BAHP: Benchmark of Assessing Word Embeddings in Historical Portuguese

Zuoyu Tian[†] **Dylan Jarrett**[†] **Juan Manuel Escalona Torres**[‡] **Patrícia Amaral**[†]
[†]Indiana University Bloomington [‡]Cornell University
{zuoytian,dsjarret,pamaral}@iu.edu; jme252@cornell.edu

Abstract

High quality distributional models can capture lexical and semantic relations between words. Hence, researchers design various intrinsic tasks to test whether such relations are captured. However, most of the intrinsic tasks are designed for modern languages, and there is a lack of evaluation methods for distributional models of historical corpora. In this paper, we conducted BAHP: a benchmark of assessing word embeddings in Historical Portuguese, which contains four types of tests: analogy, similarity, outlier detection, and coherence. We examined word2vec models generated from two historical Portuguese corpora in these four test sets. The results demonstrate that our test sets are capable of measuring the quality of vector space models and can provide a holistic view of the model's ability to capture syntactic and semantic information. Furthermore, the methodology for the creation of our test sets can be easily extended to other historical languages.

1 Introduction

Distributional semantics assumes that words with similar contexts have similar meanings (Lenci, 2018), hence, words similar in meaning should be similar in vector representation since word embeddings are learned from context. The success of word embeddings in different tasks in natural language processing verifies such assumptions to some extent. However, it has also been noticed that the quality of word embeddings is critical to different applications since the embeddings are learned in an unsupervised way. Thus, for evaluating embedding models, different tasks appear, but most of these test sets are designed for English, and the multilingual versions are usually created by translation.

As more historical texts have been digitized, we notice the surge of using computational methods to study historical language. Compared to the large amount of digital resources in modern language, researchers normally use very limited resources to generate historical embedding models. However, the quality of embedding models generated from small corpora is not always satisfying (Sahlgren and Lenci, 2016). Thus, assessing embedding models are critical to distributional semantics studies in historical languages. For a number of reasons, there is a lack of evaluation datasets for historical word embeddings. First, it is difficult to directly translate English evaluation benchmarks to historical languages because of the lack of translation tools and mismatch of concepts. Second, creating intrinsic or extrinsic evaluation tasks specific to older stages of a language demands large amounts of manual annotation and requires related expert knowledge. Meanwhile, we notice that prior studies in evaluation benchmark creation typically focus on a single type of test set, and few studies discuss the distinction between different test sets, especially for historical languages and embedding models generated from small datasets.

In this paper, we aim to create intrinsic tasks for evaluating historical Portuguese embedding models in an efficient way. Major methods for assessing word embeddings such as analogy tests and similarity tests are questioned for issues of the appropriateness and informativeness and lower inter-annotator agreement. Building a benchmark containing more types of evaluating methods can avoid the bias from a single method. Efficiency, measured in terms of human effort and amount of resources needed, is another concern for creating the benchmark. We try to minimize the cost of building test sets of historical languages and make the benchmark reliable at the same time. Four types of evaluation sets are adopted. They target different types of relations within word embeddings: syntactic, semantic, and morphological. We then use our benchmark to evaluate word2vec models generated from two historical Portuguese corpora. We find that our

Proceedings of LaTeCH-CLfL 2021, pages 113–119
Punta Cana, Dominican Republic (Online), November 11, 2021.

evaluation datasets can effectively demonstrate the quality of the distributional model.

This paper is structured as follows: we review predominant word embedding assessment methods and existing Portuguese word embedding evaluation resources in the next section. In section 3, we introduce how we created four types of word embedding assessment datasets. Then, the evaluation experiments of word2vec models generated from Medieval and Classical Portuguese corpora are discussed in section 4. We summarize our findings in the last section.

2 Related Work

Analogy tests were first introduced by Mikolov et al. (2013) to evaluate the word2vec models. This analogy test also known as Google Analogy Test (GAT) which contains 9 types of syntactic relations and 5 types of semantic categories has been widely applied in embeddings assessment. But many researchers also pointed out the disadvantages of analogy test, for example, irrelevant neighborhood structure being captured (Linzen, 2016) and drawbacks in appropriateness and informativeness (Schluter, 2018).

Compared to finding the analogy relation within the vector space, word similarity test uses the corelation score between human annotated word pairs and similarity scores from the distributional model to assess the embeddings. Based on the different definition of similarity, such test sets could be divided into word similarity like Sim999 (Hill et al., 2015), word relatedness like MEN (Bruni et al., 2014), and the mix of these two concepts like Sim353 (Finkelstein et al., 2001). The major concern toward word similarity is low human inter-annotator agreement across many datasets (Batchkarov et al., 2016).

Besides the two instrinsic tasks above, other evaluation methods include concept categorization (Baroni et al., 2010), outlier detection tests (Camacho-Collados and Navigli, 2016), and coherence tests (Zhao et al., 2018). These tests can make up for the drawbacks of analogy tests and word similarity to some degree.

Regarding evaluating Portuguese distributional models, we see different types of intrinsic tasks in Modern Portuguese. Querido et al. (2017) translated a set of English intrinsic datasets including an analogy test, word similarity, and concept categorization into Portuguese. Wilkens et al. (2016) used Portuguese BabelNet to create a TOEFL-like test set called BSG targeting semantic relations. Gamallo (2018) created an English 12-8-8 outlier detection set and then translated it into Portuguese. Analogy test set TALES created by Oliveira et al. (2020) is compiled by using existing lexical resources and examines the lexical-semantic relations in Portuguese.

Unlike the resourceful test sets in modern languages, only a few studies developed assessing tasks for historical languages. Schlechtweg et al. (2018) designed a task to measure language change in German by annotating semantic relatedness between diachronic usage of words. This method has been adopted by the following studies and expanded language change detection tasks to English, Latin and Swedish (Schlechtweg et al., 2020). Hu et al. (2021) designed analogy tests for Medieval Spanish. Their analogy test includes seven categories of morphological and semantic relations, with more tests focusing on the former. But, these studies focus only on one type of instrinsic tasks (relatedness or analogy).

3 Test Set Creation

3.1 Historical Portuguese Corpora

In order to have a holistic view of historical Portuguese and present the performance of our test sets, we use two corpora as reference to create intrinsic tasks and generate embedding models. These two corpora include texts from Medieval Portuguese and Classical to Modern Portuguese, respectively.

The *Corpus Informatizado do Português Medieval* (CIPM) is an online corpus of Medieval Portuguese (with texts from the 12th to 16th centuries), with about 2,5 million tokens.[1] Compared to modern corpora used to generate embedding models, the size of the corpus is notably small. Moreover, this corpus lacks systematic punctuation and contains a considerable amount of orthographic and morphological variation, as well as editorial annotations. Even though we normalized some variants and deleted all editorial annotations during pre-processing, [2] we suspect that the quality of the word embedding model generated from this corpus will be clearly affected, since previous work show

[1] We scraped the whole corpus from https://cipm.fcsh.unl.pt/ after getting permission from the corpus creators.

[2] For a full list of normalized variants, please see Amaral et al. (Accepted)

that embedding models generated from small corpora are less reliable (Sahlgren and Lenci, 2016; Antoniak and Mimno, 2018).

The Corpus of Historical Portuguese (COLONIA) (Zampieri and Becker, 2013) is a tagged diachronic corpus containing historical texts representing Classical Portuguese and Modern Portuguese. It contains both European and Brazilian texts from the 16th to the 20th century with about 5,1 million tokens.[3] Compared to CIPM, COLONIA is well pre-processed and contains less orthographic variation, since the degree of variation in Portuguese orthography is moderate after the 16th century (Castro, 1991). Therefore, we directly use the raw corpus in our paper.

3.2 Word Analogy Test

Word analogy tests aim to check whether the vector model successfully captures certain relations between words. The relationship could be semantic, syntactic or morphological. For example, if we have two words: *dog* and *dogs*, they are in a singular-plural relation a morphological and a semantic relation. Given a third word *cat*, we aim to find the word in the same relationship to *cat*, which should be *cats* in this case. In an analogy test, we first calculate the distance between *dog* and *dogs* in vector space and then add it to the vector of the given word *cat*. If it is an optimal vector model, *cats* should be the nearest neighbour of this new vector. For the distance calculation, the cosine similarity is used here. To evaluate the quality of distributional model, we report the accuracy in sets of analogy questions.

Since the other three assessment tests in this paper all target the capacity of semantic representation, we created an analogy test set examining the syntactic and morphological relations in historical Portuguese. Two types of grammatical relations for nouns and four types of relations for verbs are included in our dataset. We first designed some seed word pairs with analogy relations in each type consulting the CIPM corpus. Then, we combined every two word pairs in each relation type to create a standard analogy test dataset. This procedure gives us 2,994 analogy questions in total. For the model evaluation, we use the built-in function in Python package Gensim.[4] In this paper, due to time constraints, we only built an analogy test set using the vocabulary from CIPM, since this corpus is more challenging given its size and variability. We plan on developing a version of this test for COLONIA in future work.

3.3 Word Similarity Test

Considering the small size of the reference corpus we used in the paper, we do not distinguish the two concepts of relatedness and similarity in this dataset. We adopted the semantic relation annotation scheme proposed by Jurgens et al. (2014) (see Table 4), which also corresponds to the construction of SemEval 2017 word similarity test (Camacho-Collados et al., 2017).

To conduct the word similarity test, 100 pairs of common nouns in CIPM were created, resulting in a total of 200 words. During word pair creation, we take frequency factor into consideration. Fifty words were selected from the top 20 percent of vocabulary in CIPM and fifty more were selected from the bottom 20 percent. This ensured that the anchor words in each pairing consisted of some of the most frequent nouns in the corpus and some of the most infrequent. Following this, the anchor words were paired with other common nouns in the corpus. All pairings were created by hand by one of the annotators, an expert in historical Portuguese, using both intuition and the consultation of historical dictionaries of Portuguese. In order to ensure a variety of similarity ratings, approximately one third of the anchor words were paired with words that were synonymous or nearly-synonymous. Another third were paired with nouns related in some way to the anchor words (e.g. cases of meronymy, members of the same semantic category) and the final third was paired with words which bore no perceivable relationship with the anchor words, as shown in Table 1.

The annotation of word pairs were done by two experts in historical Portuguese following the annotation scheme shown in Table 4. The Pearson correlation between two annotators is 0.8928. The inter-annotator agreement measured by Cohen's Kappa is 0.5651. Only 3 pairs have a larger than 2 degree annotation difference or unclear relation perceived by our experts, and were thus discarded. In our final test set, we had 97 word pairs. We report both Pearson and Spearman correlation scores in the final evaluation. As said before, currently, we use the same word similarity test for both CIPM and COLONIA as reported in section 4.

[3] http://corporavm.uni-koeln.de/colonia/

[4] https://radimrehurek.com/gensim/index.html

	Word 1	Word 2
Synonymous	povo ’people’	gente ’people’
Related (not synonymous)	cabeça ’head’	olhos ’eyes’
Not related	morte ’death’	conselho ’advice’

Table 1: Examples of word pairings

3.4 Outlier Detection

This evaluation method was first proposed by Camacho-Collados and Navigli (2016). This task examines whether the vector space models can successfully cluster semantically homogeneous words. In this task, we start with a group of words, for example, *tuna*, *flounder*, *trout*, and *bread*, with the goal of identifying the word that does not belong to this group. We can easily find that *bread* is the outlier in this case since it is not the name of a fish. In the original paper, the authors created a 8-8-8 dataset. This means that eight categories of concepts are included in the dataset, each category is composed of eight words in that category and eight outliers not related to that category. Then, there would be $8 \times 8 \times 8$ tests, or 512 total in one test set. Following their format, we created two outlier detection datasets for Medieval Portuguese and Classical-Modern Portuguese, respectively, based on this 8-8-8 format. To assure the quality of the dataset, we asked an expert in historical Portuguese to come up with the categorization and outlier clusters. Table 3 shows the categorization in each dataset. For the evaluation portion, accuracy and outlier position percentage (OPP) is used. This task yields better results for COLONIA than for CIPM (see Table 2). For this reason, we performed an error analysis for CIPM. For example, in the body parts category the model wrongly predicted *mao* ‘hand’ to be the outlier. However, this form is also one spelling variant for the adjective ’bad’ (current spelling: *mau*). The example shows that the model is capturing distinctions between words and that abundant variations in the corpus can give us a more challenging task to generate word embedding models, which also corresponds to our assumption that the quality of models from CIPM should be lower than that of COLONIA.

3.5 Coherence Assessment

Zhao et al. (2018) proposed a new evaluation method for assessing the quality of biomedical domain-specific word embedding models. They assume that the neighbors of a given word embedding should have the same characteristics of that word.

Analogy Test

	CIPM	COLONIA
	Accuracy	Accuracy
N-Gender	0.0579	0.1215
N-Singular-Plural	0.0232	0.0801
V-1SG.Pres-3SG.Pres	0.0067	0.3911
V-3SG.Pret-3PL.Pret	0.0035	0.6338
V-Infinitive-3SG.Pres	0.0122	0.2226
V-Infinitive-Gerund	0.0252	0.3522

Word Similarity

	Pearson	Spearman
CIPM	0.2759	0.3141
COLONIA	0.5326	0.5482

Outlier Detection

	Accuracy	OPP
CIPM	0.6398	0.9093
COLONIA	0.8742	0.9809

Coherence Assessment

	Top 5 (Acc.)	Top 10 (Acc.)
CIPM	0.8333	0.7542
COLONIA	0.9280	0.8800

Table 2: Results of BAHP in embeddings models generated from two corpora.

	Categorization
CIPM	body parts; Christianity; color; food; geography; parts of buildings; titles/professions; war
COLONIA	furnature; Christianity; color; food; geography; face; titles/professions; kitchen

Table 3: Categorizations in two outlier detection datasets

For example, in their paper, they focus on drug names. The assumption is, then, that the neighbors of drug names should also be drug names if the embedding model is of good quality.

Inspired by their method, we designed a similar test dataset. Given that proper nouns, especially names of people and places, are frequent in our corpora, we decided to assess the model by reporting the percentage of neighbors generated for a proper noun which were of the same category (i.e., proper nouns). In CIPM, many texts are narrations of either historical or fictional events and hence include names of kings, knights, saints, and places. This is true of other diachronic corpora given the nature of the texts produced and the copies of the manuscripts that survived to this day. We chose the 25 most frequent proper nouns in both CIPM and COLONIA to make two coherence assessment sets. For the evaluation, we reported the percentage of proper nouns in the top 5 and 10 neighbors, following Zhao et al. (2018). To give an example from each corpus, in CIPM 7 out of 10 neighbors of the proper name "Galaaz" (the name of a knight) were proper names (names of other knights of the Arthurian legend), while 3 were common nouns, hence the percentage was 0.7. In COLONIA all 10 neighbors of the proper name "Maria" were proper names (other names of women), hence the percentage result was 1.

4 Model Evaluation

To have a better understanding of how our test sets fit into the model evaluation, we examined the quality of two distributional models trained on two corpora in this section

4.1 Word Embedding Training

We employed Skip-gram with negative sampling (SGNS) architecture of word2vec to generate the word embedding models. Considering the randomness of SGNS models, we generated 20 models for each corpus. The results presented below are the average output of these 20 models. We used the following hyper-parameters to generate all the models: minimal word frequency threshold = 20, window size = 7, vector dimension = 300.

Table 3 presents the categorizations used in two outlier detection datasets.

4.2 Results

Table 2 show the evaluation results of the word embedding models on both corpora. It is clear that word embeddings generated from COLONIA are of higher quality across all four test sets than those from CIPM. This observation is not surprising, since texts in COLONIA are well-processed, punctuated, and have less orthographic variation. It means that the benchmark we created can successfully reflect the different properties of the text we used.

Regarding the analogy test, word embeddings generated from COLONIA perform much better than CIPM. We can see that certain grammar relations, especially those concerning verbs, are correctly identified. However, it seems that the SGNS model in COLONIA is not good at capturing the gender change and singular-plural change in nouns. Meanwhile, for the distributional models in CIPM, the accuracy is notably low. It indicates that the analogy test is rather difficult for distributional models generated from small and uncleaned data.

When we look into the other three datasets targeting semantic representation, the quality of word embeddings in COLONIA is better as well in all evaluation indexes. But, unlike the poor performance in the analogy test, we see that distributional models in CIPM also achieve comparable performance in intrinsic tasks examining semantic knowledge.

5 Conclusion

In this paper, we presented a new benchmark for assessing word embeddings in historical Portuguese that includes four intrinsic tasks.[5] The results of our tests reported here show that the cleaner corpus, COLONIA, led to better results, which confirms the findings of Hu et al. (2021) regarding two Spanish corpora. In that study, the medieval corpus (a smaller corpus, with high spelling and morphological variation) was likewise more challenging than the much larger and pre-processed corpus of contemporary Spanish. Moreover, learning from the creation process of existing intrinsic tasks, we find a more economic way to develop test datasets for assessing embedding models generated from historical languages. By bringing together a few experts in the history of Portuguese and a computational linguist, we have provided a feasible method to create multiple reliable intrinsic tasks using raw corpora. Unlike previous work that has relied just on one type of task, our benchmark has the advantage of targeting different lexical relations, thus providing a more thorough assessment. In the future, we plan on doing a more thorough comparison of different embedding methods and we hope that this benchmark can be used to test models for historical corpora of other languages.

References

Patrícia Amaral, Zuoyu Tian, Dylan Jarrett, and Juan Escalona Torres. Accepted. Tracing semantic change in Portuguese: A distributional approach to adversative connectives. *Journal of Historical Linguistics*.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

[5] The entire benchmark is available at https://github.com/zytian9/BAHP.

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive science*, 34(2):222–254.

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 7–12.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47.

José Camacho-Collados and Roberto Navigli. 2016. Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 43–50.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26.

Ivo Castro. 1991. *Curso de história da língua portuguesa*. Universidade Aberta.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Pablo Gamallo. 2018. Evaluation of distributional models with the outlier detection task. In *7th Symposium on Languages, Applications and Technologies (SLATE 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

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

Hai Hu, Patrícia Amaral, and Sandra Kübler. 2021. Word embeddings and semantic shifts in historical spanish: Methodological considerations. *Digital Scholarship in the Humanities*.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. Semeval-2014 task 3: Cross-level semantic similarity. In *SemEval@ COLING*, pages 17–26.

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

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18.

Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the workshop of ICLR*.

Hugo Gonçalo Oliveira, Tiago Sousa, and Ana Alves. 2020. Tales: Test set of portuguese lexicalsemantic relations for assessing word embeddings. In *Proceedings of the ECAI 2020 Workshop on Hybrid Intelligence for Natural Language Processing Tasks*.

Andreia Querido, Rita Carvalho, João Rodrigues, Marcos Garcia, João Silva, Catarina Correia, Nuno Rendeiro, Rita Pereira, Marisa Campos, and António Branco. 2017. Lx-lr4distsemeval: A collection of language resources for the evaluation of distributional semantic models of portuguese. *Revista da Associação Portuguesa de Linguística*, 3:265–283.

Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In *Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 975–980. USA.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (durel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23.

Natalie Schluter. 2018. The word analogy testing caveat. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 242–246.

Rodrigo Wilkens, Leonardo Zilio, Eduardo Ferreira, and Aline Villavicencio. 2016. B2sg: a toefl-like task for portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3659–3662.

Marcos Zampieri and Martin Becker. 2013. Colonia: Corpus of historical portuguese. *ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research*, 5:69–76.

Mengnan Zhao, Aaron J Masino, and Christopher C Yang. 2018. A framework for developing and evaluating word embeddings of drug-named entity. In *Proceedings of the BioNLP 2018 workshop*, pages 156–160.

Rate	Relation	Explanation
4	Very similar	The two words are synonyms (e.g., midday-noon or motherboard-mainboard).
3	Similar	The two words share many of the important ideas of their meaning but include slightly different details. They refer to similar but not identical concepts (e.g., lion-zebra or firefighter-policeman).
2	Slightly similar	The two words do not have a very similar meaning, but share a common topic/domain/function and ideas or concepts that are related (e.g., house-window or airplane-pilot).
1	Dissimilar	The two words describe clearly dissimilar concepts, but may share some small details, a far relationship or a domain in common and might be likely to be found together in a longer document on the same topic (e.g., software-keyboard or driver-suspension).
0	Totally dissimilar and unrelated	The two words do not mean the same thing and are not on the same topic (e.g., pencil-frog or PlayStationmonarchy).

Table 4: The five-point Likert scale used to rate the similarity of item pairs from Camacho-Collados et al. (2017)

A Word Similarity Annotation Scheme

Table 4 presents the annotation scheme used for the similarity test. The scheme comes from Camacho-Collados et al. (2017)'s paper and is designed by Jurgens et al. (2014).