Shared task system description: Frustratingly hard compositionality prediction

Anders Johannsen, Hector Martinez Alonso, Christian Rishøj and Anders Søgaard
Center for Language Technology
University of Copenhagen
{ajohannsen|alonso|crjensen|soegaard}@hum.ku.dk

Abstract

We considered a wide range of features for the DiSCo 2011 shared task about compositionality prediction for word pairs, including COALS-based endocentricity scores, compositionality scores based on distributional clusters, statistics about wordnet-induced paraphrases, hyphenation, and the likelihood of long translation equivalents in other languages. Many of the features we considered correlated significantly with human compositionality scores, but in support vector regression experiments we obtained the best results using only COALS-based endocentricity scores. Our system was nevertheless the best performing system in the shared task, and average error reductions over a simple baseline in cross-validation were 13.7% for English and 50.1% for German.

1 Introduction

The challenge in the DiSCo 2011 shared task is to estimate and predict the semantic compositionality of word pairs. Specifically, the data set consists of adjective-noun, subject-verb and object-verb pairs in English and German. The organizers also provided the Wacky corpora for English and German with lowercased lemmas.[1] In addition, we also experimented with wordnets and using Europarl corpora for the two languages (Koehn, 2005), but none of the features based on these resources were used in the final submission.

Semantic compositionality is an ambiguous term in the linguistics litterature. It may refer to the position that the meaning of sentences is built from the meaning of its parts through very general principles of application, as for example in type-logical grammars. It may also just refer to a typically not very well defined measure of semantic transparency of expressions or syntactic constructions, best illustrated by examples:

(1) pull the plug

(2) educate people

The verb-object word pair in example (1) is in the training data rated as much less compositional than example (2). The intuition is that the meaning of the whole is less related to the meaning of the parts. The compositionality relation is not defined more precisely, however, and this may in part explain why compositionality prediction seems frustratingly hard.

2 Features

Many of our features were evaluated with different amounts of *slop*. The slop parameter permits non-exact matches without resorting to language-specific shallow patterns. The words in the compounds are allowed to move around in the sentence one position at a time. The value of the parameter is the maximum number of steps. Set to zero, it is equivalent to an exact match. Below are a couple of example configurations. Note that in order for w_1 and w_2 to swap positions, we must have slop > 1 since slop=1 would place them on top of each other.

$x\ x\ w_1\ w_2\ x\ x$ (slop=0)
$x\ x\ w_1\ x\ w_2\ x$ (slop=1)
$x\ x\ w_1\ x\ x\ w_2$ (slop=2)
$x\ x\ w_2\ w_1\ x\ x$ (slop=2)

[1] http://wacky.sslmit.unibo.it/

Proceedings of the Workshop on Distributional Semantics and Compositionality (DiSCo'2011), pages 29–32,

2.1 LEFT-ENDOC, RIGHT-ENDOC and DISTR-DIFF

These features measure the endocentricity of a word pair w_1 w_2. The distribution of w_1 is likely to be similar to the distribution of "w_1 w_2" if w_1 is the syntactic head of "w_1 w_2". The same is to be expected for w_2, when w_2 is the head.

Syntactic endocentricity is related to compositionality, but the implication is one-way only. A highly compositional compound is endocentric, but an endocentric compound need not be highly compositional. For example, the distribution of "olive oil", which is endocentric and highly compositional, is very similar to the distribution of "oil", the head word. On the other hand, "golden age" which is ranked as highly *non-compositional* in the training data, is certainly endocentric. The distribution of "golden age" is not very different from that of "age".

We used COALS (Rohde et al., 2009) to calculate word distributions. The COALS algorithm builds a word-to-word semantic space from a corpus. We used the implementation by Jurgens and Stevens (2010), generating the semantic space from the Wacky corpora for English and German with duplicate sentences removed and low-frequency words substituted by dummy symbols. The word pairs have been fed to COALS as compounds that have to be treated as single tokens, and the semantic space has been generated and reduced using singular value decompositon. The vectors for w_1, w_2 and "w_1 w_2" are calculated, and we compute the cosine distance between the semantic space vectors for the word pair and its parts, and between the parts themselves, namely for "w_1 w_2" and w_1, for "w_1 w_2" and w_2, and for w_1 and w_2, say for "olive oil" and "olive", for "olive oil" and "oil", and for "olive" and "oil". LEFT-ENDOC is the cosine distance between the left word and the compound. RIGHT-ENDOC is the cosine distance between the right word and the compound. Finally, DISTR-DIFF is the cosine distance between the two words, w_1 and w_2.

2.2 BR-COMP

To accommodate for the weaknesses of syntactic endocentricity features, we also tried introducing compositionality scores based on hierarchical distributional clusters that would model semantic compositionality more directly. The scores are referred to below as BR-COMP (compositionality scores based on Brown clusters), and the intuition behind these scores is that a word pair "w_1 w_2", e.g. "hot dog", is non-compositional if w_1 and w_2 have high collocational strength, but if w_1 is replaced with a different word w_1' with similar distribution, e.g. "warm", then "w_1' w_2" is less collocational. Similarly, if w_2 is replaced with a different word w_2' with similar distribution, e.g. "terrier", then "w_1 w_2'" is also much less collocational than "w_1 w_2".

We first induce a hierarchical clustering of the words in the Wacky corpora $cl : W \rightarrow 2^W$ with W the set of words in our corpora, using publicly available software.[2] Let the collocational strength of the two words w_1 and w_2 be $G^2(w_1, w_2)$. We then compute the average collocational strength of distributional clusters, BR-CS (collocational strength of Brown clusters):

$$\text{BR-CS}(w_1, w_2) = \frac{\Sigma^N_{x \in cl(w_1), x' \in cl(w_2)} G^2(x, x')}{N}$$

with $N = |cl(w_1)| \times |cl(w_2)|$. We now let $\text{BR-COMP}(w_1, w_2) = \frac{\text{BR-CS}(w_1, w_2)}{G^2(w_1, w_2)}$.

The Brown clusters were built with $C = 1000$ and a cut-off frequency of 1000. With these settings the number of word types per cluster is quite high, which of course has a detrimental effect on the semantic coherence of the cluster. To counter this we choose to restrict $cl(w)$ and $cl(w')$ to include only the 50 most frequently occurring terms.

2.3 PARAPHR

These features have to do with alternative phrasings using synonyms from Princeton WordNet [3] and GermaNet[4]. One word in the compound is held constant while the other is replaced with its synonyms. The intuition is again that non-compositional compounds are much more frequent than any compound that results from replacing one of the constituent words with one of its synonyms. For "hot dog" we thus generate "hot terrier" and "warm dog", but not "warm terrier". Specifically, $\text{PARAPHR}_{\geq 100}$ means

[2] http://www.cs.berkeley.edu/~pliang/software/

[3] http://wordnet.princeton.edu/

[4] GermaNet Copyright © 1996, 2008 by University of Tübingen.

that at least one of the alternative compounds has a document count of more than 100 in the corpus. PARAPHR$_{av}$ is the average count for all paraphrases, PARAPHR$_{sum}$ is the sum of these counts, and PARAPHR$_{rel}$ is the average count for all paraphrases over the count of the word pair in question.

2.4 HYPH

The HYPH features were inspired by Bergsma et al. (2010). It was only used for English. Specifically, we used the relative frequency of hyphenated forms as features. For adjective-noun pairs we counted the number of hyphenated occurrences, e.g. "front-page", and divided that number by the number of non-hyphenated occurrences, e.g. "front page". For subject-verb and object-verb pairs, we add *-ing* to the verb, e.g. "information-collecting", and divided the number of such forms with non-hyphenated equivalents, e.g. "information collecting".

2.5 TRANS-LEN

The intuition behind our bilingual features is that non-compositional words typically translate into a single word or must be paraphrased using multiple words (circumlocution or periphrasis). TRANS-LEN is the probability that the phrase's translation, possibly with intervening articles and markers, is longer than l_{min} and shorter than l_{max}, i.e.:

$$\text{TRANS-LEN}(w_1, w_2, l_{min}, l_{max}) =$$

$$\frac{\Sigma_{\tau \in trans(w_1\ w_2), l_1 \leq |\tau| \leq l_2} P(\sigma | w_1\ w_2)}{\Sigma_{\tau \in trans(w_1\ w_2)} P(\sigma | w_1\ w_2)}$$

We use English and German Europarl (Koehn, 2005) to train our translation models. In particular, we use the phrase tables of the Moses PB-SMT system[5] trained on a lemmatized version of the WMT11 parallel corpora for English and German. Below TRANS-LEN-n will be the probability of the translation of a word pair being n or more words. We also experimented with average translation length as a feature, but this did not correlate well with semantic compositionality.

[5]http://statmt.org

feat	ρ English	German
rel-type = ADJ_NN	0.0750	*0.1711
rel-type = V_SUBJ	0.0151	**0.2883
rel-type = V_OBJ	0.0880	0.0825
LEFT-ENDOC	**0.3257	*0.1637
RIGHT-ENDOC	**0.3896	0.1379
DISTR-DIFF	*0.1885	0.1128
HYPH (5)	0.1367	-
HYPH (5) reversed	*0.1829	-
G^2	0.1155	0.0535
BR-CS	*0.1592	0.0242
BR-COMP	0.0292	0.0024
Count (5)	0.0795	*0.1523
PARAPHR$_{\geq \lvert w_1\ w-2 \rvert}$	0.1123	0.1242
PARAPHR$_{rel}$ (5)	0.0906	0.0013
PARAPHR$_{av}$ (1)	0.1080	0.0743
PARAPHR$_{av}$ (5)	0.1313	0.0707
PARAPHR$_{sum}$ (1)	0.0496	0.0225
PARAPHR$_{\geq 100}$ (1)	**0.2434	0.0050
PARAPHR$_{\geq 100}$ (5)	**0.2277	0.0198
TRANS-LEN-1	0.0797	0.0509
TRANS-LEN-2	0.1109	0.0158
TRANS-LEN-3	0.0935	0.0489
TRANS-LEN-5	0.0240	0.0632

Figure 1: Correlations. Coefficients marked with * are significant ($p < 0.05$), and coefficients marked with ** are highly significant ($p < 0.01$). We omit features with different slop values if they perform significantly worse than similar features.

3 Correlations

We have introduced five different kinds of features, four of which are supposed to model semantic compositionality directly. For feature selection, we therefore compute the correlation of features with compositionality scores and select features that correlate significantly with compositionality. The features are then used for regression experiments.

4 Regression experiments

For our regression experiments, we use support vector regression with a high (7) degree kernel. Otherwise we use default parameters of publicly available software.[6] In our experiments, however, we were not able to produce substantially better results than what can be obtained using only the features LEFT-ENDOC and RIGHT-ENDOC. In fact, for German using only LEFT-ENDOC gave slightly better results than using both. These features are also those that correlate best with human compositionality scores according to Figure 1. Consequently, we only use

[6]http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

these features in our official runs. Our evaluations below are cross-validation results on training and development data using leave-one-out. We compare using only LEFT-ENDOC and RIGHT-ENDOC (for English) with using all significant features that seem relatively independent. For English, we used LEFT-ENDOC, RIGHT-ENDOC, DISTR-DIFF, HYPH (5) reversed, BR-CS, PARAPHR$_{\geq}$100 (1). For German, we used rel-type = ADJ_NN, rel-type=V_SUBJ and RIGHT-ENDOC. We only optimized on numeric scores. The submitted coarse-grained scores were obtained using average +/- average deviation.[7]

	English		German	
	dev	test	dev	test
BL	18.395		47.123	
all sign. indep.	19.22		**23.02**	
L-END+R-END	**15.89**	16.19	23.51	24.03
err.red (L+R)	0.137		0.501	

5 Discussion

Our experiments have shown that the DiSCo 2011 shared task about compositionality prediction was a tough challenge. This may be because of the fine-grained compositionality metric or because of inconsistencies in annotation, but note also that the syntactically oriented features seem to perform a lot better than those trying to single out semantic compositionality from syntactic endocentricity and collocational strength. For example, LEFT-ENDOC, RIGHT-ENDOC and BR-CS correlate with compositionality scores, whereas BR-COMP does not, although it is supposed to model compositionality more directly. Could it perhaps be that annotations reflect syntactic endocentricity or distributional similarity to a high degree, rather than what is typically thought of as semantic compositionality?

Consider a couple of examples of adjective-noun pairs in English in Figure 2 for illustration. These examples are taken from the training data, but we have added our subjective judgments about semantic and syntactic markedness and collocational strength (peaking at G^2 scores). It seems that semantic markedness is less important for scores than syntactic markedness and collocational strength. In particular, the combination of syntactic markedness and collocational strength makes annotators rank word pairs such as *happy birthday* and *open source* as non-compositional, although they seem to be fully compositional from a semantic perspective. This may explain why our COALS-features are so predictive of human compositionality scores, and why G^2 correlates better with these scores than BR-COMP.

[7]These thresholds were poorly chosen, by the way. Had we chosen less balanced cut-offs, say 0 and 72, our improved accuracy on coarse-grained scores (59.4) would have been comparable to and slightly better than the best submitted coarse-grained scores (58.5).

	sem	**syn**	**coll**	**score**
floppy disk			✓	61
free kick	✓			77
happy birthday		✓	✓	47
large scale		✓	✓	55
old school	✓	✓	✓	37
open source		✓	✓	49
real life		✓		69
small group				91

Figure 2: Subjective judgments about semantic and syntactic markedness and collocational strength.

6 Conclusions

In our experiments for the DiSCo 2011 shared task we have considered a wide range of features and showed that some of them correlate significantly and sometimes highly significantly with human compositionality scores. In our regression experiments, however, our best results were obtained with only one or two COALS-based endocentricity features. We report error reductions of 13.7% for English and 50.1% for German.

References

Shane Bergsma, Aditya Bhargava, Hua He, and Grzegorz Kondrak. 2010. Predicting the semantic compositionality of prefix verbs. In *EMNLP*.

David Jurgens and Keith Stevens. 2010. The S-Space package: an open source package for word space models. In *ACL*.

Philipp Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *MT-Summit*.

Douglas Rohde, Laura Gonnerman, and David Plaut. 2009. An improved model of semantic similarity based on lexical co-occurrence. In *Cognitive Science*.