MEASURING SIMILARITY BY LINGUISTIC FEATURES RATHER THAN FREQUENCY

RODOLFO DELMONTE, NICOLÒ BUSETTO

CA FOSCARI, UNIVERSITY OF VENICE, ITALY

ISA-18 WORKSHOP – LREC – MARSEILLE 20 JUNE 2022

OVERVIEW

- *In this work we produced an experiment to verify whether* ***word absolute frequency*** *- as reported in a frequency word list and represented in a dictionary for the creation of Deep Learning models - is more liable of the inability of BERT to predict the* **masked word** *THAN* ***word relative frequency*** *as represented by similarity cosine measures for vectors of word embeddings extracted from huge corpora.*

OVERVIEW 2

- The idea was to analyze the effect of CONTEXT and severe it from DICTIONARY. To this aim we needed to evaluated separately ABSOLUTE from RELATIVE frequency of each word form.
- We ran the Italian version of BERT – UmBERTo based on RoBERTa – available under Huggingface and we
  - Analyzed in detail the output of the FIRST LAYER, the RAW EMBEDDINGS

OVERVIEW 3

- Eventually we **measured** BERT's ability to **predict** the **MASKED** word given the sentence and the text in which it was contained
  - Similarity was measured at first solely on the basis of cosine values
  - This was then substituted by a linguistically based measure which extended the notion of similarity
  - In this way, predictability was given an overall parameter which better reflected the variables brought onto the experiment

CORPUS AND CANONICITY

- To this aim, we created a small corpus of 18 Italian sentences including both poetry - 7 sentences - and newswire texts - 11 sentences, in which the syntactic structures were non-canonical or discontinuous.
- For the experiment, the same sentences were then duplicated manually modifying the structure into a canonical configuration.

TWO IMPORTANT FACTORS: SYNTACTIC STRUCTURE AND WORD FREQUENCY

- The corpus thus contained two important factors useful for proving our hypothesis: use of uncommon words and presence of infrequent syntactic configurations.
- Then the two different genres contributed the two factors present at the same time: but in poetry in great measure, in newswire texts with less infrequent structures and and less uncommon words.

ABSOLUTE FREQUENCY IN THE DICTIONARY AND SUBWORD UNITS

- We extracted the first 50K entries in the frequency list of the corpus used to build the model for UmBERTo and the frequency associated with the last item was 1377
  - After checking all masked words, only a small number of word forms were excluded from the dictionary (10 over 236, but an additional 20 had "lower" frequency)
  - From the 10 VERY LOW frequency, 3 were Out Of Vocabulary Words and 1 was absent from the overall list
  - These had to be segmented by the tokenizer into subword units with an overall loss of semantics in the embedding containing the corresponding word form

THE EXPERIMENT: MASKED WORDS

- From the tables included in the paper the number of masked words varies as follows:
  - 54 over a total of 83 word forms in the poetry subset of sentences – 0.65
  - 86 over a total of 153 word forms in the newswire subset – 0.56
- The number of masked words per genre reflects the distribution of number of function and content words which varies in the two genres: poetry has more content and less function words. The opposite is true for the newswire domain.

THE EXPERIMENT: PREDICTED MASKED WORDS

- Now the number of masked words correctly predicted by BERT:
  - 19 over 54 word forms in the poetry subset – 28.4%
  - 69 over 86 word forms in the newswire subset – 80.2%
- Splitting the numbers per canonicity we obtain the following ratios:
  - For the poetry domain: 7 in non-canonical to 12 which amounts to 0.58
  - For the newswire domain: 31 in non-canonical to 38 amounting to 0.82

THE EXPERIMENT: FIRST CONCLUSION

- As a preliminary conclusion we can safely say that:
  - It is much harder to predict the masked word in the poetry domain
  - The contribution of the manually created canonical sentence structure is much higher in the poetry domain than in the newswire subset
- The question we must now answer is: are these two facts the effect of **CONTEXT** (i.e. *relative frequency*) or **FREQUENCY** (i.e. *absolute frequency*) or **BOTH**?

CHECKING FREQUENCY: RELATIVE OR ABSOLUTE?

- In the tables reported in the paper we show the number of HIGH, LOW and VERY LOW frequencies for the two domains and the results are as follows:
  - The ratio of HIGH to LOW+VL for the poetry subset is = 0.82
  - The ratio of HIGH to LOW+VL for the news subset is = 0.14
- In other words, the number of low frequencies in the poetry domain is almost equal to the number of high frequencies. In the news domain it is just the opposite case: the low frequencies are by far the majority.

THE EXPERIMENT: SECOND CONCLUSION

- The result obtained from the evaluation of absolute frequencies confirms the result obtained from cosine measures as derived from the first or raw layer:
  - The newswire domain sentences have a predictability score MAINLY determined by relative frequencies, i.e. THE CONTEXT
  - On the other hand, the poetry domain has a predictability score MAINLY determined by absolute frequencies, i.e. THE DICTIONARY

THE LINGUISTIC EVALUATION: DECOMPOSING SIMILARITY

- We introduced a method for the evaluation of embedding vectors where the masked word was not found in the first ten candidates. Rather than just considering cosine measures we weighted the output word on the basis of its linguistic overall features. In this way similarity was decomposed into a subset of features: morphological, lexical, syntactic and semantic.
- This was done in order to reduce the gap between the two frequency domains, the context and the dictionary and come up with a single score

THE LINGUISTIC EVALUATION: DECOMPOSING SIMILARITY

- The results are reported in the two Tables in the last column and are represented as before normalized by number of masked words:
  - News domain: 56.76/86 = 0.66
  - Poetry domain: 24.78/34 = 0.46

THE LINGUISTIC EVALUATION: CONFIRMING OUR HYPOTHESIS

- So the hypothesis was that poetry text would be highly unpredictable and remains so also in its canonical version, with some improvements though.
- Newswire texts being much more predictable and the contribution of canonical version would not increase significantly the score.

THE LINGUISTIC EVALUATION: DECOMPOSING SIMILARITY

- The news domain has a much higher predictability score than the poetry domain because:
  - It has a higher number of contextually predictable words in non-canonical structures (and canonical)
  - It has a much lower number of LOW frequency words
- AND it has an overall HIGHER linguistically based evaluation score

THE LINGUISTIC EVALUATION: DECOMPOSING SIMILARITY

- The poetry domain has a much lower predictability score than the news domain because:
  - It has a lower number of contextually predictable words in non-canonical structures (with a slight increase in canonical)
  - It has a much higher number of LOW frequency words
- AND it has an overall LOWER linguistically based evaluation score

CONTEXT AND COMPOSITIONAL MEANING

- Eventually, POETRY combines words to produce NON-LITERAL meaning composition like METAPHORS, which are highly unpredictable
- NEWS on the contrary, is more akin to produce word sequences and sentence structure that can be readily understood and obey LITERAL meaning composition

Thanks

Questions?