Using Sentence Plausibility to Learn the Semantics of Transitive Verbs

Tamara Polajnar  Laura Rimell  Stephen Clark
Computer Laboratory
University of Cambridge
United Kingdom
{tamara.polajnar, laura.rimell, stephen.clark}@cl.cam.ac.uk

Abstract

The functional approach to compositional distributional semantics considers transitive verbs to be linear maps that transform the distributional vectors representing nouns into a vector representing a sentence. We conduct an initial investigation that uses a matrix consisting of the parameters of a logistic regression classifier trained on a plausibility task as a transitive verb function. We compare our method to a commonly used corpus-based method for constructing a verb matrix and find that the plausibility training may be more effective for disambiguation tasks.

1 Introduction

The field of compositional distributional semantics seeks principled ways to combine distributional representations of words to form larger units. Representations of full sentences, besides their theoretical interest, have the potential to be useful for tasks such as automatic summarisation and recognising textual entailment. A number of recent studies have investigated ways to combine distributional representations of Subject, Verb, Object (SVO) triples to form transitive sentences, that is, sentences based on a transitive verb [2–4, 6, 7, 12, 14].

Under the functional approach [1–4], argument-taking words such as verbs and adjectives are represented as tensors, which take word vectors as arguments. A transitive verb can be viewed as a third-order tensor with input dimensions for the subject and object, and an output dimension for the meaning of the sentence as a whole. This approach has achieved promising initial results [6–9, 14], but many questions remain. Two outstanding questions are the best method of learning verb tensors from a corpus, and the best sentence space for a variety of different tasks.

This paper presents work in progress which addresses both of these questions. It compares two methods for learning verb representations, the distributional model of [7] in which positive examples of subject-object pairs for a given verb are structurally mixed, and the regression model of [14] in which positive and negative examples of subject-object pairs for a given verb are mapped into a plausibility space. A variety of methods for reducing the noun space and composing the verb, subject, and object representations are investigated. The results show that the plausibility training outperforms the distributional method on a verb disambiguation task, while the purely distributional approach performs better on sentence similarity.

2 Methods

In the definition of the functional approach to compositional distributional semantics [1–4], a transitive verb is a map that takes as arguments noun vectors representing the subject and object, and produces a vector in the sentence space. Typically, noun vectors for subject and object reside in a “topic space” where the dimensions correspond to co-occurrence features; we use a reduced space.
resulting from applying Singular Value Decomposition (SVD) to the raw co-occurrence space. The correct sentence space to use is less obvious; previous approaches have either mapped sentence meaning to the same topic-based noun space \[6, 7\] or defined a new space for sentence meaning, particularly plausibility space \[11, 14\].

If the verb function is a multi-linear map, then the verb is naturally represented by a third-order tensor. However, tensor training can be expensive and in practice, for some tasks, the verb can be approximated as a matrix \[7, 14\].

Below we describe two ways of learning a verb matrix. In the regression method, the learnt matrix consists of parameters from a plausibility classifier. The classifier is trained to distinguish plausible sentences like

animals eat plants

from implausible sentences like

animals eat planets.

In the distributional method, training is based on a sum of positive (i.e. attested) SVO triples. The acquisition of positive SVO data and plausibility training data is described in Section 2.2.

2.1 Verbs

Regression (reg) Following \[14\], we formulate regression learning as a plausibility task where the class membership can be estimated with a single variable. To produce a scalar output, we can learn the parameters for a single \(K \times K\) matrix \(V\) using standard logistic regression with the mean squared error cost function and \(K\)-dimensional subject \((n_s)\) and object \((n_o)\) noun vectors as input:

\[
O(V) = -\frac{1}{m} \left[ \sum_{i=1}^{N} t_i \log h_V (n_s^i, n_o^i) + (1 - t_i) \log h_V (n_s^i, n_o^i) \right]
\]

where \(t_i\) are the true plausibility labels of the \(N\) training examples. The function \(h_V (n_s^i, n_o^i) = \sigma((n_s^i)V(n_o^i)^T)\) is a sigmoid transformation of the scalar that results from the inner product and the objective is regularised by the parameter \(\lambda: O(V) + \frac{\lambda}{2} ||V||^2\).

All of the plausibility data is used for training and validation, as we employ only the resulting verb matrix to produce a sentence space for transitive verb phrases in the test data. The regression algorithm is trained through gradient descent with Adagrad \[5\] and 10% of the training triples are used as a validation set for early stopping.

Distributional (dist) Following \[7\], we generate a \(K \times K\) matrix for each verb as the average of outer products of subject and verb vectors from the positively labelled subset of the training data:

\[
V = \frac{1}{N_p} \left[ \sum_{i=1}^{N_p} n_s^i \otimes n_o^i \right]
\]

where \(\otimes\) is outer product and \(N_p\) is the number of positive training examples. The intuition is that the matrix encodes higher weights for contextual features of frequently attested subjects and objects; for example, multiplying by the matrix for eat may yield a higher scalar value when its subject exhibits features common to animate nouns, and its object exhibits features common to edible nouns.

2.2 Training data

In order to generate training data we find plausible SVO triples that occur in an October 2013 dump of Wikipedia. To ensure quality we choose triples whose nouns occur at least 100 times. For some verbs there are thousands of such triples, so we choose the top 300 most frequent triples for each verb. For each verb we generate negative examples by substituting plausible subjects and objects with maximally dissimilar nouns. Specifically, for a given subject (or object) noun we calculate its average sum with the centroid of plausible subject (or object) vectors, and then select the frequency-matched noun with lowest cosine similarity to this average. The noun vectors are generated from the Wikipedia corpus using the t-test weighting scheme and normalisation techniques described in \[13\]. These techniques enable us to learn using SVD reduced vectors with dimensions as low as 20.

2.3 Composition Methods

We investigate the following methods of composing the verb matrix with the subject and object vectors to form a vector representation for a transitive sentence.
Table 1: Best performance for each method on the two dataset across combination methods and noun dimensions. The results for reg show the standard deviation over 3 repetitions, with dist being deterministic. In general there was very little variance between runs.

| DATASET | METHOD | SPEARMAN | SETTINGS     |
|---------|--------|----------|--------------|
| KS2014  | reg    | 0.28 ± 0.0000 | comb: REL, K=20 |
|         | dist   | 0.35 ± 0.0000 | comb: REL, K=100 |
| GS2011  | reg    | 0.37 ± 0.0037 | comb: VO, K=200 |
|         | dist   | 0.25 ± 0.0000 | comb: CS, K=200 |

Relational: from [7, 10], the meaning of a transitive sentence is a matrix, obtained by the following formula, where ⊗ is outer product, ⊙ is elementwise product, and × is matrix multiplication:

\[
\begin{align*}
\text{subj} \, \text{verb} \, \text{obj} & = (\text{subj} \odot \text{obj}) \odot \text{verb} = (\text{subj} \times \text{obj}^T)_{ij} \cdot \text{verb}_{ij} \\
\text{verb} & = (\text{subj} \times \text{obj})_i \\
\end{align*}
\]  

(1)

Copy-subject: from [9, 10], the meaning of a transitive sentence is a vector, obtained by:

\[
\begin{align*}
\text{subj} \, \text{verb} \, \text{obj} & = \text{subj} \odot (\text{verb} \times \text{obj}) = \text{subj}_i \cdot (\text{verb} \times \text{obj})_i \\
\text{verb} & = \text{verb} \times \text{obj} \\
\end{align*}
\]

(2)

Verb-object: this method tests whether the verb and object alone are enough to measure sentence similarity. It can be compared directly to Copy-subject and reflects the linguistic generalisation that a verb selects its object more strongly than its subject. The meaning of a transitive sentence is a vector, obtained by:

\[
\begin{align*}
\text{subj} \, \text{verb} \, \text{obj} & = \text{verb} \times \text{obj} \\
\end{align*}
\]

(3)

For the Relational method, sentence similarity is measured as the Frobenius inner product of the two sentence matrices. For Copy-subject and Verb-object, sentence similarity is measured as the cosine of the two sentence vectors.

3 Tasks

We investigate the performance of the regression learning method on two tasks: verb disambiguation, and transitive sentence similarity. In each case the system must compose SVO triples and compare the resulting semantic representations.

For the verb disambiguation task we use the GS2011 dataset [7]. This dataset consists of pairs of SVO triples in which the subject and object are held constant, and the verb is manipulated to highlight different word senses. For example, the verb draw has senses that correspond to attract and depict. The SVO triple report draw attention has high similarity to report attract attention, but low similarity to report depict attention. Conversely, child draw picture has high similarity to child depict picture, but low similarity to child attract picture. The gold standard consists of human judgements of the similarity between pairs, and we measure the correlation of the system’s similarity scores to the gold standard judgements.

For the transitive sentence similarity task we use the KS2013 dataset [9]. This dataset also consists of pairs of SVO triples, but the verb as well as the subject and object vary. For example, author write book and delegate buy land are judged by most annotators to be very dissimilar, while programme offer support and service provide help are considered highly similar. Again, we measure the correlation between the system’s similarity scores and the gold standard judgements.

4 Results

Table 1 and Figure 1 summarise the results of our experiments. Overall, reg performs better than dist on verb disambiguation, while dist performs better on sentence similarity. We hypothesise the
difference lies in the nature of the two tasks. The verb disambiguation task is inherently plausibility-based, because one member of the low-similarity pairs (with the non-relevant sense of the verb) is always implausible. For example, when disambiguating the verb *meet*, the triple *system satisfy requirement* has high plausibility, while *system visit requirement* has low plausibility. In fact, the Verb-object (VO) composition method alone is enough to disambiguate these triples using \texttt{reg}.

On the other hand, both triples in the sentence similarity task tend to be highly plausible, even when their topic differs. Because \texttt{dist} uses a topic-based space, it may better capture these distinctions. For example, *girl like people* and *woman ask man* are both highly plausible, but annotators have judged the pair to have low similarity on average. The best distributional method (K=100, Relational composition) ranks this pair the middle, while the corresponding regression method (K=100, Relational composition) ranks this pair near the top of the range of similarities.

Figure 1 shows the overall trends of the two methods. The \texttt{reg} method consistently outperforms \texttt{dist} when low-dimensional vectors are used. Due to the nature of functional composition of distributional representations, where each word-type other than noun is represented by a higher-order tensor, low-dimensional representations are particularly advantageous. On the other hand, the \texttt{dist} method is very quick to train and therefore can be used with higher dimensional vectors for tasks where data storage is not an issue.

5 Conclusion

The difference in performance of the two methods underlines the need to find the appropriate sentence spaces for particular tasks. This preliminary study indicates that plausibility training may be better suited for disambiguation. Further work will consist of more in depth analysis and optimisation of the training procedure, as well as investigation into ways of low-cost learning of task-specific sentence spaces.

Acknowledgements

Tamara Polajnar is supported by the ERC Starting Grant, DisCoTex, awarded to Stephen Clark, and Laura Rimell is supported by the EPSRC grant EP/I037512/1: A Unified Model of Compositional and Distributional Semantics: Theory and Applications.
References

[1] Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, 9:5–110, 2014.

[2] Marco Baroni and Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-10)*, pages 1183–1193, Cambridge, MA, 2010.

[3] Stephen Clark. Type-driven syntax and semantics for composing meaning vectors. In Chris Heunen, Mehrnoosh Sadrzadeh, and Edward Grefenstette, editors, *Quantum Physics and Linguistics: A Compositional, Diagrammatic Discourse*, pages 359–377. Oxford University Press, 2013.

[4] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. In J. van Bentham, M. Moortgat, and W. Buszkowski, editors, *Linguistic Analysis (Lambek Festschrift)*, volume 36, pages 345–384, 2010.

[5] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July 2011.

[6] Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. Multi-step regression learning for compositional distributional semantics. *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, 2013.

[7] Edward Grefenstette and Mehrnoosh Sadrzadeh. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404, Edinburgh, Scotland, UK, July 2011.

[8] Edward Grefenstette and Mehrnoosh Sadrzadeh. Experimenting with transitive verbs in a discocat. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language*, Edinburgh, Scotland, UK, 2011.

[9] Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. Prior disambiguation of word tensors for constructing sentence vectors. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Seattle, USA, October 2013.

[10] Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. A study of entanglement in a categorical framework of natural language. In *Proceedings of the 11th Workshop on Quantum Physics and Logic (QPL)*, Kyoto, Japan, June 2014.

[11] Jayant Krishnamurthy and Tom M Mitchell. Vector space semantic parsing: A framework for compositional vector space models. In *Proceedings of the 2013 ACL Workshop on Continuous Vector Space Models and their Compositionality*, Sofia, Bulgaria, 2013.

[12] Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *Proceedings of ACL-08*, pages 236–244, Columbus, OH, 2008.

[13] Tamara Polajnar and Stephen Clark. Improving distributional semantic vectors through context selection and normalisation. In *14th Conference of the European Chapter of the Association for Computational Linguistics, EACL’14*, Gothenburg, Sweden, 2014.

[14] Tamara Polajnar, Luana Fagarasan, and Stephen Clark. Reducing dimensions of tensors in type-driven distributional semantics. In *Proceedings of EMNLP 2014*, Doha, Qatar, 2014.