ULL: Assignment 3: Evaluating Sentence Representations

Melissa Tjhia (10761071)
tjhiamelisa@gmail.com

Richard Olij (18033730)
olij.richard@gmail.com

1 Introduction

In this report, two different word embedding models are compared with each other: the skip-gram model and the Embed-Align model. This is done with classification tasks, natural language inference tasks and a semantic relatedness task from the SentEval toolkit by evaluating sentence embeddings. The skip-gram model worked better on the natural language inference tasks, while the Embed-Align model worked better in tasks in which a better understanding of the semantic meaning is more important.

2 Background

The purpose of word embeddings is to capture (the meaning of) a word, usually in the form of a vector. Words occur in a certain context, and can thus be described by their context, which means that words that are similar in context, will have more similar representations than words that do not. These word representations can be used in different tasks, such as the word similarity task and the word analogy task which were done in the first assignment. A method to obtain word embeddings is the skip-gram model, which predicts the context, given a target word. The intuition of the CBOW model is the opposite: given a context, the target word should be predicted. (Mikolov et al., 2013a) Another method is the Bayesian skip-gram model, which uses a distribution to encode the meaning of the word given a context (Brazinskas et al., 2017). Lastly, the Embed-Align model by Rios et al. (2018) also uses distributions to represent words, but it also uses aligned sentences from two languages to remove ambiguity.

3 Methods

To capture meaning, two approaches are used. The skip-gram model basically captures the meaning of a word by making tuples of the *target word* with any individual *context word* within a certain *window size* and then given the target word, the model should predict the context word. Embedding dimensions of 100 and 300, as well as window sizes of 2 and 5 are used for the skip-gram model. However, target words could be ambiguous in that language. Therefore Embed-Align uses aligned data to learn the representations with two languages since it is assumed that the target word is not ambiguous in the second language. A trained Embed-Align model was provided.

3.1 Skip-gram

For the skip-gram model, the gensim `Word2Vec` library (Řehůřek and Sojka, 2010) was used to train. To have comparable results, the data on which the Embed-Align model was trained should be comparable. It is unclear how this model was trained exactly, yet the vocabulary does consist out of 71.578 words and is supposed to be trained on the English Europarl dataset. To get as close to that as possible, the data was lowercased of which stop words[1], punctuation marks and words that occurred 2 times or less are mapped to `UNK`. Appendix A Figure 1 demonstrates that this results into 70,426 unique words, which were 155,443 originally.

As mentioned, embedding dimensions of 100 and 300 are used. Mikolov et al. (2013a) demonstrated

[1] Stopwords from `nltk.org/book/ch02.html`.

that these dimensionalities perform rather differently whereas increasing or increasing this would not differ significantly while the complexity will. Furthermore, window sizes of 2 and 5 are used to demonstrate the influence of a larger context. Our first ULL assignment concluded that Bag Of Words models with window size 2 performs better for the SimLex dataset while yielding less contextual information than a window size of 5 which performs better for the MEN dataset. To explore the influence of context for the tasks as explained in Section 5 more broadly, both window sizes are evaluated. Finally, all skip-gram models are trained for 20 epochs where the learning rates will linearly decrease from 0.025 to 0.001 (Mikolov et al., 2013a), with the default batch size of $10,000$ words.

3.1.1 Negative Sampling

Training a skip-gram is usually expensive, because it needs to sum over all the words in the vocabulary when computing the softmax. Mikolov et al. (2013b) implemented the skip-gram and used negative sampling to speed up the learning process. Instead of the whole dataset, only a random subset is used to sum over. They found that 5-20 samples worked good for small training sets, while 2-5 would be sufficient for bigger datasets. The Europarl dataset is considered big, and therefore 5 random samples were used for negative subsampling.

3.2 Embed-Align

As explained, a pretrained Embed-Align model is given. This model is trained on 2 million sentences (maximum length of 50) of the English-French Europarl dataset. The English vocabulary size is 71,578 and the French vocabulary size is 89,906. Tokens that only occurred once were mapped to `UNK`.[2] This model is trained for 10 epochs.

4 Evaluation

The SentEval toolkit consists of several tasks in which sentence representations are evaluated (Conneau and Kiela, 2018). For this assignment, downstream tasks are performed in which the embeddings are evaluated as features. Most of the tasks that are used are classification tasks, other tasks are natural language inference and semantic similarity tasks. For the classification tasks, a prediction is made based on (a part of) one sentence. For the other types, the prediction is about the relation between a premise (a sentence or an image) and a hypothesis (a sentence). The classification tasks return accuracies, the natural language inference task also gives the F1-score and the semantic similarity task returns the Pearson and Spearman correlations. The models' word2id mappings and word vectors are used, and a sentence embedding is obtained by simply averaging over the word vectors of a sentence.

5 Results and Analysis

Six classification tasks (MR, CR, MPQA, SUBJ, SST Binary Classification (SST2) and TREC), two natural language inference tasks (MRPC and SICK-Entailment (SICK-E)) and one semantic relatedness task (STS14) were performed. A brief description of each task and the results of the models are presented in the following subsections.

5.1 MR

In the MR task, the goal is to predict whether a movie review is positive or negative. Table 1 shows the results on the MR task. Embed-Align is outperformed by all versions of the skip-gram model, substantially. Changing the window size of the skip-grams from 2 to 5 does not improve the accuracy as much as changing the embedding size from 100 to 300 does.

	Dev acc.	Test acc.
Skip-gram-2-100	68.51	68.75
Skip-gram-5-100	68.18	68.17
Skip-gram-2-300	70.77	70.55
Skip-gram-5-300	71.00	70.91
Embed-Align	58.50	56.94

Table 1: The accuracies for the MR task.

5.2 CR

The CR task is similar to the MR task, but instead of movies, the sentiment of products reviews are classified. The results for the CR task are presented

[2]Note that we previously mentioned that we were unsure about the configurations and decided to map not just occurrences of one, but also two to `UNK`. Appendix A Figure 1 shows that removing only once occurring words yields a vocabulary size of roughly 90,000 instead of the 70,000 mentioned.

in Table 2. Even though the task is rather similar as well as the influences of window sizes and embedding sizes, the Embed-Align is performing more similar to the skip-gram methods, but still does not outperform them.

	Dev acc.	Test acc.
Skip-gram-2-100	72.74	72.53
Skip-gram-5-100	73.55	73.46
Skip-gram-2-300	75.13	74.86
Skip-gram-5-300	75.63	75.55
Embed-Align	70.64	70.65

Table 2: The accuracies for the CR task.

5.3 MPQA

For the MPQA task, an opinion needs to be classified as positive or negative. Even more than with the CR task, the Embed-Align is performing more similar to the skip-grams while still not outperforming them, as shown in Table 3. Also, the influence of the parameters of the skip-gram model decreases.

	Dev acc.	Test acc.
Skip-gram-2-100	84.60	84.44
Skip-gram-5-100	84.80	84.94
Skip-gram-2-300	85.47	85.52
Skip-gram-5-300	85.27	85.33
Embed-Align	83.89	83.78

Table 3: The accuracies for the MPQA task.

5.4 SUBJ

The SUBJ task is different than the previous tasks. Instead of predicting a positive or negative class, the goal is to classify whether a sentence is objective or subjective. This demands more actual understanding of a sentence instead of meaning of (almost just) individual words as with the tasks so far. As shown in Table 4, the Embed-Align model scores very high on the SUBJ task – almost perfectly – whereas the skip-gram models perform significantly lower.

5.5 SST Binary classification

The SST Binary classification task is similar to the MR task, but performed on a different datset, namely the Stanford Sentiment Treebank (hence the name SST). As shown in Table 5, similarly to the MR task, the Embed-Align model is outperformed on the SST2 task and also the difference in embedding size is more present than for the other sentiment analysis tasks.

	Dev acc.	Test acc.
Skip-gram-2-100	82.31	82.17
Skip-gram-5-100	82.68	82.86
Skip-gram-2-300	85.26	85.09
Skip-gram-5-300	85.30	85.27
Embed-Align	99.60	99.60

Table 4: The accuracies for the SUBJ task.

	Dev acc.	Test acc.
Skip-gram-2-100	69.38	69.69
Skip-gram-5-100	69.15	70.13
Skip-gram-2-300	73.05	72.76
Skip-gram-5-300	73.85	72.98
Embed-Align	67.20	67.11

Table 5: The accuracies for the SST Binary classification task.

5.6 TREC

Unlike the previous tasks, which all are binary classification tasks, the TREC task is a multiclass classification task in which a question's type is predicted. Table 6 demonstrates that all models perform poorly. Increasing the embedding size increases the performance more than with any other task in this report, indicating that having many distinctive word meanings is especially important. However, a larger window size does harm the accuracy indicating that a larger context is bad. As expected from the observation so far, the semantic and context driven Embed-Align model, performs worst.

	Dev acc.	Test acc.
Skip-gram-2-100	59.12	61.40
Skip-gram-5-100	58.14	59.40
Skip-gram-2-300	65.20	66.60
Skip-gram-5-300	63.37	62.40
Embed-Align	53.39	56.80

Table 6: The accuracies for the TREC task.

5.7 MRPC

In the MRPC task, the premise and hypothesis are both sentences and the goal is to predict whether the sentences are paraphrases. The results in Table 7 demonstrate very similar results and thus there is no clear conclusion besides that context, embedding sizes and semantics are all not (more) essential than the other.

	Dev acc.	Test acc.	F1
Skip-gram-2-100	70.85	71.77	81.18
Skip-gram-5-100	70.76	71.42	80.99
Skip-gram-2-300	70.22	71.65	80.82
Skip-gram-5-300	70.49	71.71	81.22
Embed-Align	70.64	70.96	80.10

Table 7: The accuracies and the F1 test scores for the MRPC task.

5.8 SICKEntailment

The premise and the hypothesis in the SICKEntailment tasks are also both sentences. In this task, the goal is to predict the relation between the sentences. The Embed-Align model scores the highest on the SICKEntailment task, as presented in Table 8. Also, as with the TREC task, a larger context window will harm the accuracy.

	Dev acc.	Test acc.
Skip-gram-2-100	69.00	68.74
Skip-gram-5-100	67.60	68.07
Skip-gram-2-300	71.40	70.57
Skip-gram-5-300	68.20	69.35
Embed-Align	72.60	74.75

Table 8: The accuracies for the SICKEntailment task.

5.9 STS14

Similarly to the previous two tasks, the premise and the hypothesis are also sentences. The previous tasks all used some sort of pre-defined labels. In the STS14 task, the semantic relatedness between the two sentences is given a decimal score between 0 and 5. The weighted[3] averages of the Pearson and Spearman correlations for the STS14 task are presented in Table Table 9. Again, using an embedding size of 300 for the skip-gram model shows an improvement. Although a positive correlation is found for the Embed-Align model, its performance is the worst.

[3]The averages are weighted by the amount of samples.

	Pearson	Spearman
Skip-gram-2-100	0.6476	0.6257
Skip-gram-5-100	0.6524	0.6293
Skip-gram-2-300	0.6694	0.6424
Skip-gram-5-300	0.6720	0.6424
Embed-Align	0.5951	0.5866

Table 9: The weighted averages of the Pearson and Spearman correlations for the STS14 task. Note that the correlations are between -1 and 1 where 0 means no correlation.

6 Conclusion

Four skip-gram models, which differed in the window sizes and the embedding dimensions, were evaluated. Overall, increasing the embedding dimension from 100 to 300 showed more significant changes than increasing the window size from 2 to 5, which did not necessarily lead to an evident change in most results. For the sentiment analysis tasks, no complicated understanding of the semantics is needed to correctly make a prediction. To understand sentiment in a manner without (difficult) negations, a deeper semantic understanding depresses the predictions since only meanings of individual words seem to be sufficient. Therefore, the Embed-Align model was the worst performing model for such tasks. In Rios et al. (2018), the Embed-Align model was also outperformed by the skip-gram model on most of the SentEval tasks. The Embed-Align model did score almost perfectly on the SUBJ task, in which a sentence needs to be classified as either objective or subjective. For this task the whole sentence needs to be taken into consideration. The Embed-Align model showed results that were more similar to that of the skip-gram models or even better on the natural language inference tasks. This shows that the Embed-Align model is better at tasks in which the semantic relation between sentences is predicted, meaning that it better at capturing the semantic meaning of a word.

Code is available at `github.com/meltjh/ull-3`.

References

Arthur Brazinskas, Serhii Havrylov, and Ivan Titov. 2017. Embedding words as distributions with a bayesian skip-gram model. *CoRR*, abs/1711.11027.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Miguel Rios, Wilker Aziz, and Khalil Sima'an. 2018. Deep generative model for joint alignment and word representation. *arXiv preprint arXiv:1802.05883*.

Appendix A Vocabulary

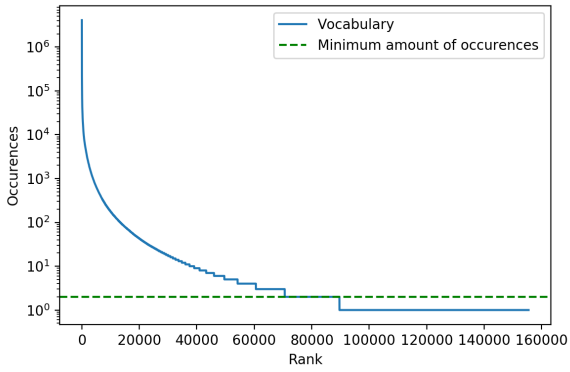

Figure 1: The English Europarl vocabulary contains of over 150,000 words of which more than half occurred less than 3 times and therefore are mapped to `UNK` (i.e. indicated by the green dashed line).