Model evaluation for diachronic semantics: A view from Portuguese and Spanish

Amaral, Patrícia*; Hu, Hai**; Tian, Zuoyu*; Kübler, Sandra*
*Indiana University; **Shanghai Jiao Tong University

For research on semantic change that spans over several centuries, assessing the accuracy of embeddings comes with two challenges: (i) native speakers who can provide judgments about meaning are not available, and (ii) historical corpora are often much smaller than contemporary datasets, which raises issues of model accuracy (Hellrich, 2019; Hu et al., 2021). This paper presents the lessons learned from developing intrinsic evaluations to test the quality of distributional models used to investigate semantic change in Medieval Spanish and Portuguese. For Spanish we experimented on a 7 million word corpus (Chronicles corpus, with texts from 13th-16th c.) (Hu et al., 2021) and for Portuguese on a ca. 2,5 million token corpus, CIPM, with texts from 12th-16th c. (Tian et al., 2021).

The lessons learned include the following: 1) We cannot use tests developed for modern languages/corpora off the shelf, since the tests' vocabulary (e.g., capitals of the world, country names and currencies) does not overlap with that of the historical corpus.

We cannot use tests developed for other historical corpora without adaptations since those corpora tend to be restricted to specific domains, which also leads to a lack of overlap in vocabulary.

We need to account for spelling and morphological variation, which are important features of many Medieval corpora. For the historical Spanish corpus, e.g., we had to delete the test "adjective to adverbs" from contemporary Spanish (Cardellino, 2016), which maps an adjective to its corresponding adverb inmente, since the variability of forms of adverbs in Medieval Spanish would have resulted in more than one possible target form, including multi-word expressions (Company and Flores Da´vila, 2014). Instead, we added tests for several types of inflection (verbal morphology, gender and number in adjectives). The morphology tests were generated by using vocabulary based on the frequency counts from the Chronicles corpus. A summary of our analogy test is given in Table 1.

If the corpora are very small, using analogy tests alone may not provide enough information. Our work on the Portuguese corpus shows that using different tests that include a range of relations is important. The tests we created include: word similarity, outlier detection, and coherence assessment (see Table 2 for a summary). The latter is based on Zhao et al. (2018), who proposed a new evaluation method for assessing the quality of domain-specific word embedding models. They assume that the neighbors of a given word embedding should have the same characteristics of that word (e.g. neighbors of drug names should be drug names). In the Portuguese corpus, names of people and places are frequent, thus we can assess coherence by reporting the percentage of neighbors generated for a proper noun that were also proper nouns.

To summarize: Given the importance of register in research on semantic and syntactic change, as well as orthographic and morphological variation in historical corpora, specific tests are re- quired for a proper assessment of distributional models in studies of semantic change. Overall, assessment of word embeddings for historical research must meet the following criteria: appropriateness (corpus vocabulary is taken into account), sustainability (i.e. not requiring extensive expert input), comprehensiveness (tasks target different types of relations, i.e. syntactic, semantic, morphological), and complementarity (avoiding the biases of individual methods).

Source	Category	Example	#Questions
MTS	Morphology nouns: kinship terms	padre madre : hijo hija	506
	Morphology verbs: third person singular	comer come : ir va	650
	Morphology verbs: infinitive to participle	saber sabido : tomar tomado	1190
	Morphology verbs: gerund to participle	sabiendo sabido : tomando tomado	1190
ours	Morphology adj.: singular to plural	negra negras : rica ricas	992
	Morphology adj.: singular to plural	negro negros : rico ricos	992
	Morphology adj.: masc to fem	negro negra : negros negras	992
	Morphology adj.: masc to fem	negros negras : ricos ricas	992
	Morphology nouns : singular to plural	casa casas: capilla capillas	1332
	Morphology/Semantics: antonyms	feliz infeliz : posible imposible	42
	Semantics: antonyms	cerca lejos : bien mal	342
Total			9220

Table 1: Structure of our analogy test; MTS denotes the analogy test from Mikolov et al. (2013), translated into Spanish.

Test	Categories	#Questions
Analogy Test	nouns: gender; nouns: singular to plural; verbs: 1st person singular to 3rd person singular; verbs: 3rd person singular to 3rd person plural; verbs: infinitive to 3rd person singular; verbs: infinitive to gerund	2994
Word Similarity	synonymous; related (not synonymous); not related	97
Outlier Detection	body parts; Christianity; color; food; geography; parts of buildings; titles/professions; war	512
Coherence Assessment	proper nouns (names of people and places)	25

Table 2: Summary of the benchmark for assessing word embeddings generated for Medieval Portuguese

References

Cardellino, C. (2016). Spanish Billion Words Corpus and embeddings. Online at https://crscardellino.github.io/SBWCE/; retrieved August 2019.

Company, C. and Flores Dávila, R. (2014). Adverbios en mente. In Company, C., editor, Sintaxis histórica de la lengua española. Tercera parte: Adverbios, preposiciones y conjunciones. Relaciones interoracionales, pages 1195–1340. Fondo de Cultura Económica y Universidad Nacional Autónoma de México.

Hellrich, J. (2019). Word Embeddings: Reliability and Semantic Change. PhD thesis, Jena University Language and Information Engineering Lab.

Hu, H., Amaral, P., and Kübler, S. (2021). Word embeddings and semantic shifts in historical Spanish: Methodological considerations. Digital Scholarship in the Humanities, 37(2):441–461.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

In Proceedings of International Conference on Learning Representations (ICLR), Scottsdale, AZ.

Tian, Z., Jarrett, D., Escalona Torres, J., and Amaral, P. (2021). BAHP: Benchmark of assessing word embeddings in historical Portuguese. In Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 113–119, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Zhao, M., Masino, A. J., and Yang, C. C. (2018). A framework for developing and evaluating word embeddings of drug-named entity. In Proceedings of the BioNLP 2018 workshop, pages 156–160, Melbourne, Australia. Association for Computational Linguistics.