Composition of Compound Nouns Using Distributional Semantics

Kyra Yee
Pomona College
333 N College Way
Claremont, CA 91711, USA
kny02014@mymail.pomona.edu

Jugal Kalita
University of Colorado, Colorado Springs
1420 Austin Bluffs Pkwy
Colorado Springs, CO 80918, USA
jkalita@uccs.edu

Abstract

The use of distributional semantics to represent the meaning of a single word has proven to be very effective, but there still is difficulty representing the meaning of larger constituents, such as a noun phrase. In general, it is unclear how to find a representation of phrases that preserves syntactic distinctions and the relationship between a compound's constituents. This paper is an attempt to find the best representation of nominal compounds in Spanish and English, and evaluates the performance of different compositional models by using correlations with human similarity judgments and by using compositional representations as input into an SVM classifying the semantic relation between nouns within a compound. This paper also evaluates the utility of different function's compositional representations, which give our model a slight advantage in accuracy over other state-of-the-art semantic relation classifiers.

Keywords compositional distributional semantics, nominal compounds, nominal compounds in Spanish

1 Introduction

The use of distributional semantics has become increasingly popular due to its effectiveness in a range of NLP tasks. The vector-based representation is computed by looking at the context of every instance of a specific word within a large corpus, which is based on the idea that the meaning of a word is determined by its associations with other words (Erk, 2012). Despite the success of vector-based representation in a wide variety on contexts, this method still has difficulty handling larger phrase structures and function words, as opposed to just isolated content words (Mitchell and Lapata, 2008). Vectors for larger phrases cannot be reliably used due to the sparseness of data (Erk, 2012).

Ways of representing compositional models for constituents larger than a single word that preserve the lexical and syntactic function of a word in a phrase and best represent the relation between the constituents of a phrase is desired in creating a more general and powerful framework for natural language semantics. (Mitchell and Lapata, 2008; Mitchell and Lapata, 2010), and (Guevara, 2010) have compared and empirically tested the effectiveness of different mathematical compositions in representing adjective-noun, verb-object, and noun-noun compounds, but there has been little research into representing nominal compounds that are longer than two words, and the vast majority of research has been in English, without cross-linguistic inquiries (Mitchell and Lapata, 2010). This paper will investigate the effectiveness of a variety of different compositional functions using two metrics: correlation of the model's cosine similarity predictions with human similarity judgments for two, three and four word Spanish and English noun compounds, and by using the composition of two vectors as input into an SVM used to classify the relations between constituent nouns for two-word English noun compounds. For the human correlation task, this paper builds on (Mitchell and Lapata, 2010) by analyzing compounds longer than two words, which is a previously unexplored topic, and by analyzing the composition of Spanish compound nouns. As far as we know, compositional models have never been applied to Spanish word vectors be-

D S Sharma, R Sangal and A K Singh. Proc. of the 13th Intl. Conference on Natural Language Processing, pages 20–29, Varanasi, India. December 2016.

fore. Previous work have utilized word embeddings as input for relation classification, but we use the composed vectors as input as well, which also has never before been tested. This paper is also the first to use noun-relation classification accuracy as a metric for the utility of compositional functions, which gives a new level of insight into the

2 Compounding in English and Spanish

What constitutes a nominal compound is contested among linguists and computational linguists(Moyna, 2011; Finin, 1980). For our purposes, we will use the definition given by Finin (Finin, 1980):

> *A nominal compound is the concatenation of two or more nominal concepts which functions as a third nominal concept.*

The structure N N in English is productive, recursive, and compositional (Bauke, 2014). In Spanish, N N compounds are rarely productive, rarely contain more than two elements and are highly stylistic (Bauke, 2014; Moyna, 2011). The process is that of lexical word-formation, as opposed to English,which has syntactic word-formation for N N compounds (Bauke, 2014). In Spanish, the creation of N N compounds more closely resembles the invention of a new morpheme, which is reflected by the fact that only 2% of N N constructions are written as two words or a hyphenated word without one-word alternates (Moyna, 2011). Because of the limitations of N N constructions in Spanish, many consider the Spanish equivalent to the English N N structure to be the N P N structure, with a semantically empty preposition. This structure, similar to the English N N structure, is productive, recursive, and compositional (Bauke, 2014). We do not consider the more theoretical qualifications for compounds nouns proposed by (Moyna, 2011) for a more linguistically rigorous definition for Spanish compound nouns.

Restricting our attention to compounds only consisting of two nouns in English, analyzing the meaning of nominal compounds computationally has proven to be a difficult task because the listener must discern the relationship between the two words, which must be inferred contextually without any syntactic clues (Finin, 1980). Consider the cases of "meeting room", "salt water" and "aircraft engine". "Room" defines the location for "meeting", "engine" is a part of the "aircraft", and "salt" is dissolved in "water" (Finin, 1980). This problem of determining relations between the constituent nouns becomes even more difficult for longer phrases, because we now must determine the parse of the compound using contextual clues. In the phrase "computer science department", "computer science" modifies "department", instead of having "computer" modify "science department". These factors pose challenges to vector-based representations of longer compound noun phrases.

In Spanish N P N constructions, despite the presence of a preposition or potentially determiners, is it still difficult to discern the relation between the constituent nouns. Spanish definite determiners are used in a much wider context than their English counterparts, so they do not provide much useful insight into the relation between the two nouns. In the majority of cases, the preposition is "de", which is semantically empty in this construction (Bauke, 2014), and is used to represent a multitude of relations, as seen from Table I (taken from (Valle, 2008)).

English	Spanish	Meaning Implied
leather shoes	zapatos de piel	shoes made of leather
sports shoes	zapatos de deporte	shoes used to play sports with
winter shoes	zapatos de invierno	shoes to be worn in winter time
high-heel shoes	zapatos de tacón	shoes with high heels
display shoes	zapatos de muestra	shoes on display
Gucci shoes	zapatos de Gucci	shoes designed by Gucci

Table 1: Spanish Semantic Relations.

Thus the Spanish N P N construction poses similar challenges to the English N N construction. Our goal is to analyze compound nouns in English (which take on the form of N N) and semantically equivalent structures in Spanish, which take on the form N P N (Girju, 2009).

3 Previous Work

3.1 Word Embeddings

A variety of methods for generating word embeddings have been proposed, most famously the GloVe, word2vec, CW, and HPCA embeddings. The word2vec model, proposed by (Mikolov and Dean, 2013), is a continuous skip-gram model, built using neural neworks, which is able to capture precise syntactic and semantic word relationships to generate a vector representation of a word. The GloVe model (Pennington et al., 2014) is a global bilinear regression model which combines the advantages of global matrix factorization and local context window methods. It utilizes statistical information by training on "non-zero elements in a word-word cooccurrence matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus". The CW model, proposed by (Collobert et al., 2011), implements a multilayer neural network, where the first layer extracts features for each word and the second layer extracts features from a window of words. The model is refined using a supervised training step utilizing data from part-of-speech tagging, chunking, named entity recognition and semantic role labeling (Collobert et al., 2011; Dima and Hinrichs, 2015). The HPCA model (Lebret and Collobert, 2013) is generated by applying Hellinger PCA to a word co-occurance matrix, which has the advantage of being much faster than training a neural net.

3.2 Compositional Models

Very little work has been done in distributional semantics for Spanish. Some studies have been done on the effectiveness of vector-based representations on Spanish (Etcheverry and Wonsever, 2016; Al-Rfou et al., 2013), but none have considered compositional models. Many studies have been done in English studying compositional models, (Mitchell and Lapata, 2010; Mitchell and Lapata, 2008; Reddy et al., 2011; Im Walde et al., 2013; Baroni and Zamparelli, 2010; Guevara, 2010; Socher et al., 2012; Polajnar and Clark, 2014) but none have considered three or four word compound nouns.

There have been many functions suggested for how to compose two vectors. The general class of models representing the vector composition is defined by:

$$\mathbf{p} = f(\mathbf{u}, \mathbf{v}, R, K) \tag{1}$$

where **u** and **v** are the constituent vectors, R represents their syntactic relation, and K represents any additional information required to interpret the semantics of **p** (Mitchell and Lapata, 2010). Since we are only considering the composition of compound nominals, we can hold R fixed. We can also ignore K for simplicity, attempting to glean as accurate of a meaning as possible without further pragmatic context (Mitchell and Lapata, 2010). From these assumptions, we arrive at several more common potential functions: additive, multiplicative, and tensor product, respectively (Mitchell and Lapata, 2010).

$$p_i = u_i + v_i \tag{2}$$

$$p_i = u_i \cdot v_i \tag{3}$$

$$p_{ij} = u_i \otimes v_j \tag{4}$$

The tensor product has interested some researchers since it does a better job of encoding syntactic information (the tensor product is not commutative, so it is seen as a representation that can distinguish "blood donor" from "donor blood"). However, the tensor product becomes very computationally expensive, as the number of dimensions grows exponentially as more constituents are composed (Mitchell and Lapata, 2010; Polajnar and Clark, 2014). The effectiveness of each of these equations, especially between the additive and multiplicative model, is still contested (Baroni and Zamparelli, 2010). Another well-known function is the weighted additive function, which is regarded as being better at representing the syntactic relation between its constituents:

$$p_i = \alpha u_i + \beta v_i. \tag{5}$$

With regard to nominal compounds, one study showed that the influence of the modifier noun has a much greater influence on the overall meaning of the compound than the head noun in German, with respect to both human ratings and vector-space models (Im Walde et al., 2013). In contrast, another study determined that the semantic contribution of the modifier and head to a compound noun are approximately equal in English (Kim and Baldwin, 2005).That being said, it could be the case that the

average contribution of the modifier and head varies between languages, so determining a weighting for Spanish and English could yield different results to obtain the optimal weighted additive model. An extreme form of this formula would be to only use the vector from either the head or modifying noun:

$$p_i = u_i \tag{6}$$

$$p_i = v_i. \tag{7}$$

It is also possible to combine the weighted additive and multiplicative model:

$$p_i = \alpha u_i + \beta v_i + \gamma u_i v_i. \tag{8}$$

One major disadvantage to the multiplicative model is that the presence of a zero in either two component vectors will lead to a zero in the resulting vector, essentially meaning information from the noun-zero entries multiplied by zero was thrown away; combining these two models could help alleviate that effect (Mitchell and Lapata, 2010).

Other models for composition include utilizing a partial least squares regression (Guevara, 2010) or using a recursive neural tensor network (Socher et al., 2012). This paper will compare different models for the composition of two, three, and four word compounds in Spanish and English.

3.3 Automatic Compound Noun Interpretation

A variety of taxonomies have been proposed for the classification of compound noun relations, some of which consist of a relatively small number of semantic relations, while others propose an unbounded number (Tratz and Hovy, 2010). The taxonomy created by (Tratz and Hovy, 2010) has been widely used because of its comparatively high level of inter-annotator agreement for its relations and the large size of the data set. (Kim and Baldwin, 2005) use wordnet similarity to classify a set of 2169 compounds into 20 semantic categories, achieving 53% accuracy. (Girju, 2007) uses cross-linguistic data and an SVM model to achieve an accuracy of 77.9% on an unseen test set. (Tratz and Hovy, 2010) use a dataset of 17509 compounds and a maximum entropy classifier to achieve 79.3% for cross-validation and 51% accuracy on an unseen test set using a set of 43 semantic relations, using wordnet, surface level, thesaurus based, and N-gram features. (Verhoeven et al., 2012) uses word embeddings to classify Dutch and Afrikaans compound nouns, achieving 47.8 % and 51.1%, respectively. (Dima and Hinrichs, 2015) use a neural net on the concatenation of CW-50, FloVe-300, HPCA-200, and word2vec embeddings on the tratz dataset to achieve 77.7% accuracy on a ten-fold cross-validation and 77.12% accuracy on an unseen test set. Although (Dima and Hinrichs, 2015) and (Verhoeven et al., 2012) use word embeddings, none of the previously proposed models used the composition of word embeddings as input for their model.

4 Experimental Setup

4.1 Overview of the Procedure

For the human similarity judgment correlation task, we first created our own dataset for two, three, and four word compounds in Spanish and English, since longer compounds and Spanish compounds have been previously unexplored. We then generated embeddings for the constituent nouns from the noun compounds using word2vec and the BNC and the Spanish Wikipedia corpus. Next, we applied the compositional functions to the constituent nouns to create a representation for the compound. To apply functions with parameters in them, we used grid search to optimize the parameters, taking the best parameters as the ones with highest correlations to the human judgments. We then took the cosine similarity of noun compound pairs and correlated this with human judgments to determine how accurately each function represents the compound noun.

For the noun classification task, we experimented with a variety of embedding types to see which had the best results. We used the Tratz dataset and used the concatenation of the constituent embeddings and the composition embedding as input for the classifier to train an SVM. We took the performance of the difference classifier using different composition functions as evidence for the accuracy of a composition function to represent noun compounds.

4.2 Materials and Tools

We will evaluate the performance of each composition function in two ways: by analyzing its correlation with human similarity judgments,

and also by seeing which composition function yields the best result for classifying compound noun relations using an SVM for English two-word compounds. For evaluating human judgments, we used the British National Corpus (http://www.natcorp.ox.ac.uk/) for English and the most recent wikidump from July 3, 2016 (https://dumps.wikimedia.org/eswiki/latest/) for Spanish to train word2vec embeddings. The WikiExtractor was used to extract and clean the wikidump (https://github.com/attardi/wikiextractor). The gensim package was used to extract 500 dimensional word2vec vectors, using the CBOW algorithm. For both corpora, stop-words were removed, and words that occurred less than 100 times for English and 50 times for Spanish were excluded from the model's vocabulary. We resorted to creating our own datasets due to the unavailability of preexisting ones for three and four word compounds and Spanish compounds. The Stanford POS tagger version 3.5.2 was used to extract Spanish nominal compounds (Toutanova et al., 2003), and the BNC's tagset was used to extract English compounds. Spanish and English compounds for the test set were randomly chosen from looking at the list of compounds that included one of the top 400 words that occurred in the most compounds. This was to ensure that for each compound in the test set, there would be a sufficient number of compounds that share one constituent word for comparison. There were six test sets: two, three and four word compounds for Spanish and English. For Spanish, this is with respect to nouns only, not counting the preposition or determiners when determining the length of the compound. 25 compounds were chosen for each test set, totaling up to 150 test compounds. For each word in the test compound, another two-word compound sharing that word was chosen for comparison. So for the four word-test sets, there were 100 pairs for comparison, for the three-word compound sets, there were 75 pairs for comparison, and for the two-word compound sets, there were 50 pairs for comparison.

For example, "periodo de expansión del imperio" was paired with "expansión del universo", "embajador del imperio", and "periodo de ausencia". "Bomb squad chief" was paired with "bomb damage","drug squad", and "police chief".

The goal of analyzing compound noun relation classification is twofold; it will serve as another metric for comparing composition functions, a previously unused metric, and we will be able to determine if using the composition vectors as input to a classifier can improve overall performance of the classifier, a previously untried strategy. We only perform this experiment in English for two-word compounds due to the availability of large preexisting annotated data sets. (Verhoeven et al., 2012) and (Dima and Hinrichs, 2015) use word embeddings for semantic classification; however, they simply concatenate the embeddings for the constituent vectors as input. For classifying semantic relations in English, we experimented with our BNC model, Google News Vectors (available at https://code.google.com/archive/p/word2vec/), GloVe vectors (Pennington et al., 2014), CW vectors (Collobert et al., 2011) and HPCA vectors (Lebret and Collobert, 2013). For the utilization of word2vec vectors, we found better results with the Google News Vectors, probably due to the amount of data used to train them, so we report only classification results utilizing those here. For each embedding type, we chose the largest possible dimensions, since that has yielded the best results in (Dima and Hinrichs, 2015). Table II gives an overview of the information used to train each set. We used the dataset described

Method	Embedding Size	Dictionary Size	Training Data Size	Support Corpora
word2vec	300	3,000,000	100.00 bn	Google News
GloVe	300	400,000	42.00 bn	Common Crawl
HPCA	200	178,080	1.65 bn	enWiki+Reuters +WSJ
CW	50	130,000	0.85 bn	enWiki+Reuters RCV1
word2vec	500	30,025	100 mn	BNC
word2vec	500	19,679	120 mn	esWiki

Table 2: Overview of different embeddings

in (Tratz, 2011) (available at http://www.isi.edu/publications/licensed-sw/fanseparser/index.html), which consists

of 37 relations and 19,158 annotated compound nouns. Compounds with words that were not included in all of the different model embeddings' vocabularies were not included in the analysis, leaving a total of 18669 compounds. The set was partitioned into a training module that was 80% of the original set and a test set that was 20%. After experimenting with a variety of different classifiers and architectures, we used the Weka machine learning software (https://weka.wikispaces.com/) to implement an SVM with a polykernel, with feature selection using a gain ratio attribute evaluator and a ranker search. To create the input features, we concatenated the vectors for the constituent nouns and the composition function vector. We experimented with using the different word embeddings individually and in conjunction, and found the best results by concatenating the constituent and composition embeddings from the Google News word2vec, GloVe, HPCA, and CW sets, similar to the work done by (Dima and Hinrichs, 2015).

4.3 Collecting Similarity Judgments

Responses were collected using Survey Gizmo (`https://www.surveygizmo.com/`), using unpaid volunteers. Subjects were asked to rate how similar or dissimilar compound noun pairs were on a Likert scale. Each pair was presented twice, once as "compound 1, compound 2" and again as "compound 2, compound 1" to account for a asymmetry of human judgments. Pairs were presented in random order. Surveys were self paced and took approximately fifteen minutes. For the English survey, there were 7 participants. For the Spanish survey, there were 4 participants. Participants ranged in age from 15-55, and were self-reportedly fluent in the language of the survey. For each pair, the average similarity was calculated on a scale of 1 to 5, 5 being most similar and 1 being the most dissimilar. Ratings from each participant were averaged to use to correlate with the model's cosine similarity predictions.

4.4 Composition Methods

For the human judgment correlation task, for each compound in the test and comparison set, representations were generated by taking the vector representations from the word2vec model using the CBOW algorithm trained from the BNC and esWiki corpora. For entries in the Spanish test set, only the nouns were considered for composing the phrase. Since the preposition is largely semantically empty and only serves to illustrate the syntactic connection between the nouns, it is ignored. As we have previously seen, the preposition "de" encodes a wide variety of semantic relations; however, there is a minority of nominal compounds that use different prepositions like "por","para", "entre", etc. We will naively assume here that the preposition does not encode semantic information and focus only on compounds using the most common preposition "de", which is a bit of a generalization. Articles were also ignored, since they also do not provide much semantic meaning, especially considering their more generalized usage in Spanish compared to English. The composition of the constituent words for each compound was then calculated using the following functions: simple additive (equation (2)), multiplicative (equation (3)), tensor product (equation (4)), head only (equation (7) for English, (6) for Spanish), modifier only(equation (6) for English, (7) for Spanish), weighted additive(equation (5)), and combined weighted additive and multiplicative(equation (8)). For three word compounds, data was parsed by hand

	Combined Model			OWA		NWA	
	α	β	γ	α	β	α	β
two-word Spanish	0.099	0.101	0.000	0.098	0.098	0.5	0.5
two-word English	0.267	0.264	9.697	0.874	0.898	0.5	0.5
three-word Spanish	1.452	1.943	-0.006	1.333	1.749	0.2	0.8
three-word English	0.842	0.724	0.000	0.821	0.719	0.9	0.1
four-word Spanish	0.949	1.387	4.422	1.065	1.639	0.1	0.9
four-word English	0.939	0.869	2.580	0.927	0.869	0.1	0.9

Table 3: Parameters for the combined, optimized weighted additive, and normalized weighted addtive models

into (n1 n2) n3 or n1 (n2 n3) so that syntactically sensitive functions could be properly applied recursively. The same method was applied to four-word compounds. For compounds longer than two words, the head only and modifier only models were not calculated, since there are multiple modifiers and heads.

4.5 Determining the Parameters of the Weighted Additive and Combined Models

The parameters of the weighted additive model were determined in two different ways. First, we considered nine models, with weights varying from 0.1 to 0.9 in a step size of 0.1, where the sum of α and β adds to one, where the model with the highest correlation to the human judgments was taken as optimal. For the purposes of this experiment, the magnitude of the vector does not matter, because the cosine similarity is taken for the final metric, which does not take magnitude into account. We used grid search to find the optimal values for α and β, but without the constraint that they had to add to one, again maximizing the correlation to human judgments. Likewise, for the combined model, we used a similar grid search, without the traditional constraint. The model parameters are described in Table III, where NWA stands for normalized weighted additive and OWA stands for optimized weighted additive.

In Spanish, the head is the first noun, and would be weighted with α, whereas the head is the second noun in English, and would be weighted with β. So we see that heavily weighting the modifier is a consistent trend across the combined, normalized additive, and optimized additive models in English and Spanish for compounds longer than two words, with the exception of the four-word normalized additive English set. This inconsistency could be due to idiosyncrasies in the relatively small data set. For two-word compounds in English and Spanish, an even weight distribution yielded the best results. This could imply that as the length of the compound noun grows, the semantic importance of the modifier increases.

5 Evaluation

For the human similarity judgments, we calculated intersubject agreement using Spearman's ρ, using leave-out one resampling as employed by (Mitchell and Lapata, 2008), with the results given in Table 4.

2W Spanish	2W English	3W Spanish	3W English	4W Spanish	4W English
0.341	0.441	0.357	0.347	0.170	0.321

Table 4: Intersubject Agreement for Human Similarity Judgments

For the two-word English set, we see that the similarity judgment is consistent with previous work, where (Mitchell and Lapata, 2010) achieved a Spearman's correlation coefficient of 0.49. As a general trend, inter-subject agreement declines as the compounds get longer.

	2W Span	2W Eng	3W Span	3W Eng	4W Span	4W Eng
simple additive	0.365	0.617	0.585	0.331	0.230	0.650
multiplicative	0.258	0.624	0.227	-0.057	0.105	0.372
tensor	0.357	0.621	0.040	-0.041	0.266	0.321
head	0.280	0.443				
modifier	0.191	0.060				
normalized weighted additive	0.365	0.617	0.521	0.312	0.289	0.336
optimized weighted additive	0.371	0.633	0.690	0.330	0.435	0.654
optimized combined	0.342	0.670	0.652	0.338	0.434	0.658

Table 5: Spearman's correlation between human similarity judgments and cosine similarity predictions

We evaluated the similarity of two compounds by taking the cosine of their vectors, a commonly used metric (Mitchell and Lapata, 2010). To test if a composition model's results were consistent with human judgments, we used Spearman's correlation, where we compared the cosine with the average human similarity judgment. Similar to (Mitchell and Lapata, 2010), the results indicate that the similarity judgment task was relatively difficult, but there still was a decent amount of consistency between partic-

ipants. Our study finds that this task becomes more difficult as the compounds get longer.

For noun relation classification, we used two metrics. We performed a ten-fold cross-validation on the training set, and also tested each model on the unseen test set. For the parameterized functions, we used the optimized parameter values from the corresponding human judgment correlation test. Since the optimal normalized parameters from the 2-word English set was 0.5 and 0.5, we did not perform a test for the normalized weighted additive set, since the proportions are the same as the simplified additive model. We also did not test the tensor product model, due to constraints in dimensionality.

6 Results

6.1 Correlation with Human Similarity Judgments

Table 5 shows the model's predictions correlated with the human judgment using Spearman's ρ.

Consistent with the work of (Mitchell and Lapata, 2010), all compositional models outperform the head-only and modifier-only models, indicating the utility of the composition functions. The simple additive model and the multiplicative model yield comparable results for two-word compounds, but the effectiveness of the multiplicative model declines for longer compounds. This could be due to the previously discussed fact that zero or low-valued entries in the vector can essentially "throw away" data in the component vector, leading to poor results as more vectors are composed. As more and more vectors are composed, this problem is exacerbated and begins to affect performance. Likewise, the tensor product performs well on two-word compounds in comparison with the additive model, but less so on longer compounds, especially three-word compounds. This may imply that in addition to dimensionality challenges, the tensor product may face similar limitations to the multiplicative model for composing larger phrases. For the optimized weighted additive and combined models, the results are very comparable, with the optimized additive model slightly outperforming the normalized additive model. The combined and weighted additive models yield the most promising results, especially since their accuracy is relatively consistent for handling longer phrases. The increasing inaccuracy of the multiplicative and tensor models and the consistency of the combined and weighted additive models for longer compounds are new insights for the effectiveness of these models, which has serious consequences for attempting to build models that can handle longer phrase structures in general. This work suggests that the utility of each function can vary with the length of the sentence, which suggests the importance of performing more work on structures longer than two-words, which has been the standard for work in compound nouns until now. This paper presents strong evidence that the multiplicative model, although promising in previous work handling two-word phrases, has serious shortcomings for handling more complex phrases.

6.2 Compound Noun Relation Classification

Table 6 gives the results for each tested function on the different word embeddings, including the concatenation of all the different embeddings. The CV column represents the 10-fold cross-validation accuracy, and the test set is comprised of unseen noun compounds. Input with only the constituent vector embeddings without the composition function was also tested to give a baseline. Adding the composition function improves the performances for every type of embedding, with the most dramatic improvement in the concatenated word2vec+HPCA+CW+GloVe model.

We achieved the best results using the concatenation of the word2vec, HPCA, CW, and GloVe embeddings. Adding the composition function improves this models performance by as much as 2.02% using the multiplicative function, demonstrating the utility of using a compositional function during classification. The simple additive and weighted additive models actually perform worse in cross-validation than using no composition function at all. The combined models γ parameter was 9.697, so the multiplicative component of the combined model mostly overpowers the additive components, which explains why its performance is similar to that of the multiplicative model.

Our model slightly outperforms (Dima and Hinrichs, 2015), with its high cross-validation score being 77.7%, and is comparable to the state of the art model of (Tratz and Hovy, 2010), achieving 79.3%.

	word2vec+HPCA+CW+GloVe		word2vec		GloVe		HPCA		CW	
	CV	test set	CV	test set	CV	test set	CV	test set	CV	test set
no composition function	76.76	77.3	75.41	76.67	73.35	73.21	71.60	72.39	61.96	62.26
simple additive	76.52	76.95	75.85	76.01	72.63	73.59	71.36	71.59	61.98	62.23
weighted additive	76.47	77.70	74.80	76.76	72.70	73.62	71.37	71.48	62.02	62.31
multiplicative	**78.78**	**78.23**	75.82	76.04	73.42	73.30	71.95	72.50	62.52	62.58
combined	78.69	78.09	75.99	76.09	73.38	73.30	71.36	71.59	62.27	62.18

Table 6: Cross-validation accuracy and accuracy on an unseen test set for semantic relation classification

However, the model of (Tratz and Hovy, 2010) only achieves 51% accuracy on an unseen test set, whereas our model is much more consistent, with 78.23% accuracy. Again, we narrowly outperform (Dima and Hinrichs, 2015), with its accuracy on an unseen test set, which was 77.12% (Dima and Hinrichs, 2015). (Tratz and Hovy, 2010) use a slightly different set of relations and data set, but similar to the work of (Dima and Hinrichs, 2015), the consistency when testing unseen compounds points to the robustness of our model in comparison to (Tratz and Hovy, 2010). It is also clear that the small performance increase spurred by the addition of the composition function gives our model its slight increase in accuracy over the model of (Dima and Hinrichs, 2015), with a 4.84% decrease in relative error for cross-validation and 4.85% decrease in relative error for an unseen test set.

7 Discussion

With regards to the effectiveness of the additive and multiplicative classes of models, this paper presents strong evidence that multiplicative class models do not perform well for longer compound nouns, which have been previously untested. This idea is further supported by the low γ parameters in the optimized combined model for three and four word compounds. However, within the context of semantic relation classification, the multiplicative model is the strongest, whereas the additive model does not improve performance significantly, and sometimes even worsens performance. One interesting direction of future study would be to see which function performs best for classifying longer compounds, since the multiplicative model did not perform well for the human similarity correlation task for longer compounds. This paper also suggests that the semantic importance of the head noun diminishes as the compound gets longer, and that the semantic importance of the modifier becomes greater, as illustrated by the optimized parameters of the weighted additive models. One future direction of study would be to implement more complex composition functions, or to incorporate information from the prepositions in Spanish compound nouns into the composition vector. Another direction of study would be to expand the noun relation classification task to a Spanish data set, and compare results, or to expand the classification task to three or four word compounds in English. This study points to the robustness of the combined model, since it is able to capture information from both the additive and multiplicative models. It performs well for three and four word compound human judgment similarity correlation, and it performs well in the relation classification task. The flexibility of its parameters, which can vary between languages and for compound nouns differing in length, makes it very promising.

8 Conclusion

The goal of this research is to find the optimal way to represent compound nouns of length two or greater using a vector-based representation. We have illustrated the utility of the multiplicative model in relation classification, but it has shortcomings in representing larger phrases in comparison to the additive class of models. Our new classification system, which incorporates composition vectors into SVMs, is comparable to other state-of-the-art models using cross-validation, or slightly outperforms them using an unseen test set.

References

[Al-Rfou et al. 2013] Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. *CoNLL-2013 (2013): 183.*.

[Baroni and Zamparelli 2010] Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *2010 Conference on Empirical Methods in NLP*, 1183-1193.

[Bauke 2014] Leah S Bauke. 2014. *Symmetry breaking in syntax and the lexicon*, volume 216. John Benjamins Publishing Company.

[Collobert et al. 2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12 (Aug):2493-2537.

[Dima and Hinrichs 2015] Corina Dima and Erhard Hinrichs. 2015. Automatic noun compound interpretation using deep neural networks and word embeddings. *IWCS 2015*,173.

[Erk 2012] Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

[Etcheverry and Wonsever 2016] Mathias Etcheverry and Dina Wonsever. 2016. Spanish word vectors from wikipedia. In *LREC 2016*, May.

[Finin 1980] Timothy Wilking Finin. 1980. The semantic interpretation of compound nominals.

[Girju 2007] Roxana Girju. 2007. Improving the interpretation of noun phrases with cross-linguistic information. In *annual meeting for computation linguistics*, volume 45, page 568.

[Girju 2009] Roxana Girju. 2009. The syntax and semantics of prepositions in the task of automatic interpretation of nominal phrases and compounds: A cross-linguistic study. *Computational Linguistics*, 35(2):185–228.

[Guevara 2010] Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *2010 Workshop on Geometric Models of NL Semantics*, pages 33-37. ACL.

[Im Walde et al. 2013] Sabine Schulte Im Walde, Stefan Müller, and Stephen Roller. 2013. Exploring vector space models to predict the compositionality of german noun-noun compounds. In 2nd Joint Conference on Lexical and Computational Semantics, 255-265.

[Kim and Baldwin 2005] Su Nam Kim and Timothy Baldwin. 2005. Automatic interpretation of noun compounds using wordnet similarity. In *ICON*,945-956.

[Lebret and Collobert 2013] Rémi Lebret and Ronan Collobert. 2013. Word embeddings through Hellinger PCA. *EACL 2014 (2014): 482.*.

[Mikolov and Dean 2013] T Mikolov and J Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

[Mitchell and Lapata 2008] Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL*, pages 236–244.

[Mitchell and Lapata 2010] Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388-1429.

[Moyna 2011] María Irene Moyna. 2011. *Compound words in Spanish: theory and history*, volume 316. John Benjamins Publishing.

[Pennington et al. 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, 1532-43.

[Polajnar and Clark 2014] Tamara Polajnar and Stephen Clark. 2014. Reducing dimensions of tensors in type-driven distributional semantics.

[Reddy et al. 2011] Siva Reddy, Ioannis P Klapaftis, Diana McCarthy, and Suresh Manandhar. 2011. Dynamic and static prototype vectors for semantic composition. In *IJCNLP*, 705-713.

[Socher et al. 2012] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *2012 Conference on Empirical Methods in NLP and Comp. Natural Language Learning*, pages 1201-1211.

[Toutanova et al. 2003] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *2003 Conference of the North Amer. Chapter of the ACL on Human Language Technology-Volume 1*, pages 173-180.

[Tratz and Hovy 2010] Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *48th Annual Meeting of the ACL*, pages 678-687.

[Tratz 2011] Stephen Tratz. 2011. *Semantically-enriched parsing for natural language understanding*. USC.

[Valle 2008] Ana Bocanegra Valle. 2008. On the teachability of nominal compounds to spanish learners of english for specific purposes. *English for Specific Purposes: Studies for Classroom Development and Implementation*, page 249.

[Verhoeven et al. 2012] Ben Verhoeven, Walter Daelemans, and Gerhard B Van Huyssteen. 2012. Classification of noun-noun compound semantics in dutch and afrikaans. In *PRASA 2012*, 121-125.