An Empirical Study on Compositionality in Compound Nouns

Abstract

A multiword is compositional if its meaning can be expressed in terms of the meaning of its constituents. In this paper, we collect and analyse the compositionality judgments for a range of compound nouns using Mechanical Turk. Unlike existing compositionality datasets, our dataset has judgments on the contribution of constituent words as well as judgments for the phrase as a whole. We use this dataset to study the relation between the judgments at constituent level to that for the whole phrase. We then evaluate two different types of distributional models for compositionality detection – constituent based models and ~~compositionality~~ composition function based models. Both the models show competitive performance though the ~~compositionality~~ composition function based models perform slightly better. In both types, additive models perform better than their multiplicative counterparts.

1 Evaluation

We evaluated all the models on the dataset developed in section sec:setup. Since our dataset has constituent level contributions along with phrase compositionality judgments, we evaluated the constituent based models against both the literality scores of the constituents (section sec:literalConst) and the phrase level judgments (section sec:literalCompound).

~~We evaluate constituent based models by correlating the constituent literality scores $s1$ and $s2$ with the constituent level human annotations and similarly for the phrase compositionality we calculate correlations of $s3$ determined using various functions with the phrase level human annotations~~ The composition function models are evaluated only on phrase level scores following [**?**, **?**, **?**]: higher correlation scores indicate better compositionality predictions.

	first constituent	second constituent
s1	0.616	–
s2	–	0.707

Table 1: Constituent level correlations

Model	ρ	R^2
Constituent Based Models		
ADD	0.686	0.613
MULT	0.670	0.428
COMB	0.682	0.615
WORD1	0.669	0.548
WORD2	0.515	0.410
Compositionality Function Based Models		
$a\mathbf{v1} + b\mathbf{v2}$	0.714	0.620
v1v2	0.650	0.501
RAND	0.002	0.000

Table 2: Phrase level Correlations of Compositionality scores

Constitunet based models evaluation

Spearman's ρ correlations of s1 and s2 with the human constituent level judgments are shown in table ~~4.~~ tab:IndividualWords. We observed that the predictions for the second constituent are more accurate than those for the first constituent. Perhaps these constitute an easier set of nouns for modelling but we need to investigate this further.

For the phrase ~~level scores~~ compositionality evaluation we did a 3-fold cross validation. The parameters of the functions f (section sec:literalCompound) are predicted by least square linear regression over the training samples and optimum values are selected. The average Spearman correlation scores ~~with the~~ of phrase compositionality scores with human judgements on the testing samples are displayed in table tab:mainResults. The goodness of fit R^2 values when trained over the whole dataset are also displayed.

~~From the results of the constituent word level literality correlations in table tab:IndividualWords, we observed that the predictions for the second constituent are more accurate than those for the first constituent. Perhaps these constitute an easier set of nouns for modelling but we need to investigate this further. For the phrase compositionality (see table tab:mainResults), the first constituent (model WORD1 i.e. $sim(v1, v3)$) was found to be a better predictor than the second (WORD2) following the behaviour of the mechanical turkers as in table tab:relGold.~~

~~Among the constituent based models (table tab:mainResults), it~~ It is clear that models ADD and COMB which use both the constituents are better predictors of phrase compositionality compared to the single word based predictors WORD1 and WORD2. Both ADD and COMB are competitive in terms of both the correlations (accuracy) and goodness of fit values. The model MULT shows good correlation but the goodness of fit is lower. First constituent (model WORD1 i.e. $sim(\mathbf{v1}, \mathbf{v3})$) was found to be a better predictor of phrase compositionality than the second (WORD2) following the behaviour of the mechanical turkers as in table tab:relGold.

Composition function based models evaluation

These models are evaluated for phrase composotionality scores. As with the constituent based models, for estimating the model parameters a and b of the compositionality function based models, we did a 3-fold cross validation~~for experimenting with parameters on the training data~~. The best results on the training ~~sample~~ samples are found at a=0.60 and b=0.40. Average Spearman correlation scores of both addition and multiplication models over the testing samples are displayed in table tab:mainResults. The goodness of fit R^2 values when trained over the whole dataset are also displayed.

~~For both constituent and compositionality function based models (table tab:mainResults), vector~~ Vector addition has a clear upper hand over multiplication in terms of both accuracy and goodness of fit for phrase compositionality prediction.

~~Comparing~~

Winner

For phrase compositionality prediction, both, constituent based and compositionality function based models (table tab:mainResults) ~~, both of them~~ are found to be competitive, though compositionality function based models ~~are~~ perfrom slightly better. The reason could be because while constituent based models use contextual information of each constituent independently, composition funciton models make use of collective evidence from the context of both the constituents simultaneously. In future, we would like to explore the importance of contexts salient to both the constituents in contrast with contexts salient to each word.

All the results when compared with random baseline (RAND in table tab:mainResults), which assigns a random compositionality score to a compound, are highly significant.