Multilingual semantic similarity

Yauheni Zviazdou
Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University
zviazyau@fel.cvut.cz, ezvezdov22@gmail.com

January 20, 2024

Abstract In certain foreign languages, a common issue arises when individuals incorrectly write words by omitting diacritics or altering letters. This problem is prevalent in social media, chatbots, and other informal written communications. As a result, embedding models face challenges in comprehending text without diacritics (hereinafter diacriticless). A potential solution involves adapting data representation to accommodate both formal and informal styles of writing. The objective of this study is to assess the quality of diacriticless models.

1 Introduction

This study relies on a corpus in the Czech language.

1.1 Problem explanation

Usual changes in informal texts

1. acute: á | é | í | ó | ú | ý → a | e | i | o | u | y
2. caron: č | ď | ě | ň | ř | š | ť | ž → c | d | e | n | r | s | t | z
3. overring: ů → u

Due to the absence of diacritics, the meanings of words may be compromised:

- byt (apartment) - být (to be)
- rád (to be glad) - řad (row or a line) - řád (order, religious order)
- krize (crisis) - kříže (crosses)

1.2 Datasets

The corpus constitutes a segment of a preprocessed Common Crawl repository, which encompasses raw data from web pages, as well as metadata and text extractions. Given the nature of this dataset, it is expected to include misspellings and text lacking diacritics.

1.2.1 Text preprocessing procedure

1. Eliminating duplicate entries.
2. Filtering out lines containing fewer than 9 characters.
3. Excluding URLs.
4. Breaking lines into individual sentences.
5. Converting all text to lowercase.
6. Removing lines containing words exceeding 30 characters.

7. Excluding HTML tags with less than 100 characters.
8. Removing lines surpassing 500 characters.
9. Omitting words longer than 21 characters.

1.2.2 Data Conversion

Initially, we need to generate a diacritic-free corpus. Before diacritic removal, I convert all characters to lowercase (to handle non-preprocessed text), thus establishing the sequence

original text → *lowercase text* → *diacriticless text*

```
sed 's/.*/\L&/' "$1" | iconv -f utf-8 -t ascii//TRANSLIT > diacriticless/"$1"
```

Listing 1: Script for removing diacritics using Unix utilities

By employing the Script 1, two datasets are generated:

- Diacritics - text with diacritics
- Diacriticless - text without diacritics

1.2.3 Preprocessed corpus summary

- The corpus contains 3.87 billion words.
- The diacritics model has 800K words in the dictionary.
- The diacriticless model has 762K words in the dictionary.

1.3 Embedding model library

For training models in this research, I will utilize the FastText.cc library [2].
The parameters for training are

- Word representation: **Word2vec**
- Vector size: **300**
- Loss: **Negative sampling loss**
- Dictionary threshold (the frequency of the word to be included in the dictionary): **130**

1.4 Metric

1.4.1 Analogies [4]

The metric enables the comparison of the accuracy of word analogies produced by the embedding model. The metric comprises 22,256 tests across 12 categories, as indicated in Table 1.

Category	Number of tests
Antonyms-nouns	1406
antonyms-adjectives	1722
antonyms-verbs	1120
state-presidents	1122
states-cities	1122
family-relations	810
nouns-plural	1332
jobs	1188
verb-past	8928
pronouns	756
antonyms-adjectives,gradation	1560
nationalities	1190
All	22256

Table 1: Information about Analogies tests

The metric results include:

- **Total accuracy**, which represents the percentage of passed tests.
- **Semantic similarity**, measuring how similar words are in meaning.
- **Syntactic accuracy**, indicating the degree to which a piece of text adheres to the grammatical rules of a language.

1.4.2 UPV FAQ [1]

The metric estimates the percentage of correctly answered questions in the test set. These tests are comprised of frequently asked questions (FAQs) from the Industrial Property Office of the Czech Republic, organized into four distinct groups, as detailed in Table 2.

Tests set	Number of tests
FAQv5	2054
FAQ50	561
FAQ76	2025
FAQ76v2	1965
All	6605

Table 2: Information about UPV FAQ tests

2 Testing

2.1 Baseline

Table 3 displays the outcomes of assessing the original (diacritics) model on text lacking diacritics. The results indicate a less satisfactory performance; however, we will juxtapose them with the original results in the subsequent analysis.

Category	Tests passed, %
Analogies Total accuracy	35.30
Analogies Semantic accuracy	4.20
Analogies Syntactic accuracy	32.42
Question matching accuracy	81.10
Answer matching accuracy	29.23

Table 3: Testing diacritics model on diacriticless tests.

2.2 Testing diacriticless model on diacritics tests

Table 4 indicates that the results fall below the baseline. This is because conducting training and testing on distinct types of data is not meaningful. The word vectors are formed using word subwords, and when diacritics are removed, these vectors differ significantly from the original ones.

Category	Tests passed, %
Analogies Total accuracy	17.84
Analogies Semantic accuracy	2.75
Analogies Syntactic accuracy	16.12
Question matching accuracy	83.31
Answer matching accuracy	28.64

Table 4: Testing diacriticless model on diacritics tests.

Conclusion To enhance performance, it is recommended to employ diacriticless tests for evaluating the diacriticless model and tests with correctly spelled diacritics for assessing the diacritics model.

2.3 Testing on the same data type as model

2.3.1 Analogies metric

Category	Diacritics, %	Diacriticless, %	Difference, %
Antonyms-nouns	10.60	9.96	↓ 0.64
antonyms-adjectives	8.07	7.78	↓ 0.29
antonyms-verbs	7.41	6.88	↓ 0.53
state-presidents	0.00	0.00	– 0.00
states-cities	9.45	7.49	↓ 1.96
family-relations	25.68	24.32	↓ 1.36
nouns-plural	69.60	69.37	↓ 0.23
jobs	88.81	77.86	↓ 10.95
verb-past	81.30	78.72	↓ 2.58
pronouns	11.91	11.11	↓ 0.80
antonyms-adjectives,gradation	87.50	87.50	– 0.00
nationalities	42.61	36.47	↓ 6.14
All	53.41	51.19	↓ 2.22

Table 5: Detailed results of Analogies metric on dataset

Table 5 reveals that the *Jobs* and *Nationalities* categories exhibit the most substantial differences. These categories adhere to the following format:

noun masculine – noun feminine → noun2 masculine – ?

The big gap between the diacritics and diacriticless model may be due to the big similarity of the masculine and feminine noun variants in the Czech language, for example *Lékař* (doctor, masculine) and *Lékař**ka*** (doctor, feminine)

2.3.2 UPV FAQ metric

Results from Table 6 are almost the same as Baseline.

	Question matching accuracy			Answer matching accuracy		
Tests set	Diacritics, %	Diacriticless, %	Difference, %	Diacritics, %	Diacriticless, %	Difference, %
FAQv5	83.11	83.65	↑ 0.54	21.27	22.92	↑ 1.65
FAQ50	88.61	88.26	↓ 0.35	27.05	30.25	↑ 3.2
FAQ76	79.66	79.96	↑ 0.30	33.46	33.51	↑ 0.05
FAQ76v2	80.77	80.93	↑ 0.16	34.18	34.13	↓ 0.05
All	83.04	83.20	↑ 0.16	28.99	30.20	↑ 1.21

Table 6: Detailed results of UPV FAQ metric on dataset

2.4 Comparison

Category	Diacritics, %	Diacriticless, %	Difference, %
Analogies Total accuracy	53.41	51.19	↓ 2.22
Analogies Semantic accuracy	10.20	9.40	↓ 0.80
Analogies Syntactic accuracy	63.61	60.17	↓ 3.44
Question matching accuracy	83.04	83.20	↑ 0.16
Answer matching accuracy	28.99	30.20	↑ 1.21

Table 7: Total testing results on dataset

Table 7 illustrates that both models perform better than the baseline. The diacriticless model exhibits slightly lower results in analogy tests, but it excels in Q&A matching quality.

3 Conclusion

The Diacriticless dataset is suitable for training an embedding model. Therefore, there are two potential solutions to the problem.

1. Apply algorithms to reintroduce diacritics to diacriticless or misspelled text, and then utilize the diacritics model.

   Text correction → *Text processing*

   Disadvantages

   - Worse Q&A matching performances.
   - The corpus containing diacritics requires more storage space compared to the Diacriticless corpus.

2. Eliminate all diacritics from the text and employ the diacriticless model. Subsequently, apply an algorithm to the model's output to reintroduce diacritics.

   Removing diacritics → *Text processing* → *Adding diacritics*

   Disadvantages

   - Worse performances in analogies tests.
   - The requirement to employ two algorithms for text processing.

References

[1] Adam Jirkovský. Upv faq semantic text similarity test. https://github.com/jirkoada/upv_faq. [Online; accessed 4-January-2024].

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[3] Meta Research. fasttext - library for efficient text classification and representation learning. https://fasttext.cc/. [Online; accessed 4-January-2024].

[4] Lukás Svoboda and Tomás Brychcín. New word analogy corpus for exploring embeddings of czech words. *CoRR*, abs/1608.00789, 2016.