Comparing Static and Contextual Distributional Semantic Models on Intrinsic Tasks: An Evaluation on Mandarin Chinese Datasets

Pranav A[1], Yan Cong[2], Emmanuele Chersoni[3], Yu-Yin Hsu[3], Alessandro Lenci[4]
Dayta AI[1], Purdue University[2], The Hong Kong Polytechnic University[3], University of Pisa[4]
cs.pranav.a@gmail.com, cong4@purdue.edu, emmanuelechersoni@gmail.com,
yu-yin.hsu@polyu.edu.hk, alessandro.lenci@unipi.it

Abstract

The field of Distributional Semantics has recently undergone important changes, with the contextual representations produced by Transformers taking the place of static word embeddings models. Noticeably, previous studies comparing the two types of vectors have only focused on the English language and a limited number of models. In our study, we present a comparative evaluation of static and contextualized distributional models for Mandarin Chinese, focusing on a range of intrinsic tasks. Our results reveal that static models remain stronger for some of the classical tasks that consider word meaning independent of context, while contextualized models excel in identifying semantic relations between word pairs and in the categorization of words into abstract semantic classes. The code and datasets are available at https://github.com/pranav-ust/chinese-dsm.

Keywords: Distributional Semantic Models, Mandarin Chinese, Semantic Similarity, Transformers

1. Introduction

Distributional Semantics, the mainstream approach to representing lexical meaning in Computational Linguistics, assumes that words appearing in similar contexts have similar meanings (Lenci, 2008; Turney and Pantel, 2010; Lenci and Sahlgren, 2023). Distributional Semantic Models (DSMs) have gained great success in the NLP community, as they provide researchers with theoretical and computational tools to derive data-driven semantic representations from large text corpora. Moreover, DSMs have become extremely popular in cognitive sciences, as the distributional estimation of semantic similarity has demonstrated a good fit to human data across various psycholinguistic tasks, including synonymy identification, generation of word associations, and semantic priming (Bullinaria and Levy, 2012; Mandera et al., 2017).

The field of Distributional Semantics has experienced a significant revolution with the introduction of language models based on the Transformer architecture (Vaswani et al., 2017). While a common criticism of traditional DSMs has been that they produce only a single, global semantic representation for each word type, disregarding its context-dependent semantic shifts, the vectors generated by Transformer language models are fully contextualized, associating each token in a sentence with a representation that is a function of the activation states of the network (Liu et al., 2020). The difference between **static** and **contextual** DSMs can be seen as analogous to the distinction between prototype-based and exemplar-based models of concept representation in cognitive psychology (Murphy, 2004; Nosofsky, 2013): On the one hand, we have a single prototype abstracted from multiple encounters with the same entity, leading to the exclusion of more idiosyncratic features. On the other hand, we have multiple exemplars, corresponding to different instances of a concept in specific contexts.

However, despite the context-sensitive nature of concepts, several cognitive phenomena seem to involve type-level representations, rather than token-level ones, such as hierarchical structure, basic-level advantage, and reasoning, among others (Murphy, 2016). Similarly, when we evaluate DSMs on many tasks that rely on the human intuition of similarity (e.g., word associations, identification of semantic relations etc.), we need a word representation at the type-level rather than the token-level. This is why recent literature has engaged in reducing the contextual embeddings from the Transformers to type-level representations, which can then be evaluated in traditional Distributional Semantics tasks (Bommasani et al., 2020; Chronis and Erk, 2020; Lenci et al., 2023). It is worth noting that such efforts often focused on a limited number of models (e.g., BERT) and specific tasks (e.g., similarity/relatedness estimation), and, above all, the evaluation of these has primarily been carried out using only English data. Although some of the previous research reported positive results for Transformers as classical DSMs, with correlations to human judgments that are similar or higher than the ones achieved with static embeddings, an important question arises regarding the generalizability of these findings to languages other than English.

In this regard, Chinese language provides a very relevant case study, because the definition of words in Chinese is not trivial (Duanmu, 2017),

LREC-COLING 2024, pages 3610–3627

and studies both in the psycholinguistic (Tsai and McConkie, 2003; Bai et al., 2008) and the NLP literature (Li et al., 2019; Wan, 2021) have suggested that characters, rather than words, are the fundamental units of Chinese language processing. Given that, to the best of our knowledge, most of the Transformer language models for Chinese are based on character-level tokenization (Si et al., 2023), it is especially interesting to assess the performance of their word-level representations for classical distributional tasks.

In our paper, we present a systematic evaluation of DSMs for Mandarin Chinese. It compares traditional static vectors with the contextualized representations generated by Transformers. Specifically, we evaluated the performance of some of the most popular Transformer models for Chinese, including BERT and GPT-2, on a range of evaluation tasks, such as similarity estimation, word associations, analogies, clustering, and semantic relations. We found that static models are consistently strong baselines, and clearly outperform contextualized vectors in tasks like semantic analogies, while the latter are the winners in semantic relation identification, and in the categorization of words in abstract semantic classes.

2. Related Work

2.1. Static and Contextual DSMs

The first generation of distributional models dates back to the 90s. During this period, the so-called *count models* were built in an unsupervised way: in order to build the semantic representation of a word, its co-occurrences with linguistic contexts were first recorded and counted. The context could consist of words co-occurring within a word window of a fixed size (Lund and Burgess, 1996), of words in a syntactic relation with the target (Padó and Lapata, 2007; Baroni and Lenci, 2010; Chersoni et al., 2016; Gamallo, 2019), or entire documents (Landauer and Dumais, 1997; Griffiths et al., 2007). In most cases, further transformations were applied to the co-occurrence matrices, such as using mutual information measures to weigh the counts (Church and Hanks, 1990; Evert, 2004), or techniques for dimensionality reduction (Landauer and Dumais, 1997; Bullinaria and Levy, 2012). The cosine similarity between the vectors of two words was the most commonly used metric. Depending on the model parameters, semantic **similarity** refers to whether two words describe similar concepts, like *tea* and *coffee*, while their **relatedness** captures the connection between two words that may describe concepts that are dissimilar but still conceptually related, e.g. *coffee* and *cup* (Hill et al., 2015).

DSMs underwent a revolution in the Deep Learning era, especially thanks to the introduction of frameworks like Word2Vec (Mikolov et al., 2013). The so-called *predict models* are neural networks that directly generate low-dimensional, dense vectors by being trained as language models that learn to predict the contexts of a target lexical item. Those dense vectors, also known as **word embeddings**, quickly replaced the traditional count models; this shift was facilitated by the availability of tools that simplified the model training process for researchers, and by their superior performance in semantic similarity task (Baroni et al., 2014) (but cf. Levy et al. (2015); Sahlgren and Lenci (2016) for alternative evaluation outcomes).

Both count models and predict models share an important feature: They both build a single, stable representation for each word type in the training corpus. In the most recent generation of embeddings, a different approach is taken; each word token in an input sentence context gets a unique representation. In models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), word representation relies on a multi-layer neural network (a bi-LSTM or a Transformer), word vectors are generated as its internal activation states, and they differ depending on the sentence contexts. Therefore, such representations are referred to as *contextualized embeddings*, in contrast to the *static* embeddings produced by the earlier generation models. Another important difference is that, with static vectors, intrinsic evaluation was the most common way to test the models (e.g., by measuring the correlation of similarity with human ratings, or by solving synonymy tests or analogy tasks), while contextual vectors are mostly used as inputs for downstream tasks (extrinsic evaluation).

Some recent work has introduced methods to obtain type-level vectors from contextual models, in order to compare their performance with traditional intrinsic methods. For example, Bommasani et al. (2020) proposed to obtain type-level vectors by pooling the contextual token vectors generated by different Transformer models and compared them with Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) across different similarity and relatedness benchmarks. They showed that the performance of the contextual vectors varied a lot across layers, but the best layers consistently outperformed the two static models. Chronis and Erk (2020) applied K-means clustering to contextual BERT embeddings, and tackled similarity/relatedness tasks by measuring the similarities between the closest clusters of the target words (this is equivalent to comparing only their most similar senses), and reported that the vectors from the middle layers are better at modeling similarity, while the ones from the later lay-

ers are better for relatedness. Finally, Lenci et al. (2023) compared static embedding models with contextual ones on multiple intrinsic and extrinsic tasks, using a similar method in Bommasani et al. (2020) to obtain type-level vectors from BERT. They found that properly optimized static vectors outperformed contextual vectors in most intrinsic tasks. However, all these studies were conducted exclusively on English data.

2.2. DSMs for Chinese

Differently from alphabetic languages, the writing system of Chinese is logographic, and a Chinese character can be a standalone word or as part of a polysyllabic word. In Chinese words, both characters and subcharacters (i.e., radicals) contain semantic information, which previous research has tried to leverage to enhance the performance of Chinese DSMs. Therefore, several scholars proposed methods to jointly learn word and character embeddings (Chen et al., 2015), or to learn word, character, and sub-character embeddings all together (Yin et al., 2016; Yu et al., 2017; Sun et al., 2019), although the evaluation benchmark is generally focusing on word-level tasks, e.g., similarity and analogy (Chen and Ma, 2018; Li et al., 2018; Huang et al., 2019).

Alongside the Transformer revolution in NLP, pre-trained language models have been developed also for Chinese, making it possible to extract contextualized representations, e.g., GPT-2 (Zhao et al., 2019), BERT, RoBERTa, XLNet (Cui et al., 2020, 2021), DeBERTa (Zhang et al., 2022), LLaMA, and Alpaca (Cui et al., 2023). The basic unit of the vocabulary, for Chinese models, is typically the character, although there are exceptions making use of sub-character tokenization (Si et al., 2023). When evaluating Chinese Transformers as DSMs, the issue might become particularly relevant in cases where the target words are not included in the model's vocabulary, and the embeddings will have to be composed by averaging their character tokens. To the best of our knowledge, our work is the first one to compare static and contextual Chinese DSMs on intrinsic tasks.

3. Experimental Settings

3.1. Static DSMs

We used the Chinese word embeddings by Li et al. (2018); Qiu et al. (2018) based on the Skip Gram architecture. The models have been trained on the Mixed-Corpus, a Chinese corpus combining the Baidu Encyclopedia, Wikipedia, People's Daily News, and other corpora, encompassing over 4 billions tokens. These Chinese models are of particular interest as they have been trained with different context features, i.e., words, n-grams, characters, or a combination of the latter two. Therefore, we can use them to test the effects of different contextual features on the overall performance. Table 1 summarizes the models and their features.

Model	Context Features
Skip Gram	Words
Skip Gram N	Words, Ngrams
Skip Gram C	Words, Characters
Skip Gram N+C	Words, Ngrams, Characters

Table 1: Summary of static embedding models.

3.2. Contextual DSMs

Notoriously, contextual DSMs output a different word vector for each sentence context in which a word is found. Therefore, we have sampled n sentences (at least $n = 10$, and $max = 100$)[1] containing the target dataset words, and then we used the Transformers library[2] to extract contextual vectors from each Transformer model. The target words with fewer than 10 occurrences in the corpus were discarded. The vectors were then averaged using mean pooling to create a single type-level representation for each word, following the procedure in Lenci et al. (2023). If the target word is not found in the Transformer's vocabulary, we also average the embeddings of the subwords composing it.

We selected three Transformer models for Chinese: **GPT-2 Zh** (Zhao et al., 2019), based on GPT-2 Base (Radford et al., 2019); **BERT Base Zh** (Cui et al., 2021), based on BERT Base (Devlin et al., 2019); **DeBERTa Zh** (Zhang et al., 2022), based on DeBERTa Base (He et al., 2020). We chose those models because they are available in Chinese in their Base version with 12 layers, which makes it easier to compare their layer-wise performance. Following Lenci et al. (2023), we extracted the word embeddings by aggregating different sets of layers, in order to understand what are the best layers for each task:

- *last*, we used the embeddings of the last layer;
- *F4*, we used the average of the embeddings of the first four layers (1-4);
- *M4*, we used the average of the embeddings of the middle four layers (5-8);

[1] Vulić et al. (2020) showed that improvements by sampling a bigger number of occurrences per word are marginal. Similarly to Lenci et al. (2023), we limited the sentence length to be between 4 and 21 tokens (or characters, in Chinese language models).

[2] https://huggingface.co/docs/transformers/index.

- *L4*, we used the average of the embeddings of the last four layers (9-12).

The layer-wise analysis is noteworthy. It has been shown that Transformers models tend to encode morphological and syntactic information in early layers, while capturing more abstract and context-specific information (e.g., semantic) in later layers (Tenney et al., 2019), and that the degree of contextualization increases with the layers.

3.3. Similarity Metrics

Vector cosine is the standard similarity metric in Distributional Semantics (Turney and Pantel, 2010). However, recent studies did not recommend the usage of this metric to compare Transformer models, because of its sensitivity to the anisotropy of their contextualized vector spaces (Ethayarajh, 2019; Timkey and van Schijndel, 2021). On the other hand, metrics based on the rank of vector dimensions were proved to be competitive against cosine, and more correlated with human judgements (Santus et al., 2016a,b, 2018; Zhelezniak et al., 2019; Timkey and van Schijndel, 2021). Therefore, we also report results obtained with the **Spearman correlation** between vectors.

3.4. Benchmarks

As in traditional DSM evaluation, we focus on intrinsic tasks: the superiority of contextualized models in extrinsic settings is a result consistently accepted in the literature (Lenci et al., 2023).

We selected five tasks that address different aspects of word-level semantics, also based on the availability of benchmarks for Mandarin Chinese: Similarity Estimation, Word Associations, Semantic Analogies, Identification of Semantic Relations and Semantic Clustering.

3.4.1. Similarity Estimation

We evaluated our models on the **COS960** dataset by Huang et al. (2019). The dataset includes a total of 960 Mandarin Chinese word pairs (480 nouns, 240 verbs and 240 adjectives), each of them rated for similarity by 30 native speakers on a scale ranging from 0 (not similar at all) to 4 (very similar). We take the average similarity score for each word pair and we measure the *Pearson* and the *Spearman* correlation with the similarity scores computed by the distributional models.

3.4.2. Word Associations

We use the **FAST-zh** dataset, derived from the Small World of Words project data[3], for Chinese word associations (Kwong et al., 2022). FAST-zh consists of 300 tuples of four words (see Table 3 for an example). In each tuple, there is a stimulus word, a first associate (i.e., the word that was produced as the response to the stimulus by most of the speakers), and a higher associate (i.e., a word ranked the n-th, but not the most frequent, typically a response produced by just two subjects), and a random word without any association with the stimulus. We proposed two different evaluation tasks, inspired by Evert and Lapesa (2021) on word association data. The first is **Multiple Choice**: given a tuple, a model should assign a similarity score to the pairs formed by the stimulus word and each one of the other words, getting a hit every time the stimulus-first pair has the highest score. This task is evaluated with *Accuracy* (i.e., the fraction of correct responses out of the total of dataset tuples).

The second task is **Open Access Vocabulary**: for each stimulus in the dataset, a word embedding model has to retrieve the right FIRST associate out of a list of candidates including all the other FIRST associates in the dataset (e.g., for each language, there will be around 300 candidates). For each stimulus, we measure the similarity with all the other FIRST associates in the dataset and we compile a ranking based on decreasing similarity values. *Mean Rank* is the task metric: we compute the average rank of the right first associate for each stimulus (see Equation 1). For $rank_i$, we use the index of instance i if the right first associate is in the top 3 of the rank, and 4 otherwise.[4]

$$MeanRank = \frac{1}{n} * \sum_{i=1}^{n} rank_i \quad (1)$$

For this metric, the lower the score, the better it is, as we want the models to push the right first associates as close as possible to rank 1.

Word Pair	Translation	Score
共存 - 溶合	co-exist - integrate	0.6
窥视 - 窥探	peak-observe - peak-detect	3.86
船头 - 船尾	ship-head - ship-tail	0.93

Table 2: Examples of word pairs and average human scores in the COS960 dataset.

Stimulus	First	Higher	Random
活 (live)	死 (die)	人生 (life)	人才 (talent, talented person)

Table 3: Example of a tuple from the FAST-Zh dataset.

[3] https://smallworldofwords.org/en/project/home

[4] This setting was adopted in the SemEval task 2018 on hypernymy discovery (Camacho-Collados et al., 2018) to avoid penalizing too much systems with a small number of outputs far away from the first ranking spots.

Analogy	Target Word
中国: 北京 = 意大利:??? (China:Beijing=Italy:???)	罗马 (Rome)

Table 4: Example of a semantic analogy for the country-capital relation from the CA8 dataset.

3.4.3. Semantic Analogies

When the Word2Vec model was originally introduced (Mikolov et al., 2013), one of the most cited findings was the success of word embeddings in **analogical reasoning**. Given a vector space, analogies such as $man : king \sim woman :?$ could be automatically solved by looking for the word whose vector had the highest similarity (cosine or Spearman) with $king - man + woman$ (with $queen$ being the target word). The **CA8** dataset for analyzing morphological and semantic regularities for Chinese was presented and evaluated by Li et al. (2018) and Qiu et al. (2018).

In our work, we focus on the semantic subset of their data, composed of $7,363$ semantic questions representing 28 different types of relations (e.g., country-capital, dynasty-emperor, and book-author). We evaluated all our models using *Accuracy* as the percentage of the correct answer. To make the results comparable between static and contextual models, we have limited the search space for the closest vector to the vocabulary of the dataset itself.[5]

3.4.4. Identification of Semantic Relations

A commonly-cited shortcoming of DSMs is that measuring proximity in vector spaces only provides an underspecified notion of semantic similarity/relatedness, whereas there are different ways in which words can be semantically related (Lenci and Sahlgren, 2023). The problem of discriminating the semantic relations between nominals (e.g., synonyms, hypernyms, meronyms, etc.) received a lot of attention in the literature (Baroni and Lenci, 2011; Xiang et al., 2020; Schulte Im Walde, 2020), leading to the publication of datasets in several languages, including Mandarin Chinese. Inspired by the evaluation dataset for English (Santus et al., 2015), Liu et al. (2019) introduced **EVALution-MAN**, a dataset for evaluating the identification of semantic relations in Chinese.

The dataset contains $3,923$ word pairs, covering the relations of *synonymy*, *hypernymy* and *antonymy*. To introduce some noise and make the dataset more challenging, an equal number of *random* pairs was generated, for a total of $7,846$ items. Since the original data are in Traditional Chinese characters, we first converted them to Simplified Chinese during the preprocessing phase.

Word Pair	Translation	Rel.
不僅僅 - 不單單	not only - not just	syno.
海獅 - 水中生物	sea lion - marine animals	hyper.
男性 - 女性	male - female	anto.
新光 - 勸告	new light - advice	random

Table 5: An example of semantic *relata* for each relation in EVALution-MAN dataset.

DSMs for Chinese can be evaluated on this dataset in an unsupervised fashion for a specific semantic relation: similarity metrics can be computed for each pair, and models can be assessed in terms of *Average Precision* (AP) (Kotlerman et al., 2010). Specifically, given the list of the dataset pairs sorted in a decreasing order for the model/similarity metric, AP measures the extent to which the most similar pairs belong to the target semantic relation. If $AP = 1$, all the instances of a given semantic relation in the dataset are at the top of the ranking, while if $AP = 0$ all the instances are at the bottom.

We first evaluate our distributional models for their capacity of discriminating between related and unrelated words, considering synonyms, hypernyms and antonyms as members of a **related** target class. We expect most systems to be able to put the **related** words to the top of the distributional similarity ranking, and the random pairs at the bottom. Then, we also evaluate how good they are in identifying genuine semantic similarity (Hill et al., 2015), and in this case we consider **synonymy** as our target class. Notice that this latter task should be much more difficult, because antonyms and hypernyms are also likely to have highly similar vector representations.

3.4.5. Semantic Clustering

If word embeddings represent word semantics accurately, we expect them to group together in coherent regions of the semantic space. With this goal, we used two datasets from the recent psycholinguistic literature on Mandarin Chinese.

The **Zhong22** dataset (Zhong et al., 2022) includes 664 nouns annotated with their sensorimotor associations and a wide range of psycholinguistic variables. Moreover, they have been annotated with the "abstract" and "concrete" classes. The Chinese Binder norms (**Binder-zh**) (Qiu et al., 2023) include a pool of 535 words for three different parts-of-speech (nouns, verbs and adjectives) and 11 semantic classes (see the full list in Table 6). The dataset is a translation of the brain-based English norms introduced by Binder et al. (2016) and contains ratings for the words across 65 different experiential domains.

[5]For this reason, our results with static embeddings might diverge from previous work using this benchmark.

We use the word embeddings of the dataset words as inputs to a hierarchical agglomerative clustering algorithm, and assess the extent to which a model reproduces the gold clustering of abstract vs. concrete classes in Zhong22, and of the 11 semantic classes in Binder-zh (in both cases, the number of clusters in the gold standard is fed as a parameter to the clustering algorithm).

Type-POS	No. of items
Concrete Objects - Nouns	275
Living Things - Nouns	126
Other Natural Objects - Nouns	19
Artifacts - Nouns	130
Concrete Events - Nouns	60
Abstract Entities - Nouns	99
Concrete Actions - Verbs	52
Abstract Actions - Verbs	5
States - Verbs	5
Abstract Properties - Adjectives	13
Physical Properties - Adjectives	26

Table 6: Concept classes, parts-of-speech and number of words in the Binder-zh norms.

The evaluation metrics are *homogeneity* and *completeness*. Homogeneity is defined in terms of the entropy of the cluster C given the class K (Equation 2), and it achieves 1 as its highest score if all the clusters contain only data points belonging to single class. Completeness is defined in terms of the entropy of the class K given the cluster C (Equation 2), and it achieves 1 as highest score if all the data points belonging to a single class are elements of the same cluster.

$$h = 1 - \frac{H(C|K)}{H(C)} \qquad c = 1 - \frac{H(K|C)}{H(K)} \quad (2)$$

It should be noticed that the static models do not have full coverage for all the datasets (see Table 7). Therefore, we exclude the items that are not included in the Skip Gram vocabulary and we evaluate the models on the remaining ones, in order to guarantee a fair comparison on the same items.

Dataset	Missing words	Covered items
COS960	34	898
FAST-zh	5	285
CA8	0	7363
EVALution-MAN	222	7095
Zhong22	0	664
Binder-zh	46	489

Table 7: Missing words for the static models, and remaining items in each dataset. Metrics are computed on the covered items.

	Spearman		Cosine	
Model	ρ	r	ρ	r
BERT-Base First 4	0.73	0.70	0.71	0.68
BERT-Base Last	0.74	0.70	0.74	0.69
BERT-Base Last 4	0.73	0.69	0.73	0.67
BERT-Base Middle 4	0.72	0.69	0.72	0.67
DeBERTa First 4	0.74	0.72	0.73	0.69
DeBERTa Last	0.74	0.71	0.72	0.66
DeBERTa Last 4	0.74	**0.72**	0.73	0.67
DeBERTa Middle 4	0.73	0.70	0.74	**0.70**
GPT-2 First 4	0.73	0.70	0.70	0.66
GPT-2 Last	0.70	0.67	0.44	0.37
GPT-2 Last 4	0.71	0.68	0.61	0.56
GPT-2 Middle 4	0.72	0.69	0.71	0.66
SkipGram N+C	**0.75**	0.70	**0.75**	**0.70**
SkipGram N	0.69	0.64	0.69	0.64
SkipGram C	0.71	0.67	0.71	0.67
SkipGram	0.65	0.60	0.65	0.61

Table 8: Similarity Estimation on COS960 dataset. We perform Spearman (ρ) and Pearson (r) correlations using Spearman/Cosine as similarity metrics. The best performance is shown in bold.

4. Results

The results for the similarity estimation task and the word association task are shown in Tables 8 and 9. At a glance, we can see that on COS960 static and embeddings models perform similarly, with the Skip Gram with word, ngrams, and characters as contexts being on par with the best contextualized models. It is evident that while the Skip Gram model with only words as context is the weakest one, incorporating characters and ngrams as extra contexts is strongly beneficial, with the full model being the best performing one in both similarity metrics. Model scores are not particularly affected by the frequency of the word pairs, i.e., the *hubness effect* in distributional models (Dinu et al., 2014; Schnabel et al., 2015), as the Spearman correlation between the similarity metrics and the average log frequencies of the words in each pair consistently shows weak associations (< 0.2 for all models).[6] This confirms that the models are not simply assigning higher scores to more frequent words.

For the word association task on FAST-zh, static models clearly exhibit the best performance, regardless of features used in training Skip Gram. Among the contextualized models, we have not found any striking difference across layers (previous study had reported a general better performance in early layers, cf. Chronis and Erk (2020); Lenci et al. (2023)) and similarity metrics. We analyzed the errors on this dataset, and the percent-

[6] Frequencies from a combination of corpora were extracted via the *wordfreq* Python library (Speer, 2022).

	Accuracy		Mean Rank	
Model	cos	ρ	cos	ρ
BERT-Base First 4	0.68	0.68	2.60	2.61
BERT-Base Last	0.69	0.69	2.28	2.42
BERT-Base Last 4	0.71	0.71	2.40	2.25
BERT-Base Middle 4	0.70	0.71	2.41	2.32
DeBERTa First 4	0.69	0.68	2.56	2.49
DeBERTa Last	0.71	0.70	2.40	2.28
DeBERTa Last 4	0.69	0.72	2.41	2.25
DeBERTa Middle 4	0.68	0.69	2.42	2.28
GPT-2 First 4	0.68	0.66	2.61	2.60
GPT-2 Last	0.58	0.68	2.87	2.42
GPT-2 Last 4	0.66	0.68	2.53	2.39
GPT-2 Middle 4	0.66	0.66	2.54	2.53
SkipGram	0.73	**0.73**	2.16	2.17
SkipGram C	0.72	0.71	**2.10**	**2.12**
SkipGram N	**0.74**	**0.73**	**2.10**	2.13
SkipGram N+C	0.73	0.72	2.19	2.20

Table 9: Word Associations results on the FAST-zh dataset. We show accuracy (the higher the better) and mean rank (the lower the better) using cosine (cos) and Spearman (ρ) as similarity metrics.

Model	**FIRST**	**HIGHER**	**RAND**
BERT First 4	0.64	0.30	0.05
BERT Last	0.65	0.32	0.03
BERT Last 4	0.67	0.31	0.02
BERT Middle 4	0.67	0.30	0.03
DeBERTa First 4	0.64	0.32	0.04
DeBERTa Last	0.66	0.32	0.02
DeBERTa Last 4	0.68	0.30	0.02
DeBERTa Middle 4	0.65	0.32	0.03
GPT-2 First 4	0.62	0.34	0.04
GPT-2 Last	0.64	0.33	0.03
GPT-2 Last 4	0.64	0.34	0.02
GPT-2 Middle 4	0.62	0.34	0.04
SkipGram	0.72	0.27	0.01
SkipGram C	0.71	0.28	0.01
SkipGram N	0.73	0.26	0.01
SkipGram N+C	0.73	0.26	0.01

Table 10: Percentage of items with highest similarity with the stimulus in the FAST-zh dataset.

age of items with the highest similarity to the stimulus can be found in Table 10. As expected, FIRST is correctly recognized as the strongest associate in most cases, but HIGHER associates are still effective confounders, misleading the models on average in approximately 30% of the cases.

Some major differences appear in the analogy task on the CA8 dataset, as shown in Table 11. First of all, Skip Gram models achieve much higher scores compared to the contextualized models, showing near-perfect performance across all settings. Notice that our scores even surpass most of the Chinese models evaluated on this dataset, but this outcome is expected due to a methodological adjustment. To ensure comparability between

Model	cos	ρ
BERT-Base First 4	0.39	0.39
BERT-Base Last	0.84	0.84
BERT-Base Last 4	0.82	0.82
BERT-Base Middle 4	0.67	0.68
DeBERTa First 4	0.44	0.44
DeBERTa Last	0.63	0.63
DeBERTa Last 4	0.63	0.63
DeBERTa Middle 4	0.57	0.57
GPT-2 First 4	0.41	0.41
GPT-2 Last	0.38	0.38
GPT-2 Last 4	0.44	0.43
GPT-2 Middle 4	0.44	0.43
SkipGram N+C	0.93	0.93
SkipGram N	**0.97**	**0.97**
SkipGram C	0.91	0.91
SkipGram	0.93	0.92

Table 11: Semantic analogies results for the CA8 dataset. We show accuracy using cosine (cos) and Spearman (ρ) as similarity metrics.

static and contextualized vectors, we limited the search space for both models to only the words that are present in the dataset, rather than using the entire vocabulary of the vector space. On the other hand, among the Transformers, BERT performs better, while GPT-2 vectors significantly lag behind. Excluding GPT-2, it can also be seen that the other two Transformer models show better results when using the later layers. This observation aligns with intuition, as the meaning of the relations between entities in the analogy task is likely to be better grasped via lexico-syntactic patterns in specific sentence contexts (e.g., ***x** is the capital of **y***).

In the semantic relations task (Table 12), contextualized models, particularly BERT, consistently outperform all the competitors by a large margin for all metrics and layer settings. Interestingly, in this task, Spearman as a metric is much more reliable for the contextualized models, always yielding higher scores for both DeBERTa and GPT-2. Static models, while not as powerful as contextualized models, still exhibit strong performance. In general, all models efficiently discriminate related words from random ones (Rel scores) as well as synonyms from other word pairs (Syn) (Figure 1). This aligns with previous results of the CogALex shared task (Xiang et al., 2020), where embeddings-based supervised systems achieved much better performance in Chinese relation identification compared to English. Possibly, the semantic radicals of the characters provide additional information about the word categories and their associations (Wang et al., 2018).

Finally, in the Semantic Clustering task (Table 13), contrasting results were found. In the Zhong22 dataset, static models seem to perform better in the abstract-concrete distinction, but

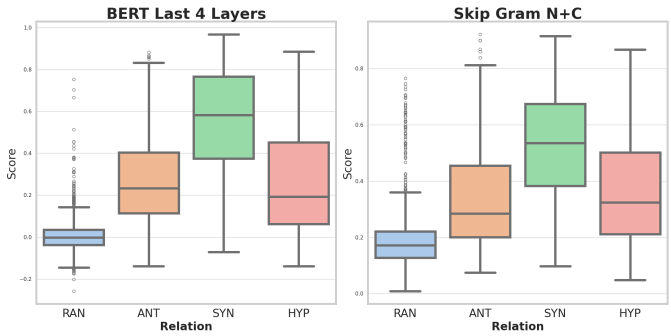

Figure 1: Distribution of Spearman scores of the best contextualized and static models (BERT and Skip Gram N+C) across the relations in the EVALution-MAN dataset.

	cos		ρ	
Model	**Rel.**	**Syn.**	**Rel.**	**Syn.**
BERT-Base First 4	0.95	0.50	0.92	0.52
BERT-Base Last	0.93	0.56	**0.94**	**0.61**
BERT-Base Last 4	**0.96**	0.58	**0.94**	**0.61**
BERT-Base Middle 4	**0.96**	**0.59**	**0.94**	**0.61**
DeBERTa First 4	0.69	0.25	0.83	0.42
DeBERTa Last	0.76	0.33	0.91	0.56
DeBERTa Last 4	0.72	0.29	0.89	0.54
DeBERTa Middle 4	0.71	0.29	0.87	0.52
GPT-2 First 4	0.89	0.49	0.91	0.53
GPT-2 Last	0.78	0.44	0.92	0.59
GPT-2 Last 4	0.87	0.49	0.92	0.59
GPT-2 Middle 4	0.91	0.52	0.91	0.57
SkipGram	0.80	0.26	0.80	0.25
SkipGram C	0.82	0.33	0.81	0.32
SkipGram N	0.79	0.25	0.79	0.25
SkipGram N+C	0.80	0.30	0.79	0.29

Table 12: Semantic Relations results on EVALution-MAN. We show average precision using cosine (cos) and Spearman (ρ) as metrics on related (Rel.) and synonymy (Syn.) classes.

	Zhong		**Binder-zh**	
Model	**H**	**C**	**H**	**C**
BERT-Base First 4	0.11	0.12	0.10	0.31
BERT-Base Last	0.28	0.31	0.14	0.37
BERT-Base Last 4	0.16	0.24	0.12	0.35
BERT-Base Middle 4	0.25	0.28	0.12	0.34
DeBERTa First 4	0.09	0.16	0.13	0.37
DeBERTa Last	0.21	0.23	0.13	0.36
DeBERTa Last 4	0.29	0.32	0.14	0.36
DeBERTa Middle 4	0.29	0.29	**0.17**	**0.46**
GPT-2 First 4	0.07	0.07	0.11	0.33
GPT-2 Last	0.03	0.03	0.04	0.07
GPT-2 Last 4	0.04	0.04	0.11	0.28
GPT-2 Middle 4	0.07	0.11	0.15	0.40
SkipGram N+C	0.20	0.25	0.08	0.22
SkipGram N	0.21	0.25	0.07	0.20
SkipGram C	0.34	0.35	0.03	0.21
SkipGram	**0.37**	**0.37**	0.03	0.21

Table 13: Semantic Clustering results on Zhong22 and Binder-zh. We show homogeneity (H) and completeness (C) using agglomerative clustering.

when Skip Gram uses only words as the contexts, while the inclusion of ngrams and characters deteriorates the performance. Contextualized models, except for BERT and DeBERTa in the middle-to-later layers, achieve lower scores. However, for the fine-grained semantic distinctions of the Binder dataset, BERT and DeBERTa always do better in the middle and late layers, and static models lag behind. The result may be due to the fact that the abstract-concrete distinction is closely related to the out-of-context property of word meaning, strongly determining the distributional behaviour of words (e.g., in the selectional preferences), whereas attribution to a Binder class may require idiosyncratic contextual cues that are only available to contextualized models.

Upon suggestions of the reviewers, we have run some further experiments including also BERT-Large, to test the effect of the model size, and an additional study on dimensionality reduction using Singular Value Decomposition (SVD, Deerwester et al. (1990); Landauer and Dumais (1997)) on the contextualized embeddings. We found that the quality of the contextualized embeddings decreases with SVD, and that BERT-Large constantly improves over the Base model. The full results can be seen in the Appendix.

5. Conclusion

We presented the first extensive comparison of static and contextualized embedding models in Mandarin Chinese, including different Transformer models (BERT, DeBERTa, and GPT-2) and static models with different levels of contextual granularity. Our results align with those of Lenci et al. (2023), possibly with an even larger edge in favor of static models in tasks that require a representation of out-of-context word meaning (such as similarity, word associations, analogy, abstract/concrete clustering). This is perhaps due to the fact that Transformers' vocabularies for Chinese are made of characters, so the models do

not have a 'native' representation of word meanings, but have to combine character-based ones (and multi-character words are more than 84% of the combined vocabulary of our datasets). Notice that, although cognitive research (Tsai and McConkie, 2003; Bai et al., 2008) claimed that the character could be the main unit for Chinese language processing, current benchmarks for distributional models are based on words. We want to stress that this issue is not limited to distributional semantics evaluation: for example, an increasing number of computational psycholinguistics works use word surprisals derived from pre-trained language models for modeling human reading behavior, often focusing on English; it is unclear if results obtained for English can be reproduced on typologically different languages (see e.g. Kuribayashi et al. (2021)), and whether different outcomes can be explained by the fact that not all languages have the same notion of what counts as a "word" (Nair and Resnik, 2023). In the case of Chinese, future benchmarks might possibly have to be conceived with this issue in mind, and the consideration could be extended to other Chinese modeling studies using annotations at the word level (e.g. prediction of eye movements in reading Chinese text, Li et al. (2023, 2024)).

On the other hand, contextualized vectors led to improvements in relation classification and categorization in abstract semantic classes, which probably benefit from context-specific semantic cues. For the task of semantic relations, scores were much higher than recent work on English (cf. the scores of the multilingual shared task in Xiang et al. (2020)). A possible explanation is that the Chinese dataset may be simpler for character-based models: for some relations (e.g. antonymy), there are word pairs formed with the same root-morpheme, and therefore such pairs share at least one identical character (e.g. 以下'below' vs. 以上'above'; 上山'uphill' vs. 下山'down hill'); moreover, some pairs also exhibit a Modern Chinese-Archaic Chinese alternation, and they also share a character (e.g. the SYN pair 來自於 vs. 源於, 'originate'). Again, such findings suggest that linguistic specificity has to be taken into account when planning benchmarks for distributional models. We have to stress, among the limitations of our study, that our evaluation is not multilingual as we focus just on Mandarin - so the conclusions we draw are limited to that language and to English via comparison with previous results. Multilingual distributional models will have to take into account the compatibility of the tokenizations between different languages (Maronikolakis et al., 2021), and deal with the fact that such languages may rely on different basic linguistic units.

In times when language models seem to have definitively shifted the attention toward evaluation with downstream tasks, we hope our work can pave the way for the rediscovery of distributional semantics in new and unseen languages.

Limitations

Our study has some clear limitations in that we tested a relatively small number of models. Additionally, we focused just on a single language, so it is possible that our findings do not generalize to DSM performance across languages.

Acknowledgements

We would especially like to thank the PolyU-CBS community, Chu-Ren Huang and John Sie Yuen Lee for the fruitful discussion on a preliminary version of this paper, plus the three anonymous reviewers for their insightful comments.

This research was also partly funded by PNRR—M4C2—Investimento 1.3, Partenariato Esteso PE00000013— "FAIR—Future Artificial Intelligence Research"—Spoke 1 "Human-centered AI," funded by the European Commission under the NextGeneration EU programme, and by the General Research Fund (4-ZZRX) of the Department of Chinese and Bilingual Studies of the Hong Kong Polytechnic University.

6. Bibliographical References

Xuejun Bai, Guoli Yan, Simon P Liversedge, Chuanli Zang, and Keith Rayner. 2008. Reading Spaced and Unspaced Chinese Text: Evidence from Eye Movements. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5):1277.

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 ACL*.

Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A General Framework for Corpus-based Semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni and Alessandro Lenci. 2011. How We BLESSed Distributional Semantic Evaluation. In *Proceedings of the GEMS Workshop on GEometrical Models of Natural Language Semantics*.

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. 2016. Toward a Brain-Based Componential Semantic Representation. *Cognitive Neuropsychology*, 33(3-4):130–174.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of ACL*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas. *Behavior Research Methods*, 46(3):904–911.

John A Bullinaria and Joseph P Levy. 2012. Extracting Semantic Representations from Word Co-occurrence Statistics: Stop-lists, Stemming, and SVD. *Behavior Research Methods*, 44(3):890–907.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 Task 9: Hypernym Discovery. In *Proceedings of SemEval*.

Chi-Yen Chen and Wei-Yun Ma. 2018. Word Embedding Evaluation Datasets and Wikipedia Title Embedding for Chinese. In *Proceedings of LREC*.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huanbo Luan. 2015. Joint Learning of Character and Word Embeddings. In *Proceedings of IJCAI*.

Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2016. Representing Verbs with Rich Contexts: An Evaluation on Verb Similarity. In *Proceedings of EMNLP*.

Gabriella Chronis and Katrin Erk. 2020. When Is a Bishop Not Like a Rook? When it's like a Rabbi! Multi-prototype BERT Embeddings for Estimating Semantic Relationships. In *Proceedings of CONLL*.

Kenneth Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of EMNLP*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. *arXiv preprint arXiv:2304.08177*.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "Small World of Words" English Word Association Norms for over 12,000 cue Words. *Behavior Research Methods*, 51:987–1006.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving Zero-shot Learning by Mitigating the Hubness Problem. *arXiv preprint arXiv:1412.6568*.

San Duanmu. 2017. Word and Wordhood, Modern. *Encyclopedia of Chinese Language and Linguistics*, 4:543–49.

Kawin Ethayarajh. 2019. How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of EMNLP*.

Stefanie Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

Stefanie Evert and Gabriella Lapesa. 2021. FAST: A Carefully Sampled and Cognitively Motivated Dataset for Distributional Semantic Evaluation. In *Proceedings of CONLL*.

Pablo Gamallo. 2019. A Dependency-based Approach to Word Contextualization Using Compositional Distributional Semantics. *Journal of Language Modelling*, 7(1):99–138.

Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. 2007. Topics in Semantic Representation. *Psychological Review*, 114(2):211.

Zellig S. Harris. 1954. Distributional Structure. *Word*, 10(2-3):146–162.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv preprint arXiv:2006.03654*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695.

Junjie Huang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, and Maosong Sun. 2019. COS960: A Chinese Word Similarity Dataset of 960 Word Pairs. *arXiv preprint arXiv:1906.00247*.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional Distributional Similarity for Lexical Inference. *Natural Language Engineering*, 16(4):359–389.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower Perplexity Is Not Always Human-like. In *Proceedings of ACL*.

Trina Kwong, Emmanuele Chersoni, and Rong Xiang. 2022. Evaluating Monolingual and Crosslingual Embeddings on Datasets of Word Association Norms. In *Proceedings of the LREC Workshop on Building and Using Comparable Corpora*.

Thomas K Landauer and Susan T Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211.

Yu-Yin Hsu Le Qiu and Emmanuele Chersoni. 2023. Collecting and Predicting Neurocognitive Norms for Mandarin Chinese. In *Proceedings of IWCS*.

Alessandro Lenci. 2008. Distributional Semantics in Linguistic and Cognitive Research. *Italian Journal of Linguistics*, 20(1):1–31.

Alessandro Lenci. 2018. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4:151–171.

Alessandro Lenci. 2022. Understanding Natural Language Understanding Systems. *Sistemi Intelligenti*, pages 1–26.

Alessandro Lenci and Magnus Sahlgren. 2023. *Distributional Semantics*. Cambridge University Press.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2023. A Comparative Evaluation and Analysis of Three Generations of Distributional Semantic Models. *Language Resources and Evaluation*, pages 1–45.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Junlin Li, Yu-Yin Hsu, Emmanuele Chersoni, and Bo Peng. 2024. Predicting Mandarin and Cantonese Adult Speakers' Eye-Movement Patterns in Natural Reading. In *Proceedings of the EACL Workshop on Research in Computational Linguistic Typology and Multilingual NLP*.

Junlin Li, Bo Peng, Yu-Yin Hsu, and Emmanuele Chersoni. 2023. Comparing and Predicting Eye-tracking Data of Mandarin and Cantonese. In *Proceedings of the EACL Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical Reasoning on Chinese Morphological and Semantic Relations. In *Proceedings of ACL*.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is Word Segmentation Necessary for Deep Learning of Chinese Representations? In *Proceedings of ACL*.

Yanran Li, Wenjie Li, Fei Sun, and Sujian Li. 2015. Component-enhanced Chinese Character Embeddings. In *Proceedings of EMNLP*.

Hongchao Liu, Emmanuele Chersoni, Natalia Klyueva, Enrico Santus, and Chu-Ren Huang. 2019. Semantic Relata for the Evaluation of Distributional Models in Mandarin Chinese. *IEEE Access*, 7:145705–145713.

Qi Liu, Matt J Kusner, and Phil Blunsom. 2020. A Survey on Contextual Embeddings. *arXiv preprint arXiv:2003.07278*.

Kevin Lund and Curt Burgess. 1996. Producing High-dimensional Semantic Spaces from Lexical Co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

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.

Antonis Maronikolakis, Philipp Dufter, and Hinrich Schütze. 2021. Wine is Not vi n.–On the Compatibility of Tokenizations Across Languages. In *Findings of EMNLP*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

George A Miller and Walter G Charles. 1991. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 6(1):1–28.

Gregory L Murphy. 2004. *The Big Book of Concepts*. MIT Press.

Gregory L Murphy. 2016. Is There an Exemplar Theory of Concepts? *Psychonomic Bulletin & Review*, 23:1035–1042.

Sathvik Nair and Philip Resnik. 2023. Words, Subwords, and Morphemes: What Really Matters in the Surprisal-Reading Time Relationship? In *Findings of EMNLP*.

Robert M Nosofsky. 2013. Exemplars, Prototypes, and Similarity Rules. In *From Learning Theory to Connectionist Theory*, pages 149–167. Psychology Press.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of EMNLP*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of NAACL*.

Le Qiu, Yu-Yin Hsu, and Emmanuele Chersoni. 2023. Collecting and Predicting Neurocognitive Norms for Mandarin Chinese. In *Proceedings of IWCS*.

Yuanyuan Qiu, Hongzheng Li, Shen Li, Yingdi Jiang, Renfen Hu, and Lijiao Yang. 2018. Revisiting Correlations between Intrinsic and Extrinsic Evaluations of Word Embeddings. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 209–221. Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.

Magnus Sahlgren. 2008. The Distributional Hypothesis. *Italian Journal of Linguistics*, 20:33–53.

Magnus Sahlgren and Alessandro Lenci. 2016. The Effects of Data Size and Frequency Range on Distributional Semantic Models. In *Proceedings of EMNLP*.

Enrico Santus. 2016. *Making Sense: From Word Distribution to Meaning*. Ph.D. thesis, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University.

Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, Chu-Ren Huang, and Philippe Blache. 2016a. Testing APsyn against Vector Cosine on Similarity Estimation. In *Proceedings of PACLIC*.

Enrico Santus, Tin-Shing Chiu, Qin Lu, Alessandro Lenci, and Chu-Ren Huang. 2016b. What a Nerd! Beating Students and Vector Cosine in the ESL and TOEFL Datasets. In *Proceedings of LREC*.

Enrico Santus, Hongmin Wang, Emmanuele Chersoni, and Yue Zhang. 2018. A Rank-Based Similarity Metric for Word Embeddings. In *Proceedings of ACL*.

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. Evalution 1.0: An Evolving Semantic Dataset for Training and Evaluation of Distributional Semantic Models. In *Proceedings of the ACL Workshop on Linked Data in Linguistics: Resources and Applications*.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation Methods for Unsupervised Word Embeddings. In *Proceedings of EMNLP*.

Sabine Schulte Im Walde. 2020. Distinguishing between Paradigmatic Semantic Relations across Word Classes: Human Ratings and Distributional Similarity. *Journal of Language Modelling*, 8(1):53–101.

Chenglei Si, Zhengyan Zhang, Yingfa Chen, Fanchao Qi, Xiaozhi Wang, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2023. Subcharacter Tokenization for Chinese Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 11:469–487.

Robyn Speer. 2022. rspeer/wordfreq: v3.0. *Zenodo*.

Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. VCWE: Visual Character-enhanced Word Embeddings. In *Proceedings of NAACL-HLT*.

Yaming Sun, Lei Lin, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. 2014. Radical-enhanced Chinese Character Embedding. In *Proceedings of NeurIPS*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of ACL*.

William Timkey and Marten van Schijndel. 2021. All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality. In *Proceedings of EMNLP*.

Jie-Li Tsai and George W McConkie. 2003. Where Do Chinese Readers Send Their Eyes? In *The Mind's Eye*, pages 159–176. Elsevier.

Peter D Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention As All You Need. *Advances in Neural Information Processing Systems*, 30.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing Pretrained Language Models for Lexical Semantics. In *Proceedings of EMNLP*.

Ada Wan. 2021. Fairness in Representation for Multilingual NLP: Insights from Controlled Experiments on Conditional Language Modeling. In *International Conference on Learning Representations*.

Xiaoxi Wang, Xie Ma, Yun Tao, Yachen Tao, and Hong Li. 2018. How Semantic Radicals in Chinese Characters Facilitate Hierarchical Category-based Induction. *Scientific Reports*, 8(1):5577.

Rong Xiang, Emmanuele Chersoni, Luca Iacoponi, and Enrico Santus. 2020. The CogALex Shared Task on Monolingual and Multilingual Identification of Semantic Relations. In *Proceedings of the COLING Workshop on the Cognitive Aspects of the Lexicon*.

Jian Xu, Jiawei Liu, Liangang Zhang, Zhengyu Li, and Huanhuan Chen. 2016. Improve Chinese Word Embeddings by Exploiting Internal Structure. In *Proceedings of NAACL-HLT*.

Rongchao Yin, Quan Wang, Peng Li, Rui Li, and Bin Wang. 2016. Multi-granularity Chinese Word Embedding. In *Proceedings of EMNLP*.

Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. 2017. Joint Embeddings of Chinese Words, Characters, and Fine-grained Subcharacter Components. In *Proceedings of EMNLP*.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. *CoRR*, abs/2209.02970.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. UER: An Open-Source Toolkit for Pre-training Models. In *Proceedings of EMNLP-IJCNLP*.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Hammerla. 2019. Correlation Coefficients and Semantic Textual Similarity. In *Proceedings of NAACL*.

Yin Zhong, Mingyu Wan, Kathleen Ahrens, and Chu-Ren Huang. 2022. Sensorimotor Norms for Chinese nouns and their Relationship with Orthographic and Semantic Variables. *Language, Cognition and Neuroscience*, 37(8):1000–1022.

Appendix

We explored the impact of dimensionality on the performance of contextualized word embeddings, since the vectors produced by all the Transformer models have 768 dimensions vs. the 300 dimensions of static word embeddings. With this goal, we reduce to 300 dimensions the contextualized vectors with SVD (Deerwester et al., 1990; Landauer and Dumais, 1997) and repeat all the experiments. Moreover, we include results for the BERT-Large model to assess the effect of model size. All scores can be seen in Table 14-18.

It is immediately noticeable that, after SVD compression is applied, the scores of contextualized embeddings slightly decreases for almost all the tasks and settings, suggesting that the higher dimensionality of those vectors might have played a role in their performance in previous studies.

On the other hand, BERT-Large seems to small but constant improvements over the Base version in the same settings in almost all the tasks. Although the observed pattern does not change substantially, the scores achieved by this model suggests that representation quality of contextualized embeddings might improve with the increase of the model size.

			Spearman		Cosine	
Model	**Config**	**Compression**	ρ	r	ρ	r
BERT-Base	First 4	with SVD	0.70	0.67	0.68	0.65
		w/o SVD	0.73	0.70	0.71	0.68
BERT-Base	Last	with SVD	0.71	0.67	0.71	0.66
		w/o SVD	0.74	0.70	0.74	0.69
BERT-Base	Last 4	with SVD	0.70	0.66	0.70	0.64
		w/o SVD	0.73	0.69	0.73	0.67
BERT-Base	Middle 4	with SVD	0.69	0.66	0.69	0.64
		w/o SVD	0.72	0.69	0.72	0.67
BERT-Large	First 4	with SVD	0.72	0.68	0.68	0.66
		w/o SVD	0.74	0.72	0.73	0.68
BERT-Large	Last	with SVD	0.73	0.68	0.72	0.67
		w/o SVD	0.75	0.70	0.72	0.71
BERT-Large	Last 4	with SVD	0.71	0.68	0.71	0.65
		w/o SVD	0.75	0.70	0.75	0.68
BERT-Large	Middle 4	with SVD	0.70	0.68	0.71	0.65
		w/o SVD	0.73	0.7	0.73	0.69
DeBERTa	First 4	with SVD	0.72	0.69	0.70	0.66
		w/o SVD	0.74	0.72	0.73	0.69
DeBERTa	Last	with SVD	0.72	0.68	0.69	0.63
		w/o SVD	0.74	0.71	0.72	0.66
DeBERTa	Last 4	with SVD	0.72	0.69	0.70	0.64
		w/o SVD	0.74	0.72	0.73	0.67
DeBERTa	Middle 4	with SVD	0.73	0.70	0.71	0.67
		w/o SVD	0.73	0.70	0.74	0.70
GPT-2	First 4	with SVD	0.70	0.67	0.68	0.63
		w/o SVD	0.73	0.70	0.70	0.66
GPT-2	Last	with SVD	0.67	0.64	0.41	0.34
		w/o SVD	0.70	0.67	0.44	0.37
GPT-2	Last 4	with SVD	0.68	0.65	0.58	0.53
		w/o SVD	0.71	0.68	0.61	0.56
GPT-2	Middle 4	with SVD	0.69	0.66	0.68	0.63
		w/o SVD	0.72	0.69	0.71	0.66
SkipGram N+C			0.75	0.70	0.75	0.70
SkipGram N			0.69	0.64	0.69	0.64
SkipGram C			0.71	0.67	0.71	0.67
SkipGram			0.65	0.60	0.65	0.61

Table 14: Similarity Estimation on COS960 dataset. We perform Spearman (ρ) and Pearson (r) correlations using Spearman/Cosine as similarity metrics. The best performance is shown in bold.

			Accuracy		Mean Rank	
Model	**Layers**	**Compression**	cos	ρ	cos	ρ
BERT-Base	First 4	with SVD	0.64	0.64	2.74	2.75
		w/o SVD	0.68	0.68	2.60	2.61
BERT-Base	Last	with SVD	0.65	0.65	2.42	2.56
		w/o SVD	0.69	0.69	2.28	2.42
BERT-Base	Last 4	with SVD	0.67	0.67	2.54	2.39
		w/o SVD	0.71	0.71	2.40	2.25
BERT-Base	Middle 4	with SVD	0.66	0.67	2.55	2.46
		w/o SVD	0.70	0.71	2.41	2.32
BERT-Large	First 4	with SVD	0.66	0.66	2.69	2.70
		w/o SVD	0.70	0.70	2.55	2.56
BERT-Large	Last	with SVD	0.67	0.67	2.37	2.51
		w/o SVD	0.71	0.71	2.23	2.37
BERT-Large	Last 4	with SVD	0.69	0.69	2.49	2.34
		w/o SVD	0.73	0.73	2.35	2.20
BERT-Large	Middle 4	with SVD	0.68	0.69	2.50	2.41
		w/o SVD	0.72	0.73	2.36	2.27
DeBERTa	First 4	with SVD	0.65	0.64	2.70	2.63
		w/o SVD	0.69	0.68	2.56	2.49
DeBERTa	Last	with SVD	0.67	0.66	2.54	2.42
		w/o SVD	0.71	0.70	2.40	2.28
DeBERTa	Last 4	with SVD	0.65	0.68	2.55	2.39
		w/o SVD	0.69	0.72	2.41	2.25
DeBERTa	Middle 4	with SVD	0.64	0.65	2.56	2.42
		w/o SVD	0.68	0.69	2.42	2.28
GPT-2	First 4	with SVD	0.64	0.62	2.75	2.74
		w/o SVD	0.68	0.66	2.61	2.60
GPT-2	Last	with SVD	0.54	0.64	3.01	2.56
		w/o SVD	0.58	0.68	2.87	2.42
GPT-2	Last 4	with SVD	0.62	0.64	2.67	2.53
		w/o SVD	0.66	0.68	2.53	2.39
GPT-2	Middle 4	with SVD	0.62	0.62	2.68	2.67
		w/o SVD	0.66	0.66	2.54	2.53
SkipGram			0.73	0.73	2.16	2.17
SkipGram C			0.72	0.71	2.10	2.12
SkipGram N			0.74	0.73	2.10	2.13
SkipGram N+C			0.73	0.72	2.19	2.20

Table 15: Word Associations results on the FAST-zh dataset. We show accuracy (the higher the better) and mean rank (the lower the better) using cosine (cos) and Spearman (ρ) as similarity metrics.

Model	**Layers**	**Compression**	cos	ρ
BERT-Base	First 4	with SVD	0.36	0.36
		w/o SVD	0.39	0.39
BERT-Base	Last	with SVD	0.80	0.80
		w/o SVD	0.84	0.84
BERT-Base	Last 4	with SVD	0.78	0.78
		w/o SVD	0.82	0.82
BERT-Base	Middle 4	with SVD	0.63	0.64
		w/o SVD	0.67	0.68
BERT-Large	First 4	with SVD	0.38	0.38
		w/o SVD	0.41	0.41
BERT-Large	Last	with SVD	0.82	0.82
		w/o SVD	0.86	0.86
BERT-Large	Last 4	with SVD	0.80	0.80
		w/o SVD	0.84	0.84
BERT-Large	Middle 4	with SVD	0.65	0.66
		w/o SVD	0.69	0.70
DeBERTa	First 4	with SVD	0.41	0.41
		w/o SVD	0.44	0.44
DeBERTa	Last	with SVD	0.59	0.59
		w/o SVD	0.63	0.63
DeBERTa	Last 4	with SVD	0.59	0.59
		w/o SVD	0.63	0.63
DeBERTa	Middle 4	with SVD	0.53	0.53
		w/o SVD	0.57	0.57
GPT-2	First 4	with SVD	0.38	0.38
		w/o SVD	0.41	0.41
GPT-2	Last	with SVD	0.35	0.35
		w/o SVD	0.38	0.38
GPT-2	Last 4	with SVD	0.41	0.40
		w/o SVD	0.44	0.43
GPT-2	Middle 4	with SVD	0.41	0.40
		w/o SVD	0.44	0.43
SkipGram N+C			0.93	0.93
SkipGram N			0.97	0.97
SkipGram C			0.91	0.91
SkipGram			0.93	0.92

Table 16: Semantic analogies results for the CA8 dataset. We show accuracy using cosine (cos) and Spearman (ρ) as similarity metrics.

			cos		ρ	
Model	**Layers**	**Compression**	**Rel.**	**Syn.**	**Rel.**	**Syn.**
BERT-Base	First 4	with SVD	0.92	0.47	0.89	0.49
		w/o SVD	0.95	0.50	0.92	0.52
BERT-Base	Last	with SVD	0.90	0.53	0.91	0.58
		w/o SVD	0.93	0.56	0.94	0.61
BERT-Base	Last 4	with SVD	0.93	0.55	0.91	0.58
		w/o SVD	0.96	0.58	0.94	0.61
BERT-Base	Middle 4	with SVD	0.93	0.56	0.91	0.58
		w/o SVD	0.96	0.59	0.94	0.61
BERT-Large	First 4	with SVD	0.94	0.49	0.91	0.51
		w/o SVD	0.97	0.52	0.94	0.54
BERT-Large	Last	with SVD	0.92	0.55	0.93	0.60
		w/o SVD	0.95	0.58	0.96	0.63
BERT-Large	Last 4	with SVD	0.95	0.57	0.93	0.60
		w/o SVD	0.98	0.60	0.96	0.63
BERT-Large	Middle 4	with SVD	0.95	0.58	0.93	0.60
		w/o SVD	0.98	0.61	0.96	0.63
DeBERTa	First 4	with SVD	0.66	0.22	0.80	0.39
		w/o SVD	0.69	0.25	0.83	0.42
DeBERTa	Last	with SVD	0.73	0.30	0.88	0.53
		w/o SVD	0.76	0.33	0.91	0.56
DeBERTa	Last 4	with SVD	0.69	0.26	0.86	0.51
		w/o SVD	0.72	0.29	0.89	0.54
DeBERTa	Middle 4	with SVD	0.68	0.26	0.84	0.49
		w/o SVD	0.71	0.29	0.87	0.52
GPT-2	First 4	with SVD	0.86	0.46	0.88	0.50
		w/o SVD	0.89	0.49	0.91	0.53
GPT-2	Last	with SVD	0.75	0.41	0.89	0.56
		w/o SVD	0.78	0.44	0.92	0.59
GPT-2	Last 4	with SVD	0.84	0.46	0.89	0.56
		w/o SVD	0.87	0.49	0.92	0.59
GPT-2	Middle 4	with SVD	0.88	0.49	0.88	0.54
		w/o SVD	0.91	0.52	0.91	0.57
SkipGram			0.80	0.26	0.80	0.25
SkipGram C			0.82	0.33	0.81	0.32
SkipGram N			0.79	0.25	0.79	0.25
SkipGram N+C			0.80	0.30	0.79	0.29

Table 17: Semantic Relations results on EVALution-MAN. We show average precision using cosine (cos) and Spearman (ρ) as metrics on related (Rel.) and synonymy (Syn.) classes.

			Zhong		Binder-zh	
Model	**Layers**	**Compression**	**H**	**C**	**H**	**C**
BERT-Base	First 4	with SVD	0.08	0.09	0.07	0.27
		w/o SVD	0.11	0.12	0.10	0.31
BERT-Base	Last	with SVD	0.25	0.28	0.11	0.33
		w/o SVD	0.28	0.31	0.14	0.37
BERT-Base	Last 4	with SVD	0.13	0.21	0.09	0.31
		w/o SVD	0.16	0.24	0.12	0.35
BERT-Base	Middle 4	with SVD	0.22	0.25	0.09	0.30
		w/o SVD	0.25	0.28	0.12	0.34
BERT-Large	First 4	with SVD	0.10	0.11	0.09	0.29
		w/o SVD	0.13	0.14	0.12	0.33
BERT-Large	Last	with SVD	0.27	0.30	0.13	0.35
		w/o SVD	0.30	0.33	0.16	0.39
BERT-Large	Last 4	with SVD	0.15	0.23	0.11	0.33
		w/o SVD	0.18	0.26	0.14	0.37
BERT-Large	Middle 4	with SVD	0.24	0.27	0.11	0.32
		w/o SVD	0.27	0.30	0.14	0.36
DeBERTa	First 4	with SVD	0.06	0.13	0.10	0.33
		w/o SVD	0.09	0.16	0.13	0.37
DeBERTa	Last	with SVD	0.18	0.20	0.10	0.32
		w/o SVD	0.21	0.23	0.13	0.36
DeBERTa	Last 4	with SVD	0.26	0.29	0.11	0.32
		w/o SVD	0.29	0.32	0.14	0.36
DeBERTa	Middle 4	with SVD	0.26	0.26	0.14	0.42
		w/o SVD	0.29	0.29	0.17	0.46
GPT-2	First 4	with SVD	0.04	0.04	0.08	0.29
		w/o SVD	0.07	0.07	0.11	0.33
GPT-2	Last	with SVD	0.00	0.00	0.01	0.03
		w/o SVD	0.03	0.03	0.04	0.07
GPT-2	Last 4	with SVD	0.01	0.01	0.08	0.24
		w/o SVD	0.04	0.04	0.11	0.28
GPT-2	Middle 4	with SVD	0.04	0.08	0.12	0.36
		w/o SVD	0.07	0.11	0.15	0.40
SkipGram N+C			0.20	0.25	0.08	0.22
SkipGram N			0.21	0.25	0.07	0.20
SkipGram C			0.34	0.35	0.03	0.21
SkipGram			0.37	0.37	0.03	0.21

Table 18: Semantic Clustering results on Zhong22 and Binder-zh. We show homogeneity (H) and completeness (C) using agglomerative clustering.