Research Project Report
Modeling Morphology with Distributional Semantics

Zhiyin Tan
tan.zhiyin@ims.uni-stuttgart.de

1 Introduction

The word bag model is commonly used to form a distributed semantic space. The theoretical basis of this model is that the words adjacent to each other are also semantically related to some extent. There is also another kind of sound that replaces the window of the bag model with the path of syntactic relationship, that is, establishes a dependency-based construction of semantic space models. The research path of this study, replicate Lazaridou et al. 2013, differs in the use of syntax-based DSMs as an underlying model to train cDSMs.

2 Hypothesis

The hypothesis of this study is that the syntax-based distributed semantic model is different from the bag-of-word based.

3 Experimental setup

3.1 Material

The corpus comes from ukWaC, a total of 98 million tokens. Because the data is relatively large, this study extract data separately and finally puts them together (see the code set for details). Dependency parsing is done using the python package spacy. Paths of length is 1.

The parsed data format is that each token is accompanied by its head and the syntactic relationship between them. The token is treated as the target word, and the token's head acts as the context word in the bag-o-word, which can be called a syntax element.

And then calculated how often each pair of syntactic relations occurs. It specifically sets weights for different syntactic relationships. Among them, reference Pado at el. 2007, the weight of the subj (subject) is 5, the weight of the obj (object) is 4 (including the direct object and the indirect object), the weight of the agent (implementer) is 3, the adv (adverb) is 2, the mod (modifier) is 2, and the rest is 1. The total frequency of each pair of syntactic relationships is equal to the weight multiplied by the frequency.

There were three steps to clean the data. The first step was to remove the syntactic relationship with a total frequency below 3. The second step was based on the reference, The second step is based on the reference, target words in the form of lemma, leaving only part-of-speech are nouns, verbs, adjectives and adverbs. The third step was to generat a word types list according to the target words. Then checked if each of the syntactic elements exist in the list, deleted the syntactic relationship that does not exist. The list would also be used in chapter 3.

3.2 Distributional semantic space

Distributed semantic space is done with tookit DISSECT. I apply Positive Point-wise Mutual Information (PPMI) as the weighting scheme. Then dimensionality reduction uses Singular Value Decomposition (SVD) to set the number of reduced-space dimensions to 350.

3.3 Morphological data

Upload the word types list of target word which consist of the row of the semantic space, to obtained a list of stem/derived-form pairs from the CELEX English Lexical Database. The data set consists of affixes, stem/derived pairs matching its most common part-of-speech signature.

In order to clean the data, the word whose stem is the same as its own have been removed. This study did not do a crowdsourcing survey. The validation set is the intersection of the dataset of this study and the dataset of the reference (http://clic.cimec.unitn.it/composes), a total

of 805 pairs of words.

3.4 Implementation of composition methods

Dinu at el. (2013) has suggested that dissect could generate a series of cDSM. See Figure **??** for the currently available composition models, their definitions and parameters. This study selected lexfun as a method of composition.

Model	Composition function	Parameters
Add.	$w_1\vec{u} + w_2\vec{v}$	$w_1(=1), w_2(=1)$
Mult.	$\vec{u} \odot \vec{v}$	-
Dilation	$\|\|\vec{u}\|\|_2^2\vec{v} + (\lambda - 1)\langle\vec{u}, \vec{v}\rangle\vec{u}$	$\lambda(=2)$
Fulladd	$W_1\vec{u} + W_2\vec{v}$	$W_1, W_2 \in \mathbf{R}^{m\times m}$
Lexfunc	$A_u\vec{v}$	$A_u \in \mathbf{R}^{m\times m}$

Figure 1: Composition Models

4 Experiment

Based on the established distributed semantic space model, the training set is put into lexfunc for training. The place where the function originally belongs to the word is replaced by affix as a function. The result of the training is the correlation between each pair of stem/derived-form pairs. Calculate the average by taking the same affix as a group. This result represents the correlation coefficient of this affix as a function. The result shows on Table **??** column 4.

The result of column 3 is a crowdsourced survey from Lazaridou et al. 2013. The rating on the left is 7-points scale. On the right is the score based on the score interval of cosine similarity, calculated from the percentage, used to compare with the results of this study.

Column 5 is the result of building a distributed semantic space using bag-of-word and performing the same model operation. As can be seen from the table, the results of the two models generated in this study are very close. Different places are represented by affix -ful (syntax ¡ word), -less (syntax ¡ word), -ly (syntax ¡ word), in- (syntax ¡ word), -un (syntax ¡ word), -ness (syntax ¿ word). Basically, in the syntax-based semantic space, these affix-related words are less relevant than word-based, except affix -ness.

5 Conclusions and Disscussion

In summary, at present, the two different semantic spaces are not much different in this experiment. But it is worth exploring that there is a very interesting difference between the results of this study and the crowdsourcing results of the reference papers.

Specifically in affix in-, un- and -ness. In the crowdsourcing score, the scores of the first two affix were quite low. In the present study, the two affix did not specifically show their distinctiveness. Conversely, affix -ness is generally considered to have a high correlation coefficient, and the score is the lowest in this study. This inevitably makes me think about whether people will be biased by subjectivity when making judgments. Furthermore, when we are using the results of the crowdsourcing survey, should we be cautious about whether it can be used as a gold standard?

Affix	Stem/Der Pos	Avg. SDR	cos. Sim(syntax)	cos. Sim(word)
-able	verb/adj	5.96 / 0.70	0.34	0.36
-al	noun/adj	5.88 / 0.68	0.41	0.41
-er	verb/noun	5.51 / 0.57	0.49	0.52
-ful	noun/adj	6.11 / 0.75	0.32	0.37
-ic	noun/adj	5.99 / 0.71	0.37	0.38
-ion	verb/noun	6.22 / 0.78	0.52	0.54
-ist	noun/noun	6.16 / 0.76	0.37	0.39
-ity	adj/noun	6.19 / 0.77	0.30	0.31
-ize	noun/verb	5.96 / 0.70	0.35	0.36
-less	noun/adj	3.72 / 0.06	0.29	0.34
-ly	adj/adv	6.33 / 0.81	0.33	0.39
-ment	verb/noun	6.06 / 0.73	0.45	0.48
-ness	adj/noun	6.29 / 0.80	0.27	0.22
-ous	noun/adj	5.94 / 0.70	0.31	0.33
-y	noun/adj	5.25 / 0.5	0.49	0.51
in-	adj/adj	3.39 / -0.03	0.50	0.55
re-	verb/verb	5.28 / 0.51	0.41	0.46
un-	adj/adj	3.23 / -0.08	0.49	0.57

Table 1: Results of relateness