Knowledge-Base Enriched Word Embeddings for Social Media Text Analysis

VYTAUTAS MAGNUS UNIVERSITY
Faculty of Informatics

M. Petkevičius[1], D. Vitkutė-Adžgauskienė[2]
Vytautas Magnus University [1]mindaugas.petkevicius@vdu.lt [2]daiva.vitkute-adzgauskiene@vdu.lt

ABSTRACT

Performance of deep learning models for social media text analysis tasks depends on the quality of pre-defined word embeddings. Embeddings, that are obtained from regular public domain texts, such as Wikipedia, do not reflect the specifics of social media language with irregular words and jargon. Use of social texts for embedding training is problematic, as they are typically brief (comments, forum posts, etc.), thus limiting the context.
We demonstrate the advantage of semantically enriched word embeddings for solving natural language processing (NLP) tasks. We use a retrofitting approach for an intrinsic word similarity evaluation task, using semantic information from a linguistic LitWordNet ontology for word vector pre-training by pulling semantically related words closer together.
Dual convolutional neural network (CNN) with additional semantic layer is applied for solving more complex extrinsic social text classification tasks.

METHODS

RETROFITTING - ENRICHING WORD EMBEDDING MODELS

The retrofitting approach is used, adding information from semantic knowledge bases, and, thus, pulling similar pre-trained FastText word vectors closer together. The method leverages cosine similarity between a word and each of its synonyms for adjusting corresponding word vectors. Figure 1 shows an example of original vectors (in blue) vs the retrofitted vectors (in red).

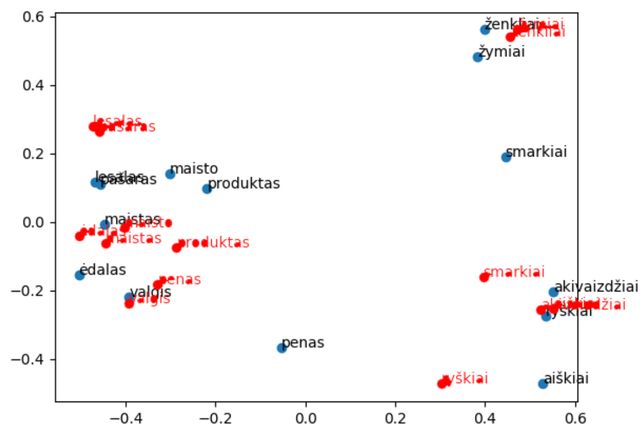

Figure 1. Word vectors before and after retrofitting

ADDING ADDITIONAL SEMANTIC LAYER TO A DEEP LEARNING MODEL

A convolutional neural network (CNN) model is supplemented with an additional semantic layer. We combine two CNN networks: one for traditional word embeddings, and another for semantic information. The semantic information layer can compensate for sparse contextual information in brief social texts, thus leading to better accuracy in text classification tasks, network model is shown on Figure 2.

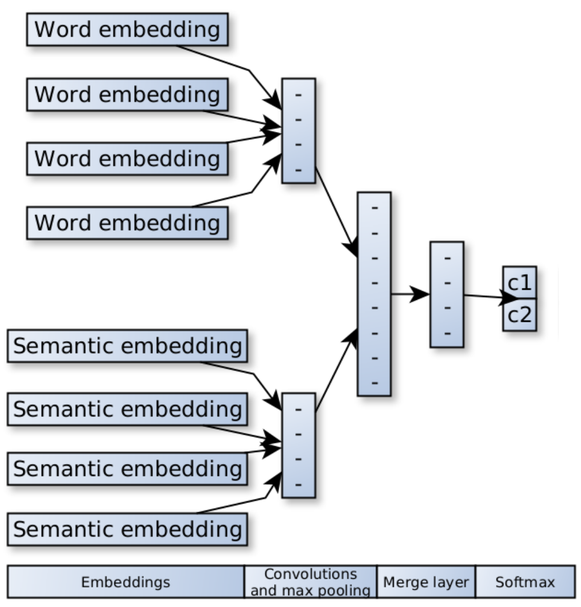

Figure 2. Dual CNN model with semantic embeddings

ENRICHMENT

A deep learning model is capable of capturing several information layers. Our hypothesis is that by incorporating a semantic layer into the deep learning model, we can improve contextual information, which is typically sparse in brief, irregular social media texts, thus increasing accuracy in domain-specific text classification tasks.

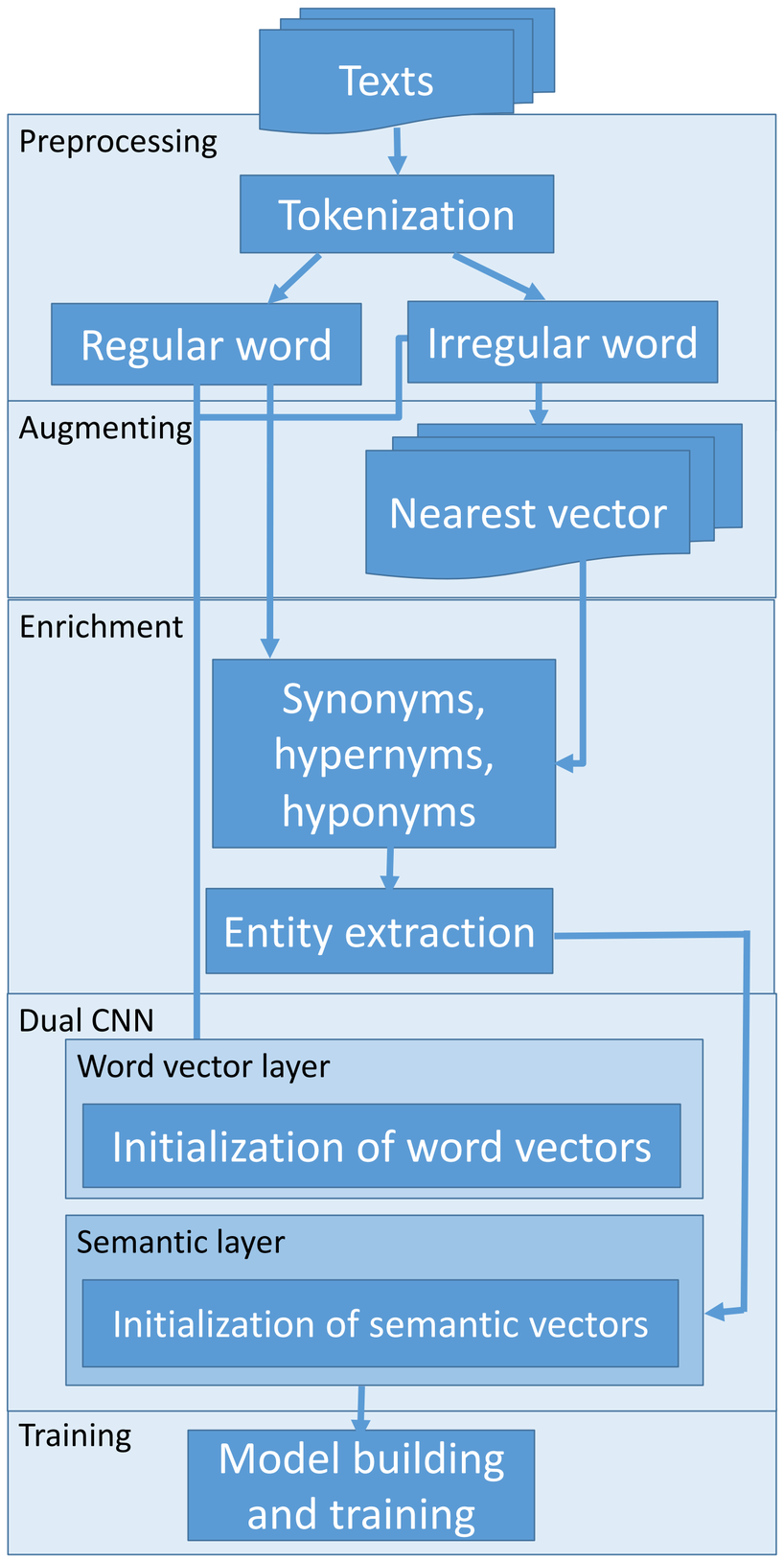

Figure 3. The enrichment process

Word categorization by the nearest word vector

In case of a misspelled or irregular word, we can use the word embedding model for obtaining the nearest words (vectors), and then try to categorize this word using typical NER categories.
E.g., Word '*pamaskvė*' -> [*pamaskvėje*, *Kamčatka*, *Peterburgų, etc*], NER -> '**location**'
Word '*minde*' -> [*Mindaugas*, *Kestas*, *Raimis*, *etc*]
NER -> '**person**'

DATA RESOURCES

FastText embeddings (Lithuanian social text corpus 15m, Lithuanian media corpus 60m, Wikipedia 22m words)

- LitWordNet ontology:
  - ✓ synsets (50048)
  - ✓ word definitions (43070)
- Dictionaries related to social media language specifics:
  - ✓ jargon dictionary (6452)
- Domain-specific semantic data sources - NER knowledge base.

RESULTS

WORD SIMILARITY TASK (INTRINSIC)

The enriched word representations were evaluated against the best rated Symlex999 benchmark for a word similarity task.
The cosine similarity between the vectors of two words is computed for obtaining the Spearman's rank correlation coefficient between the retrofitted model ranks and human rankings.
When comparing Spearman correlation results for traditional embeddings and knowledge-base enriched word embeddings, we observe an increase in Spearman correlation accuracy, as shown in Table 1.

Table 1. Benchmarking results

Model	*SimLex999*	ρ *diff %*
FastText $_{\text{Similarity, jargon, definition}}$	0.4059	+2.56%
FastText $_{\text{Similarity}}$	0.4213	+6.12%
FastText $_{\text{Original}}$	0.3955	-

TEXT CLASSIFICATION TASK (EXTRINSIC)

For text classification task, the training dataset included 1000 social media comments, covering different event types:

- global events;
- local events;
- elections;
- COVID-19 related events, etc.

Additionally, 2000 comments and posts unrelated to any specific events, were used.

The purpose of the text classification experiment was to distinguish between posts related to a domain-specific event or events, and those not related to events.
Events are typically identified by domain specific named entities, e.g., "*Vakar Lukašenka nuleido Ryenair lėktuvą*."
Classification results for a typical CNN model and for the suggested dual CNN model with an additional semantic layer are presented.

Table 2. Experiment results

Model	*Precision*	*Recall*	*F1*	*F1 diff %*
Typical CNN	0.718	0.740	0.729	-
Dual CNN with semantic layer	0.751	0.799	0.774	+5.81%

Precision, recall, as well as F1 score were improved when using the enriched embeddings and a Dual CNN model.

CONCLUSIONS

The results, of our study show, that the enrichment of word embeddings with additional semantic information from external knowledge bases can be beneficial for both simple intrinsic language tasks, such as word-level similarity evaluation, as well as for complex extrinsic text analysis tasks, such as domain-specific text classification for social media texts.

FURTHER RESEARCH

Further research intends to investigate the performance of transformers (BERT) when using semantically enriched data. The challenges in this case are related to the availability of specific Lithuanian language transformers and multilingual model incompatibility with social media texts.

12th Conference **Data Analysis Methods for Software Systems**, Druskininkai, Lithuania,