

Assignment 2: Distributional Word Embeddings

Filip Stefaniuk
filipste@student.matnat.uio.no

October 2, 2018

Introduction

This assignment had four different parts, each is described in detail in separate section of this report. To solve the problems I have implemented a number of scripts:

- `similar_wv.py` - finds nearest neighbours of the list of given words
- `eval_wv_intinsic.py` - evaluates models on intrinsic evaluation datasets
- `train_word2vec.py` - trains word embeddings model
- `train_model.py` - trains neural network model for classifying texts
- `eval_on_test.py` - evaluates model on test data

Every script has interface with support of `--help` command, that shortly explains script and other command line arguments. Most of the results are saved in json files. The code snippets to preprocess the data and process the results are in notebooks. Scripts, notebooks, results and latex source for this report are available in github repository.

1 Basic operations with word embeddings

For this part of the assignment, I have used script `play_with_model.py` which I have slightly modified, to allow me specify the range of indexes of the nearest words.

To produce input file, I copied the first sentence from Chen and Manning paper, and used `nlTK` to lemmatize, tag and remove stop words from that sentence. Results are saved in the file `data/words_auto.txt`. Unfortunately not everything was done correctly, namely one word was badly lemmatized and one had wrong POS tag. I manually corrected the mistakes and saved it as `data/words.txt`.

Finally I run the script using 2 different word models *Wikipedia* and *Gigaword*. I saved the results in `results/similar_words/`. Some of the words are presented in the tables below:

Top Words: classify_VERB				
Gigaword			Wikipedia	
11	redacted_ADJ	0.451433	reclassification_NOUN	0.502967
12	iv2_NOUN	0.446738	classified_ADJ	0.498774
13	unredacted_ADJ	0.441288	group_VERB	0.498190
14	nonpublic_ADJ	0.440909	sub-type_NOUN	0.489708
15	declassify_VERB	0.435394	label_VERB	0.484267

Figure 1: 11th to 15th nearest neighbours for word 'classify', with corresponding scores.

Top Words: current_ADJ				
	Gigaword		Wikipedia	
11	latest_ADJ	0.405347	2008-2011_NUM	0.443892
12	overall_ADJ	0.402181	2011-2013_NUM	0.439801
13	original_ADJ	0.401413	past_ADJ	0.437877
14	outgoing_ADJ	0.400801	change_VERB	0.433897
15	existing_ADJ	0.397309	2014-2017_NUM	0.427117

Figure 2: 11th to 15th nearest neighbours for word 'current', with corresponding scores.

It looks like *Wikipedia* model captures the meaning of 'classify' quite well with the similar words like 'group' or 'label'. Moreover, because the words presented above are not the closest neighbours it is also interesting to see that the other meaning of 'classify' was captured with the words such as 'nonpublic' and 'classified'. With word 'current' another phenomena occurs, while in *Gigaword* we can see words with quite similar meaning, in *Wikipedia* we observe the dates! This is not wrong, since when talking about *current* events, we often think about last days or years, but this example illustrates that with anomalies like this one it is difficult to train models, that will stay valid for a longer period of time.

2 Intrinsic evaluation of pre-trained word embeddings

I have downloaded both **SimLex999** and **Google Analogies** datasets, and I evaluated on them three different models of word embeddings. I have used builtin gensim methods `evaluate_word_pairs()` and `evaluate_word_analogies()`.

2.1 SimLex999

I found out that, the simplest way to split **SimLex999** into sections, was to simply split dataset into separate files. Because there were no clear separators between sections I have checked the original paper and discovered, that the dataset is structured in the way that the first 111 word pairs are adjectives next 666 are nouns, and the last 222 are verbs. I split the dataset accordingly.

	40 CoNLL-17	75 Oil and Gas	82 Common Crawl
adj	0.436081	0.405065	0.570734
noun	0.410926	0.263471	0.434926
verb	0.154409	0.186691	0.157278
total	0.361796	0.268299	0.398917

Figure 3: Pearson correlation on SimLex999 dataset.

	40 CoNLL-17	75 Oil and Gas	82 Common Crawl
adj	0.00000	41.441441	0.00000
noun	0.00000	12.462462	0.00000
verb	0.45045	31.981982	0.45045
total	0.10010	20.020020	0.10010

Figure 4: Ratio of OOV words (%).

From the evaluation results we can observe that comparing verb pairs is the most difficult. It is very interesting, since the authors of the **Simlex999** mentioned in their work, that when testing word pairs with children, they also had the most difficulty with the verbs. It is also worth mentioning that even though the second model (trained on Oil and Gas data) achieved better correlation on verbs than the other two, 30% of the words were not in the vocabulary.

2.2 Google Analogies

Gensim methods for evaluation are not very flexible, and with word analogies there was no simple method to extract the information about oov tokens, that was logged but not returned as a result from function. I present the results of the evaluation below, but I want to emphasise that in case of the second model 35% of words were not in vocabulary.

	40 CoNLL-17	75 Oil and Gas	82 Common Crawl
capital-common-countries	0.776680	0.177489	0.916996
capital-world	0.722370	0.084122	0.906883
currency	0.146040	0.000000	0.206304
city-in-state	0.462910	0.022505	0.697608
family	0.863636	0.000000	0.938735
gram1-adjective-to-adverb	0.218750	0.173810	0.357863
gram2-opposite	0.280788	0.158333	0.347291
gram6-nationality-adjective	0.762351	0.479290	0.893684
Total accuracy	0.574914	0.153178	0.740230

Figure 5: Accuracy scores, evaluation on Google Analogies dataset.

2.3 Analysis

On both test sets the clear winner is the model trained on *Common Crawl*. There is a simple explanation for that fact, this model was trained on the largest number of data which really matters in case of word embeddings.

3 Training a word embedding model on in-domain data

I have trained two models using `train_word2vec.py`, the models were trained with skipGram algorithm, window size of 2 and vector sizes of **300** (model1) and **100** (model2). Both models were trained for 2 iterations.

3.1 Top most frequent words

I have compared the most frequent words from models trained on *SignalMedia*, *CoNLL17*, *Oil and Gas* and *Common Crawl*. The most common words in models trained on large corporas are either interpunction signs or stop words. The explanation is that the large amount of data is more difficult to process and the quality of it drops. The models achieve good results because the amount of data compensates its quality. In report I include only 10 most common but file `results/top_words.csv` contains all 20.

	SignalMedia	CoNLL17	Oil and Gas	Common Crawl
0	say_VB	</s>	lrb	the
1	year_NN	,	rrb	,
2	make_VB	the	sediment	.
3	company_NN	.	fault	and
4	also_RB	of	datum	to
5	get_VB	and	basin	of
6	new_JJ	to	sample	a
7	time_NN	a	area	in
8	include_VB	in	study	is
9	take_VB	-	model	that

Figure 6: Most common words in models

3.2 Intrinsic Evaluation

I have evaluated both models in the same manner as the models in the previous section:

	SignalMedia1	SignalMedia2		SignalMedia1	SignalMedia2
adj	0.574247	0.496217	adj	0.000000	0.000000
noun	0.401974	0.377345	noun	0.600601	0.600601
verb	0.320678	0.191505	verb	0.900901	0.900901
total	0.390785	0.340145	total	0.600601	0.600601

Figure 7: Pearson correlation and percentage of OOV words on SimLex999 dataset.

Interestingly correlation on nouns were as good as in *Common Crawl* model, even though this model was trained on very little data. The total correlation is not really comparable, since 90% of verbs, which lowers the result were out of vocabulary. Comparing two models trained on in-domain data it can be seen that increasing vector size improves the results.

	SignalMedia1	SignalMedia2
capital-common-countries	-	-
capital-world	-	-
currency	0.000000	0.166667
city-in-state	-	-
family	0.620915	0.565359
gram1-adjective-to-adverb	-	-
gram2-opposite	-	-
gram6-nationality-adjective	-	-
Total accuracy	0.586420	0.543210

Figure 8: Accuracy scores, evaluation on Google Analogies dataset.

It is hard to get useful information from evaluating the models on analogies dataset, since almost all the classes contained words that were not in the vocabulary of this models. The only meaningful result was from evaluating on *family* section, on which model performs worse than pre-trained embeddings. The reason for that is that the domain on which the model was trained is much different from the categories in this dataset.

4 Document classification with word embeddings