are spread out across the scale. In Spanish (Subfigure 2b), scores around zero elicit a neutral response, but scores below zero do not seem associated to a specific response. In English (Subfigure 2a), we see a linear trend: the very lowest BERT scores tend to elicit a strong preference for the distractor.

To provide a more quantitative outlook, we turn to a dominance analysis (Budescu, 1993) to determine the factor most closely related to annotators' behavior. Dominance analysis consists in learning a simple linear regression, computing the associated r^2 to measure its fitness, and deriving the proportion of this r^2 that can be imparted on each predictor. Here, we predict the average Likert score from the original average success rate for the word pair and the BERT score for the context. This allows us to compare these two metrics as competing explanations for the collected Likert annotations. We also include other likely predictors: the original source of the sentence shown to the Likert annotators, the time taken for the word pair and the frequency of target and distractor.

Results are presented in Table 8. The r^2 of each linear regression is given in the last column; columns 1 through 5 detail the proportion of this r^2 imparted on each predictor (in %). The fitness of the regression, as measured by r^2 scores, suggests that more than half of the variance in annotations is not explained by a simple linear relation between predictors. This is especially striking in Spanish, where the r^2 score is at 0.21. Yet all models consistently rank the BERT score as the most important predictor. The French and English both impart more than 75% of the explained variance on BERT scores and 15% to 20% to average success rates and time taken on the previous dataset. The Spanish model emphasizes more the frequency of the target and distractor (21.24%) and the corpora from which the presented context originate (11.16%).

	BERT	succ.	time	freq.	src.	r^2
en	76.86	8.97	9.01	3.80	1.36	0.28
es	59.18	4.22	4.20	21.24	11.16	0.21
fr	81.38	7.15	7.30	1.94	2.23	0.44

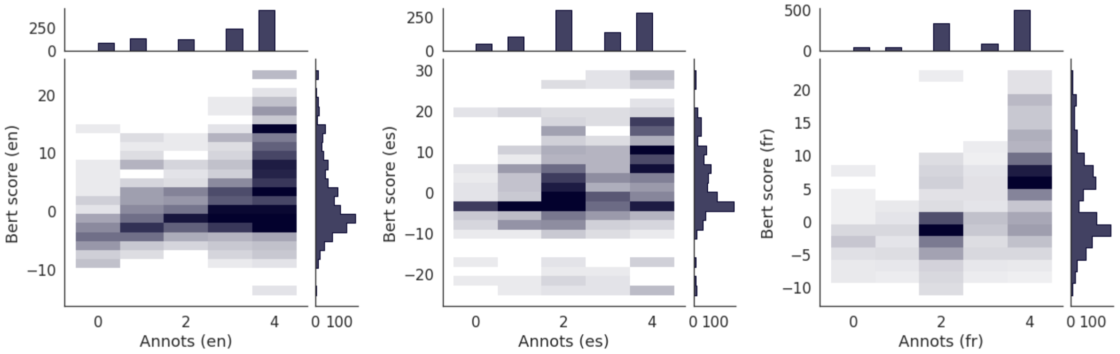
Table 8: Prop. of r^2 explained by type of predictor

6.3 Discussion

In short, this last experiment stresses that in specific conditions BERT models can prove to be useful tools to manipulate the distributional hypothesis. This is especially visible on the case of the French data, which yields the most obvious bimodal distribution (Subfigure 2c), the highest r^2, and the largest proportion of variance explained by the BERT model scores. These elements suggest that the CamemBERT model was able to select sentences that strongly cued the target.

On the other hand, we are not able to reliably find French contexts that elicit a strong preference for the distractor. This is something we only tentatively observe for English, where paradoxically the dominance analysis suggests that our current predictors are less well-suited to explain the phenomena we recorded ($r^2 = 0.28$). This is in line with previous experiments: while high BERT scores translate into a confident preference for the target, much remains to be done in order to accurately depict the full breadth of human behaviors, ranging from strong preferences in the distractor and accurately depicting less confident human judgments.

Opposite to this is Spanish: the lowest scores from BETO do not bias the annotators towards neutral or negative responses. The main reason of this difference is unclear: the quality of the sentences presented to annotators might play a role, but so



(a) en

(b) es

(c) fr

Figure 2: Word pair difficulty compared to BERT scores

might the quality of BETO. We also find a much lower inter-annotator agreement for this language: the Pearson r correlation coefficient for our two Spanish annotators is of only 0.11, compared to the 0.59 we observe for English or the 0.74 for French.

Improvements could be made on our analyses: one could use as predictors the average success rate and the average time taken restricted to items with $k = 1$ contexts, as these would be more representative. We leave this to future investigations, as we haven't collected enough data to establish such baselines (cf. Table 1). Another point we leave for future study is the number of datapoints in our original datasets: some predictors are derived from them (average time and success rate) and it is unclear how size discrepancy impacts them.

7 Related Works

Distributional semantics was first suggested by Harris (1954). A wealth of work has focused on characterizing it (Sahlgren, 2008; Lenci, 2018; Boleda, 2020; Emerson, 2020; Rogers et al., 2020). as well as its limitations (Miller, 1967; Westera and Boleda, 2019); related works point out the hardships of semantically grounding distributional representations (Searle, 1980; Jackson, 1982; Harnad, 1990; Bender and Koller, 2020). Other researchers concern themselves with the empirical foundations of the distributional hypothesis, often from the point of view of psycholinguistics (Rubenstein and Goodenough, 1965; Mandera et al., 2017).

Also relevant to our work are studies that attempt to evaluate the performance or quality of distributional models of semantics. These can generally be grouped in two categories: works introducing evaluation procedures or benchmarks (Mikolov et al., 2013b; Wang et al., 2018; Ferret, 2021) and works proposing large surveys across multiple models (Vulić et al., 2020; Lenci et al., 2021).

8 Conclusions

We broached the question of how to quantify our expectations with respect to distributional semantics and DSMs, using the distributional substitution task. In short, distributional information would allow humans to retrieve about 82% of pairwise meaning distinctions. In contrast, embedding models like BERT and word2vec would only reach 75% accuracy (Section 4), and how they achieve these performances begs the question of whether we should consider them as models of distributional semantics, seeing that they do not seem to match human annotators' judgments (Section 5). In specific circumstances BERT-like models can however be used as tools to manipulate the distributional hypothesis (Section 6). More research is needed before any firm conclusion can be reached, given the limits of the dataset we currently have (Section 3: 5 languages, about 15 000 items).

Taking a more linguistic-oriented point of view, our experiments suggest that much remains to be done before we can confidently say that modern NLP models can be construed as linguistically valuable models of distributional semantics. This opens a number of perspectives for future research: how would this translate to other languages, especially non-European ones? What is required of DSMs for them to accurately describe human behavior? Which factors are necessary to model human behavior on the distributional substitution task?

Ethical impact

We propose a dataset derived from human judgments. The data collection process has been approved by the relevant instances within the research structures of the authors. As such, a number of considerations apply.

The authors of this work are based in an area where GDPR laws apply.[7] The data was therefore collected in a manner that guarantees the anonymity of the annotators; in particular, all identifiers associated to annotators in the dataset correspond to randomly generated strings. Time of annotation creation, geographic location of annotators, contact information of annotators are not provided in the released dataset.

The data was collected through a gamified online platform. As such, annotators of the base dataset described in Section 3 were not financially compensated for their work; and the whole collection project was constructed to ensure this voluntary work is conducted in as ethical a manner as possible. In particular:

- Participation to the annotation platform requires informed consent of how their in-game behavior will be used. As such, the gamified platform was systematically advertised as a research project.
- Annotators are free to opt out of the task at any moment. Annotators retain the right to have all records of their activity automatically destroyed at any time.
- Annotators are provided with the means to convey feedback.

The contexts presented to annotators were automatically collected from large corpora. Hence, they may contain unwanted biases discriminating against a specific race, sex, gender, ethnicity, age, religion or any other social criterion. Such contexts have not been removed: (a) the dataset was constructed with the intent of collecting human judgments on the sort of data presented to distributional model, including socially biased data; (b) a manual evaluation of a sample of 100 contexts did not reveal any downright problematic sentences, although we found one sentence expressing bias against the handicapped, and three sentences displaying female characters in stereotypically gendered situations. Also note that contexts were drawn from sources such as Europarl, which contains a majority of male speakers, or Wikipedia, which is known for its high proportion of white, male, college-educated writers.

Word pairs are constructed through multiple strategies, which include automatic means and crowd-sourced propositions. It is therefore possible that these contain unwanted associations that could reflect systemic biases. Although we have not identified any such item in our analyses , a more in-depth study is required.

The dataset proposed in this paper is highly Eurocentric. All the languages we propose correspond to European countries. This choice stems from practical considerations—namely due to the availability of experts and data for these languages. We nonetheless ensured that the gamified platform itself would be easily transposable to other languages, by including website translation mechanisms and externalizing language processing pipelines.

Lastly, the authors stress that they make no claim with respect to measuring social impact through their proposed dataset. It is important to acknowledge that models that yield interesting results and high performances on the present dataset may very well display unwanted biases. The present dataset is not constructed to assess such aspects of an NLP system.

[7] https://gdpr.eu/what-is-gdpr/

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 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, Baltimore, Maryland. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

David V. Budescu. 1993. Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, 114(3):542–551.

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

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

Guy Emerson. 2020. What are the goals of distributional semantics? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7436–7453, Online. Association for Computational Linguistics.

Olivier Ferret. 2021. Using distributional principles for the semantic study of contextual language models.

Juan Luis Gastaldi. 2021. Why can computers understand natural language? *Philosophy & Technology*, 34(1):149–214.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335 – 346.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Frank Jackson. 1982. Epiphenomenal qualia. *Philosophical Quarterly*, 32(April):127–136.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2021. A comprehensive comparative evaluation and analysis of distributional semantic models.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

George Miller. 1967. Empirical methods in the study of semantics. *Journeys in Science: Small Steps – Great Strides*, pages 51–73.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Radim Řehůřek and Petr Sojka. 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. ELRA. http://is.muni.cz/publication/884893/en.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Herbert Rubenstein and John Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8:627–633.

Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.

Ferdinand de Saussure. 1916. *Cours de linguistique générale*, 1995 edition. Payot & Rivage, Paris.

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–424.

Wilson Taylor. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Matthijs Westera and Gemma Boleda. 2019. Don't blame distributional semantics if it can't do entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 120–133, Gothenburg, Sweden. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).