Ana Mestre - MIIS

Word similarity with distributional semantics

Exercise 1. The classier needs to decide when two words are highly similar, that is, when they are synonyms or near synonyms. Your classifier will decide that two words are highly similar if their cosine similarity is larger than 0.5.

The following table shows an overview of the results obtained for this model:

Part-of-speech results			
	Precision	**Recall**	**Accuracy**
Overall	0.73958	0.41764	0.63529
Adjectives	0.93333	0.58333	0.76595
Nouns	0.69354	0.43	0.62189
Verbs	0.73684	0.30434	0.59782

The precision is quite high. This means that 74% of positive predictions were actually positive.

The recall is not so high. This means that 42% of the actual positives our model captured them as positive.

With these two values, the high precision and low recall we can see that our classifier doesn't have a lot of positives, but the ones that identifies as positives are more likely to be true positives.

The general accuracy is not so high. In our case, this means that 64% of the time our model is right. This percentage is quite near to the threshold, so nearly half of the time our model predicts different results from the ground truth.

In this case, we can check how many were right predicted (124 out of 340 pairs), so this means that only 36% of the time our model was right, which doesn't seem to be a really good general classifier. However, if we consider that a > 50% for precision, recall and accuracy is a high number, we can think that in general these values are moderately good, although it obviously depends on what you are training your model for. As an overview of the different numbers, this is a moderately good model although could be improved. An optimization would be to increase the threshold, although this depends on what we really want to achieve. If we want to increase the precision or recall and accuracy.

Some pattern observations from the data are:

- Our model tends to classify as similar, pairs of words that are total synonyms, such like:

W1	W2	Classified	PoS	Similarity	Ground truth
Huge	Immense	C	A	0.688	1
Hymn	Anthem	C	N	0.604	1
Investigate	Examine	C	V	0.609	1
Liquor	Booze	C	N	0.61	1

However, there are words which are synonyms but the classifier fails at rating their similarity, this is because of the different kind of contexts (and texts) where these words can be found:

W1	W2	Classified	PoS	Similarity	Ground truth
Water	h2o	X	N	0.377	1
Sinner	Evildoer	X	N	0.347	1
Animal	Creature	X	N	0.463	1

- Our model usually correctly classifies antonyms as different:

W1	W2	Classified	PoS	Similarity	Ground truth
unnecessary	necessary	C	A	0.428	0
weird	normal	C	A	0.277	0
wide	narrow	C	A	0.457	0

Although there are some exceptions:

W1	W2	Classified	PoS	Similarity	Ground truth
smart	Dumb	X	A	0.579	0

- Our model classifies as similar those words that share the same lexema. This is generally a correct behaviour since these words are usually near synonyms:

W1	W2	Classified	PoS	Similarity	Ground truth
winter	wintertime	C	N	0.697	1
alley	alleyway	C	N	0.740	1
aunt	auntie	C	N	0.687	1

- Our model fails at classifying topic-related words that aren't neither synonyms nor near synonyms, they are just related and shouldn't be classified as high similar:

W1	W2	Classified	PoS	Similarity	Ground truth
son	father	X	N	0.893	0
king	princess	X	N	0.516	0

Taking a deeper look into the Part-of-speech results, we can see that the tendency of the overall is also followed. The precision tends to be high, the accuracy is not so high and recall not that high either. However, it can be seen that the model obtains better results for adjective pairs (Precision 0.93, recall 0.58 and accuracy 0.77).

Here are some examples for Adjectives who usually make a good classification, compared with nouns or verbs that tend to make more mistakes.

W1	W2	Classified	PoS	Similarity	Ground truth
difficult	hard	**C**	**A**	0.602	1
difficult	simple	**C**	**A**	0.306	0
cow	cattle	**C**	**N**	0.632	1
cow	goat	**X**	**N**	0.636	0
create	destroy	**C**	**V**	0.377	0
create	make	**X**	**V**	0.444	1

Exercise 2. Similarity is a graded notion. In the second part of the assignment, you are going to work with graded similarity judgments. You will use correlation to evaluate the model.

Spearman Results		
	Correlation	**P-value**
Whole dataset	0.4419	5.1181e-49
PoS = A	0.5922	7.6142e-12
PoS = N	0.4520	7.4475e-35
PoS = V	0.3219	9.5990e-07

We are trying to check whether our assumed hypothesis is true or false. We can see a correlation of 0.44 which is a positive number and a quite moderate value in this case. So the correlation, that is the relationship between the predicted values and the real ones, is basically moderate. Neither weak nor high, falls in between.

For the state-of-the-art comparison, let's take a look at the following table:

State-of-the-art	**Spearman Correlation**	**Higher/Lower than ours**
Our skipgram-GoogleNews model	***0.44***	
Skipgram (Word2Vec) model trained on 1bn words of Wikipedia text	0.37	Lower
model trained on running monolingual text	0.56	Higher
Neural Machine Translation Model (En->Fr) trained on a relatively small bilingual corpus	0.52	Higher
A model that exploits curated knowledge-bases (WordNet, Framenet etc)	0.58	Higher
A model that uses rich paraphrase data	0.68	Higher
A hybrid model trained on features from various word embeddings and two lexical databases	0.76	Higher

As we can see in the table above, our model is below the average of the other models listed. I assume that this is because these models are newer, more advanced and have been optimized achieving higher correlation values. However, it is higher than the Skipgram (Word2Vec) model trained on 1bn words of Wikipedia since the data used for the training is different, one uses Wikipedia data and the other one uses Google News. The Google News model obtained a better correlation because of the diversity of the news used for training. Wikipedia has an article per topic while Google News may have different articles per news, which makes it easier to find synonyms and therefore, get a better trained model.

Now paying attention to the Part-of-speech values, we can observe that higher correlation is obtained for adjectives, meaning that adjectives are the best predicted kind of words in this models. Adjectives are followed by nouns and verbs, respectively. This is also what happened on the first exercise with binary classification (threshold 0.5). The previous model predicted the adjectives better than nouns and verbs.

If we look at individual cases, comparing our predictions to the SimLex-999 values (range [0-10]), we can observe the following statements:

- For synonyms, our model usually gets a high similarity prediction, which is related to the SimLex-999 value (values usually > 8)

W1	W2	PoS	Similarity	SimLex-999
smart	intelligent	A	0.649	9.2
stupid	dumb	A	0.817	9.58
hard	difficult	A	0.603	8.77
Happy	glad	A	0.741	9.17

Although, there are some exceptions:

W1	W2	PoS	Similarity	SimLex-999
father	parent	N	0.257	7.07
boundary	border	N	0.386	9.08

- Our model doesn't do really good with antonyms, the similarity prediction it gets is quite high. However, the SimLex-999 value tends to be weak. (< 3)

W1	W2	PoS	Similarity	SimLex-999
short	long	A	0.577	1.23
winter	summer	N	0.716	2.38

- There are some cases where our model gets high similarity because of how the words are related among themselves but this doesn't make them synonyms. Comparing these values with the SimLex-999 grade, we'll see that they're not quite similar.

W1	W2	PoS	Similarity	SimLex-999
employer	employee	N	0.655	3.65
decade	century	N	0.603	3.48
dinner	breakfast	N	0.701	3.33

- For near synonyms, our model gets high similarity, which is a good prediction. If we compare to the SimLex-999 value, we'll see that this number is pretty high as well.

W1	W2	PoS	Similarity	SimLex-999
priest	monk	N	0.633	6.28
sea	ocean	N	0.764	7.08
band	orchestra	N	0.503	7.08

Also, looking at the predictions themselves, we can conclude that total synonyms usually obtain a higher similarity prediction (as well as SimLex-999 value) although the difference between these numbers doesn't seem to be that as much meaningful as expected. For example:

W1	W2	PoS	Similarity	SimLex-999	Synonyms
attorney	lawyer	N	**0.820**	9.35	Total
sea	ocean	N	**0.764**	7.08	Near

Exercise 3. General conclusions.

- There is some pattern identification or analysis that should be done in order to understand the model better and how the individual results behave.
- The structure and size of the model is also important to take into account, this can be easily be seen comparing with a smaller dataset.
- Using a graded similarity judgement is a better option rather than the binary one, which can be too simple for some cases. The graded similarity helps to determine how close those words are to be synonyms.
- When using some kind of threshold for binary classification, it is important to consider what case is more critical. Whether we want low/high false negatives, low/high false positives, low/high true positives, low/high true negatives. The case we are working for is really decisive to take into consideration before starting.
- Part-of-the-speech results are different between them. That is another reason for studying the behaviour and patterns of the model. For example, our model usually gets better results with adjectives.
- Our model it’s moderately good. However some improvements should be done in order to try to handle better the high prediction between antonyms and semantic related words which aren’t synonyms at all.