Advanced Morphology 2022

Impact of incorporating subword information on the relationship between wordform similarity and word vector similarity

Lina Conti and Isaac Murphy

1. Introduction

In Dautriche et al (2017), the authors question the commonly held idea that the relation between the form of a word and its meaning is completely arbitrary. They propose instead that, from an information-theoretic point of view, there are substantial ways in which the meaning of a word may affect its phonological form. Those include but are not limited to: frequency of use, acoustic salience, ease of articulation, and distinction from competing forms. To test whether there is a correlation between the form and semantics of a word, the authors compared words of the same length within each language for samples from 101 languages, using Levenshtein distance as a measure of formal similarity and Latent Semantic Analysis (LSA), computed from distributional word vectors, as a measure of semantic similarity. The authors found that there was a stronger connection between formal and semantic meaning than would be predicted by just chance, although they were unable to completely make sure that shared morphology/etymology were not impacting the results.

More recently, Bojanowski et al (2017) applied the idea of a link between formal and semantic similarity to the creation of FastText, a new neural model for learning word representations. The goal was to capture morphological features to improve the handling of unknown words. The model is a distributional space model with vectors representing complete wordforms, but also with vector representations of sequences of characters within each wordform — the so-called character n-grams or 'subwords'. The final vector representation of a wordform is the sum of the vectors for the full wordform and for some of the subwords it contains. While this model does achieve state-of-the-art results at approximating semantic similarity, whether the encoded subword information is actually capturing something about morphology remains an open question.

Our goal was to evaluate the FastText model to determine whether the use of subword information to train the model captures some generalizations about morphology or whether the noise generated by shared subwords in unrelated words results in artificially boosted similarity scores for words which are semantically unrelated but may share some formal features.

2. Methods

We tested the hypothesis that taking into account subword information would artificially boost the correlation between form similarity and vector similarity. We used Levenshtein (edit) distance as a measure of form similarity between words. To measure the similarity between vectors we used either cosine similarity or Euclidean distance.

We used pre-trained word vectors for English from two different FastText models by Mikolov et al. (2017). One represents words as bags of character n-grams (subwords) and the other directly computes a vector for each wordform. This way, we could see how using subwords or not for learning the word vectors would impact our results.

To get the word pairs we were going to compare, we sampled only from the 100,000 most frequent words in FastText models. This way, rare words were left out. The vector representation of rare words for the model that does not use subword information is not reliable. If a word does not appear often enough in the corpus, it is not possible to capture meaningful information about its distribution to encode in the vectors. Therefore, for rare words, vector similarity would not be a good proxy for semantic similarity, so we decided to leave them out.

From the 100,000 most frequent words, we started by randomly sampling two lists of 1,000 words each. To get each word pair, we randomly sampled a word from the first list and one from the second list. We used 10,000 word pairs in total. For each word pair, we computed the edit distance between the wordforms and the cosine similarity and the Euclidean distance between the vector representations of each word in each of the models.

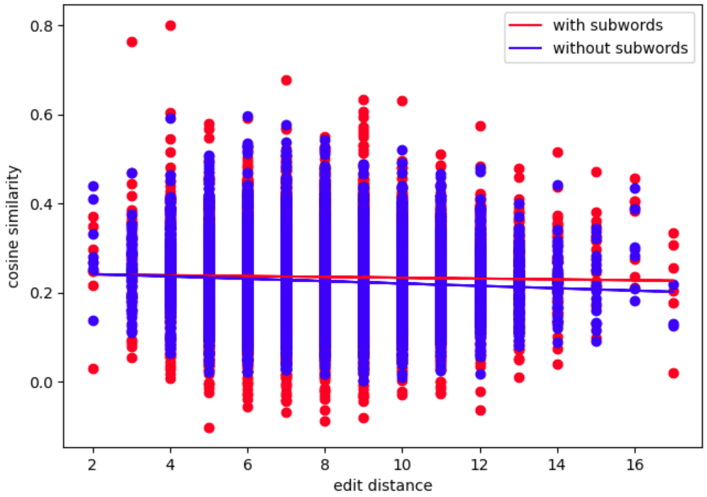
To model the relationship between wordform similarity and vector similarity, we fitted a linear regression model using SciPy, with edit distance as the explanatory variable and cosine similarity or Euclidean distance as the response variable. This was done separately for the vectors computed using subword information and the other ones.

Given that we wanted to know if any potential correlation we might observe was due to morphological reasons or to other factors, we decided to repeat the experiment using only pairs of morphologically related words. For this, we used the Categorial-Variation Database (Catvar) by Habash and Dorr (2003). We selected all word pairs that belong to the same derivational family according to Catvar and are not identical in form. We kept only the word pairs for which both words were present among the 100,000 most frequent words of the FastText models. We thus obtained 20,185 word pairs, from which we randomly sampled 10,000 so as to have the same sample size as for the previous experiment. Again, we computed edit distances, cosine similarities, Euclidean distances and fitted linear regression models for each vector representation model (with or without subword information).

3. Results

3.1. Random word pairs

3.1.1. Cosine similarity



	slope	rvalue	pvalue
using subword information	-0.000921965	-0.021801029	0.029258917
not using subword information	-0.002652044	-0.075027962	5.803178652e-14

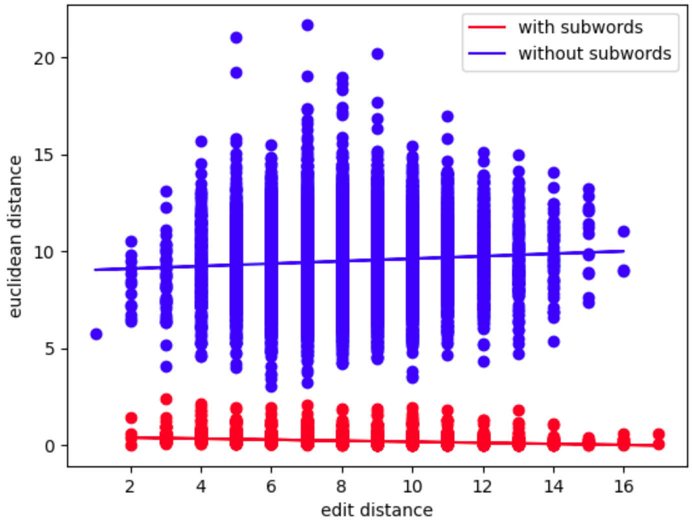
Figure 1: Cosine similarity between the vector representation of random pairs of words predicted from their edit distance. This graph combines a scatter plot of the real values measured from our sample and the linear regression line computed by SciPy.

Table 1: Slope, Pearson correlation coefficient, and p-value of the linear least-squares regression between the edit distance and the cosine similarity for random pairs of words for each of the word representation models, calculated with SciPy

Our results when modeling the relationship between edit distance and cosine similarity, for the randomly sampled frequent words, are shown in Figure 1 and Table 1. The Pearson correlation coefficients are close to zero for both models, so there does not seem to be a significant linear correlation between edit distance and cosine similarity. Nevertheless, the slopes, however small, go in the expected direction. We expected vector similarity to increase with form similarity. Since the slopes are negative, the similarity between vectors decreases as the edit distance between the wordforms increases. With p-values below .05, we can be confident in the statistical significance of these results.

For this experiment, the use of subword information did not seem to make a significant difference. In Figure 1, the values are spread out in a similar manner and the lines for both models are very close to each other. If anything, the slope is more negative for the model that does not use subword information, which does not support our hypothesis that including subword information would artificially boost the correlation between wordform and vector similarity.

3.1.2. Euclidean distance



	slope	rvalue	pvalue
using subword informatio n	-0.02698 8266	-0.31159 2159	4.736289 351e-224
not using subword informatio n	0.063971 818	0.068446 0009	7.293645 325e-12

Figure 2: Euclidean distance between the vector representation of random pairs of words predicted from their edit distance. This graph combines a scatter plot of the real values measured from our sample and the linear regression line computed by SciPy.

Table 2: Slope, Pearson correlation coefficient, and p-value of the linear least-squares regression between the edit distance and the Euclidean distance for random pairs of words for each of the word representation models, calculated with SciPy

Using Euclidean distance instead of cosine similarity as a measure of vector similarity, we obtain very different results for both models. The Euclidean distances between the vectors that were created with subword information are much lower than between those of the other model, as can be seen in Figure 2. The reason for this is unclear, but it could be linked to the fact that vectors for the first model are a sum of many vectors of subwords. One of the main differences between Euclidean distance and cosine similarity is that the first takes into account the magnitude of vector components whereas the second does not. Maybe summing the subword vectors somehow gives end-vectors with similar magnitudes.

Without subword information, the slope is positive, albeit small, which indicates that Euclidean distance increases with edit distance, as expected. However, for the vectors that use subword information, the slope is negative. So surprisingly, by this metric, vector similarity seems to decrease with wordform similarity here. But the Euclidean distances are all so similar for this model that it is difficult to draw any conclusion from them — except, maybe, that Euclidean distance is not the most adequate measure of vector similarity to use for the subword model.

3.2. Related word pairs

3.2.1. Cosine similarity

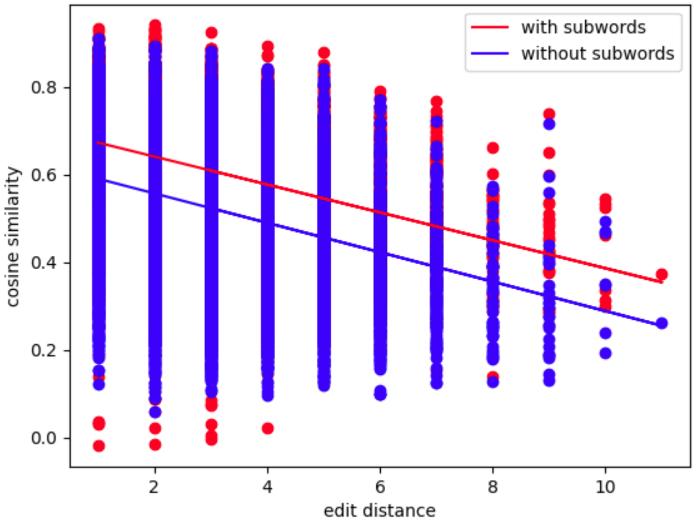


Figure 3: Cosine similarity between the vector representation of pairs of morphologically related words predicted from their edit distance. This graph combines a scatter plot of the real values measured from our sample and the linear regression line computed by SciPy.

	slope	rvalue	pvalue
using subword information	-0.031857516	-0.316556998	1.434908131e-231
not using subword information	-0.033546534	-0.295705676	5.931272986e-20

Table 3: Slope, Pearson correlation coefficient, and p-value of the linear least-squares regression between the edit distance and the cosine similarity for pairs of morphologically related words for each of the word representation models, calculated with SciPy

Figure 3 shows the result of comparing morphologically related words using cosine distance as the similarity measure. Compared to the results for random word pairs, morphologically related words show a much stronger relationship between edit distance and cosine similarity for both models. The slopes for the two models are negative, which matches our observations for random words, but the Pearson coefficient is significantly closer to -1, indicating a closer negative correlation between the two values. Notably, the similarity scores predicted by the regression model for vectors with subword information are always higher than for the vectors without it. This makes sense: it’s logical that morphologically related words will share more character n-grams than unrelated words and so an increase in similarity is not unexpected. It’s unclear from just the two datasets (random words and related words) whether this is an artificial increase in similarity due to related wordforms or whether the subword model does capture some morphological generalizations. What seems more important here is that the relationship between form and meaning is consistent across the two models. Again low p-values show a strong level of confidence in these results.

3.2.2. Euclidean distance

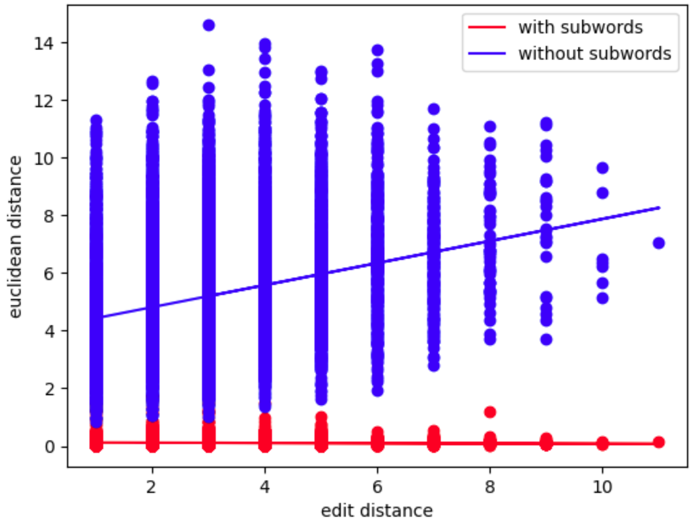


Figure 4: Euclidean distances between the vector representation of pairs of morphologically related words predicted from their edit distance. This graph combines a scatter plot of the real values measured from our sample and the linear regression line computed by SciPy.

	slope	rvalue	pvalue
using subword information	-0.004682622	-0.056030251	2.061858407e-08
not using subword information	0.382892711	0.265179220	1.5028816884963133e-160

Table 4: Slope, Pearson correlation coefficient, and p-value of the linear least-squares regression between the edit distance and the Euclidean distance for pairs of morphologically related words for each of the word representation models, calculated with SciPy

Figure 4 shows the results for related word pairs using Euclidean distance as the similarity measure. Similarly to the results for random word pairs, there is a much greater separation between the two models, with Euclidean distances much lower in the vectors from the subword model than for the ones without subword information. The similarity scores for the model without subword information are consistent with the cosine scores for related word pairs:

we see a similar Pearson coefficient which indicates that the relationship captured with the two similarity measures is approximately the same. However, for the subword model, the Euclidean distances are just as uninformative as they are for random word pairs: both the slope and r-value are close to zero, showing that any correlation between edit distance and Euclidean distance is very weak for this model. While the p-value for both models remains low, it is clear that Euclidean distance is once again not a good measurement for the subword model.

4. Discussion

4.1. Conclusion

Our goal with this project was to evaluate the effect of including subword information when training word vectors: does the subword information artificially increase similarity, or does it accurately capture morphological information? As shown in the results above, there is a small but significant correlation between edit distance and vector similarity for both models. This correlation is much stronger for morphologically related words than for unrelated words, but in both cases it is present with a high degree of confidence.

When using cosines as our similarity measure, the relationship between edit distance and similarity is consistent between models, both for random and morphologically related words. For related words, cosine similarity is generally higher but the rate of similarity is approximately equivalent for the two models. To know whether this similarity is artificially boosted or not we would need to also evaluate formally similar but semantically different word pairs, which was unfortunately outside the scope of this project.

When using Euclidean distance for similarity, the two models had very different distributions for both random and related word pairs. In general, the model without subword information behaved the same with Euclidean distances as it did for cosine similarity, but the scores for the subword model were all extremely similar among themselves and showed little to no correlation with edit distance. It is possible that this is a result of how these vectors were derived (i.e by summing subword vectors together), but what is clear is that Euclidean distance is not a good similarity measure for the subword model.

4.2. Future work

Our r-values being close to zero, linear regression does not seem to be the best way to model the phenomenon we are studying. It would be interesting to look for other regression models that better fit the data. Another possibility would be to keep using linear regression but to first apply some transformation on the vector similarity values so that their distribution is closer to a normal distribution.

Our data shows that the vector similarity for morphologically related words is globally higher when subword information is used to compute the vectors. However, our experiments are not sufficient to fully interpret this. This could mean that the subword model is capturing morphological information better than the other one. But this boost could also be artificial and comes just from having n-grams in common (formal similarity). To decide between the two interpretations, it would be interesting to run our experiment again, but only on pairs of words that are similar in form but not in meaning (not morphologically related), to see whether vector similarity for those is boosted by the subword model as well. However, a corpus of such words would be very difficult to collect, since it is hard to show that two words are "unrelated", be it in meaning or in morphology. The Catvar corpus, for example, contains many false negatives, with "use" and "reuse" not considered to be part of the same derivational family.

References

- Bojanowski, Piotr, et al. "Enriching word vectors with subword information." Transactions of the association for computational linguistics 5 (2017): 135-146.
- Dautriche, Isabelle, et al. "Wordform similarity increases with semantic similarity: An analysis of 100 languages." Cognitive science 41.8 (2017): 2149-2169.
- Habash, Nizar, and Bonnie Dorr. A categorial variation database for English. MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES, 2003.
- Mikolov, Tomas, et al. "Advances in pre-training distributed word representations." arXiv preprint arXiv:1712.09405 (2017).

Other materials

- All our code is available on Github