Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing

Angeliki Lazaridou, Eva Maria Vecchi and **Marco Baroni**
Center for Mind/Brain Sciences
University of Trento, Italy
first.last@unitn.it

Abstract

In this work, we argue that measures that have been shown to quantify the degree of semantic plausibility of phrases, as obtained from their compositionally-derived distributional semantic representations, can resolve syntactic ambiguities. We exploit this idea to choose the correct parsing of NPs (e.g., (*live fish*) *transporter* rather than *live* (*fish transporter*)). We show that our plausibility cues outperform a strong baseline and significantly improve performance when used in combination with state-of-the-art features.

1 Introduction

Live fish transporter: A transporter of live fish or rather a fish transporter that is not dead? While our intuition, based on the meaning of this phrase, prefers the former interpretation, the Stanford parser, which lacks semantic features, incorrectly predicts the latter as the correct parse.[1] The correct syntactic parsing of sentences is clearly steered by semantic information (as formal syntacticians have pointed out at least since Fillmore (1968)), and consequently the semantic plausibility of alternative parses can provide crucial evidence about their validity.

An emerging line of parsing research capitalizes on the advances of compositional distributional semantics (Baroni and Zamparelli, 2010; Guevara, 2010; Mitchell and Lapata, 2010; Socher et al., 2012). Information related to compositionally-derived distributional representations of phrases is integrated at various stages of the parsing process to improve overall performance.[2] We are aware of two very recent studies exploiting the semantic information provided by distributional models to resolve syntactic ambiguity: Socher et al. (2013) and Le et al. (2013).

Socher et al. (2013) present a recursive neural network architecture which jointly learns semantic representations and syntactic categories of phrases. By annotating syntactic categories with their distributional representation, the method emulates lexicalized approaches (Collins, 2003) and captures similarity more flexibly than solutions based on hard clustering (Klein and Manning, 2003; Petrov et al., 2006). Thus, their approach mainly aims at improving parsing by capturing a richer, data-driven categorial structure.

On the other hand, Le et al. (2013) work with the output of the parser. Their hypothesis is that parses that lead to less semantically plausible interpretations will be penalized by a reranker that looks at the composed semantic representation of the parse. Their method achieves an improvement of 0.2% in F-score. However, as the authors also remark, because of their experimental setup, they cannot conclude that the improvement is truly due to the semantic composition component, a crucial issue that is deferred to further investigation.

This work aims at corroborating the hypothesis that the semantic plausibility of a phrase can indeed determine its correct parsing. We develop a system based on simple and intuitive measures, ex-

[1] http://nlp.stanford.edu:8080/parser/index.jsp

[2] Distributional representations approximate word and phrase meaning by vectors that record the contexts in which they are likely to appear in corpora; for a review see, e.g., Turney and Pantel (2010).

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1908–1913,
Seattle, Washington, USA, 18-21 October 2013.

Type of NP	#	Example
A (N N)	1296	*local phone company*
(A N) N	343	*crude oil sector*
N (N N)	164	*miracle home run*
(N N) N	424	*blood pressure medicine*
Total	2227	-

Table 1: NP dataset

tracted from the compositional distributional representations of phrases, that have been shown to correlate with semantic plausibility (Vecchi et al., 2011).

We develop a controlled experimental setup, focusing on a single syntactic category, that is, noun phrases (NP), where our task can be formalized as (left or right) bracketing. Unlike previous work, we compare our compositional semantic component against features based on n-gram statistics, which can arguably also capture some semantic information in terms of frequent occurrences of meaningful phrases. Inspired by previous literature demonstrating the power of metrics based on Pointwise Mutual Information (PMI) in NP bracketing (Nakov and Hearst, 2005; Pitler et al., 2010; Vadas and Curran, 2011), we test an approach exploiting PMI features, and show that plausibility features relying on composed representations can significantly boost accuracy over PMI.

2 Setup

Noun phrase dataset To construct our dataset, we used the Penn TreeBank (Marcus et al., 1993), which we enriched with the annotation provided by Vadas and Curran (2007a), since the original treebank does not distinguish different structures inside the NPs and always marks them as right bracketed, e.g., *local (phone company)* but also *blood (pressure medicine)*. We focus on NPs formed by three elements, where the first can be an adjective (A) or a noun (N), the other two are nouns. Table 1 summarizes the characteristics of the dataset.[3]

Distributional semantic space As our source corpus we use the concatenation of ukWaC, the English Wikipedia (2009 dump) and the BNC, with a total of about 2.8 billion tokens.[4] We collect co-occurrence statistics for the top 8K Ns and 4K As, plus any other word from our NP dataset that was below this rank. Our context elements are composed of the top 10K content words (adjectives, adverbs, nouns and verbs). We use a standard bag-of-words approach, counting within-sentence collocates for every target word. We apply (non-negative) Pointwise Mutual Information as weighting scheme and dimensionality reduction using Non-negative Matrix Factorization, setting the number of reduced-space dimensions to 300.[5]

Composition functions We experiment with various composition functions, chosen among those sensitive to internal structure (Baroni and Zamparelli, 2010; Guevara, 2010; Mitchell and Lapata, 2010), namely dilation (**dil**), weighted additive (**wadd**), lexical function (**lexfunc**) and full additive (**fulladd**).[6] For model implementation and (unsupervised) estimation, we rely on the freely available *DISSECT* toolkit (Dinu et al., 2013).[7] For all methods, vectors were normalized before composing, both in training and in generation. Table 2 presents a summary description of the composition methods we used.

Following previous literature (Mitchell and Lapata, 2010), and the general intuition that adjectival modification is quite a different process from noun combination (Gagné and Spalding, 2009; McNally, 2013), we learn different parameters for noun-noun (NN) and adjective-noun (AN) phrases. As an example of the learned parameters, for the **wadd** model the ratio of parameters w_1 and w_2 is 1:2 for ANs, whereas for NNs it is almost 1:1, confirming the intuition that a non-head noun plays a stronger role in composition than an adjective modifier.

[3]The dataset is available from: `http://clic.cimec.unitn.it/composes`

[4]`http://wacky.sslmit.unibo.it`, `http://en.wikipedia.org`, `http://www.natcorp.ox.ac.uk`

[5]For tuning the parameters of the semantic space, we computed the correlation of cosines produced with a variety of parameter settings (SVD/NMF/no reduction, PMI/Local MI/raw counts/log transform, 150 to 300 dimensions in steps of 50) with the word pair similarity ratings in the MEN dataset: `http://clic.cimec.unitn.it/~elia.bruni/MEN`

[6]We do not consider the popular multiplicative model, as it produces identical representations for NPs irrespective of their internal structure.

[7]`http://clic.cimec.unitn.it/composes/toolkit/`

Model	Composition function	Parameters
wadd	$w_1\vec{u} + w_2\vec{v}$	w_1, w_2
dil	$\|\|\vec{u}\|\|_2^2\vec{v} + (\lambda - 1)\langle\vec{u}, \vec{v}\rangle\vec{u}$	λ
fulladd	$W_1\vec{u} + W_2\vec{v}$	$W_1, W_2 \in \mathbf{R}^{m\times m}$
lexfunc	$A_u\vec{v}$	$A_u \in \mathbf{R}^{m\times m}$

Table 2: Composition functions of inputs (u, v).

Recursive composition In this study we also experiment with recursive composition; to the best of our knowledge, this is the first time that these composition functions have been explicitly used in this manner. For example, given the left bracketed NP *(blood pressure) medicine*, we want to obtain its compositional semantic representation, $\overrightarrow{blood\ pressure\ medicine}$. First, *basic composition* is applied, in which $\overrightarrow{blood}$ and $\overrightarrow{pressure}$ are combined with one of the composition functions. Following that, we apply *recursive composition*; the output of basic composition, i.e., $\overrightarrow{blood\ pressure}$, is fed to the function again to be composed with the representation of $\overrightarrow{medicine}$.

The latter step is straightforward for all composition functions except **lexfunc** applied to left-bracketed NPs, where the first step should return a matrix representing the left constituent (*blood pressure* in the running example). To cope with this nuisance, we apply the **lexfunc** method to *basic composition* only, while recursive representations are derived by summing (e.g., $\overrightarrow{blood\ pressure}$ is obtained by multiplying the *blood* matrix by the *pressure* vector, and it is then summed to $\overrightarrow{medicine}$).

3 Experiments

Semantic plausibility measures We use measures of semantic plausibility computed on composed semantic representations introduced by Vecchi et al. (2011). The rationale is that the correct (wrong) bracketing will lead to semantically more (less) plausible phrases. Thus, a measure able to discriminate semantically plausible from implausible phrases should also indicate the most likely parse. Considering, for example, the alternative parses of *miracle home run*, we observe that *home run* is a more semantically plausible phrase than *miracle home*. Furthermore, we might often refer to a baseball player's miracle home run, but we doubt that even a miracle home can run! Given the composed representation of an AN (or NN), Vecchi et al. (2011) define the following measures:

- *Density*, quantified as the average cosine of a phrase with its (top 10) nearest neighbors, captures the intuition that a deviant phrase should be isolated in the semantic space.
- *Cosine of phrase and head N* aims to capture the fact that the meaning of a deviant AN (or NN) will tend to diverge from the meaning of the head noun.
- *Vector length* should capture anomalous vectors.

Since length, as already observed by Vecchi et al., is strongly affected by independent factors such as input vector normalization and the estimation procedure, we introduce *entropy* as a measure of vector quality. The intuition is that meaningless vectors, whose dimensions contain mostly noise, should have high entropy.

NP Parsing as Classification Parsing NPs consisting of three elements can be treated as binary classification; given *blood pressure medicine*, we predict whether it is left- (*(blood pressure) medicine*) or right-bracketed (*blood (pressure medicine)*).

We conduct experiments using an SVM with Radial Basis Function kernel as implemented in the *scikit-learn* toolkit.[8] Our dataset is split into 10 folds in which the ratio between the two classes is kept constant. We tune the SVM complexity parameter *C* on the first fold and we report accuracy results on the remaining nine folds after cross-validation.

Features Given a composition function $\boldsymbol{f}$, we define the following feature sets, illustrated with the usual *blood pressure medicine* example, which are used to build different classifiers:

- $\boldsymbol{f_{basic}}$ consists of the semantic plausibility measures described above computed for the two-word phrases resulting from alternative bracketings, i.e., 3 measures for each bracketing, evaluated on *blood pressure* and *pressure medicine* respectively, for a total of 6 features.
- $\boldsymbol{f_{rec}}$ contains 6 features computed on the vectors resulting from the recursive compositions

[8] http://scikit-learn.org/

Features	Accuracy
$right$	65.6
pos	77.3
$lexfunc_{basic}$	74.6
$lexfunc_{rec}$	74.0
$lexfunc_{plausibility}$	76.2
$wadd_{basic}$	75.9
$wadd_{rec}$	78.2
$wadd_{plausibility}$	78.7
pmi	81.2
pmi+$lexfunc_{plausibility}$	82.9
pmi+$wadd_{plausibility}$	**85.6**

Table 3: Evaluation of feature sets from Section 3

(blood pressure) medicine and *blood (pressure medicine)*.

- $f_{plausibility}$ concatenates f_{basic} and f_{rec}.
- pmi contains the PMI scores extracted from our corpus for *blood pressure* and *pressure medicine*.[9]
- pmi + $f_{plausibility}$ concatenates pmi and $f_{plausibility}$.

Baseline Model Given the skewed bracketing distribution in our dataset, we implement the following majority baselines: *a*) $right$ classifies all phrases as right-bracketed; *b*) pos classifies NNN as left-bracketed (Lauer, 1995), ANN as right-bracketed.

4 Results and Discussion

Table 3 omits results for **dil** and **fulladd** since they were outperformed by the $right$ baseline. That **wadd**- and **lexfunc**-based plausibility features perform well above this baseline is encouraging, since it represents the typical default behaviour of parsers for NPs, although note that these features perform comparably to the pos baseline, which would be quite simple to embed in a parser (for English, at least). For both models, using both basic and recursive features leads to a boost in performance over basic features alone. Note that recursive features (f_{rec}) achieve at least equal or better performance than basic ones (f_{basic}). We expect indeed that in many cases the asymmetry in plausibility will be sharper when considering the whole NP rather than its sub-parts; a *pressure medicine* is still a conceivable concept, but *blood (pressure medicine)* makes no sense whatsoever. Finally, **wadd** outperforms both the more informative baseline pos and **lexfunc**. The difference between **wadd** and **lexfunc** is significant ($p < 0.05$)[10] only when they are trained with recursive composition features, probably due to our suboptimal adaptation of the latter to recursive composition (see Section 2).

The pmi approach outperforms the best plausibility-based feature set $wadd_{plausibility}$. However, the two make only a small proportion of common errors (29% of the total $wadd_{plausibility}$ errors, 32% for pmi), suggesting that they are complementary. Indeed the pmi + $wadd_{plausibility}$ combination significantly outperforms pmi alone ($p < 0.001$), indicating that plausibility features can improve NP bracketing on top of the powerful PMI-based approach. The same effect can also be observed in the combination of pmi + $lexfunc_{plausibility}$, which again significantly outperforms pmi alone ($p < 0.05$). This behaviour further suggests that the different types of errors are not a result of the parameters or type of composition applied, but rather highlights fundamental differences in the kind of information that PMI and composition models are able to capture.

One hypothesis is that compositional models are more robust for low-frequency NPs, for which PMI estimates will be less accurate; results on those low-frequency trigrams only (20% of the NP dataset, operationalized as those consisting of bigrams with frequency ≤ 100) revealed indeed that $wadd_{plausibility}$ performed 8.1% better in terms of accuracy than pmi.

5 Conclusion

Our pilot study showed that semantic plausibility, as measured on compositional distributional representations, can improve syntactic parsing of NPs. Our results further suggest that state-of-the-art PMI features and the ones extracted from compositional representations are complementary, and thus, when combined, can lead to significantly better results. Besides paving the way to a more general integration

[9] Several approaches to computing PMI for these purposes have been proposed in the literature including the *dependency model* (Lauer, 1995) and the *adjacency model* (Marcus, 1980). We implement the latter since it has been shown to perform better (Vadas and Curran, 2007b) on NPs extracted from Penn TreeBank.

[10] Significance values are based on t-tests.

of compositional distributional semantics in syntactic parsing, the proposed methodology provides a new way to evaluate composition functions.

The relatively simple-minded **wadd** approach outperformed more complex models such as **lexfunc**. We plan to experiment next with more linguistically motivated ways to adapt the latter to recursive composition, including hybrid methods where ANs and NNs are treated differently. We would also like to consider more sophisticated semantic plausibility measures (e.g., supervised ones), and apply them to other ambiguous syntactic constructions.

6 Acknowledgments

We thank Georgiana Dinu and German Kruszewski for helpful discussions and the reviewers for useful feedback. This research was supported by the ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

References

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, pages 1183–1193, Boston, MA.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational linguistics*, 29(4):589–637.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. DISSECT: DIStributional SEmantics Composition Toolkit. In *Proceedings of the System Demonstrations of ACL 2013*, Sofia, Bulgaria.

Charles Fillmore. 1968. The case for case. In Emmon Bach and Robert Harms, editors, *Universals in Linguistic Theory*, pages 1–89. Holt, Rinehart and Winston, New York.

Christina Gagné and Thomas Spalding. 2009. Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, 60:20–35.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of GEMS*, pages 33–37, Uppsala, Sweden.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*, pages 423–430. Association for Computational Linguistics.

Mark Lauer. 1995. Corpus statistics meet the noun compound: Some empirical results. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 47–54, Cambridge, MA.

Phong Le, Willem Zuidema, and Remko Scha. 2013. Learning from errors: Using vector-based compositional semantics for parse reranking. In *Proceedings of the ACL 2013 Workshop on Continuous Vector Space Models and their Compositionality*, Sofia, Bulgaria.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Mitchell P Marcus. 1980. *Theory of syntactic recognition for natural languages*. MIT press.

Louise McNally. 2013. Modification. In Maria Aloni and Paul Dekker, editors, *Cambridge Handbook of Semantics*. Cambridge University Press, Cambridge, UK. In press.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Preslav Nakov and Marti Hearst. 2005. Search engine statistics beyond the n-gram: Application to noun compound bracketing. In *Proceedings of CoNLL*, pages 17–24, Stroudsburg, PA, USA.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of COLING-ACL*, pages 433–440, Stroudsburg, PA, USA.

Emily Pitler, Shane Bergsma, Dekang Lin, and Kenneth Church. 2010. Using web-scale n-grams to improve base NP parsing performance. In *Proceedings of the COLING*, pages 886–894, Beijing, China.

Richard Socher, Brody Huval, Christopher Manning, and Andrew Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP*, pages 1201–1211, Jeju Island, Korea.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing with compositional vector grammars. In *Proceedings of ACL*, Sofia, Bulgaria.

Peter Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

David Vadas and James Curran. 2007a. Adding noun phrase structure to the Penn Treebank. In *Proceedings of ACL*, pages 240–247, Prague, Czech Republic.

David Vadas and James R Curran. 2007b. Large-scale supervised models for noun phrase bracketing. In *Proceedings of the PACLING*, pages 104–112.

David Vadas and James R. Curran. 2011. Parsing noun phrases in the penn treebank. *Comput. Linguist.*, 37(4):753–809.

Eva Maria Vecchi, Marco Baroni, and Roberto Zamparelli. 2011. (Linear) maps of the impossible: Capturing semantic anomalies in distributional space. In *Proceedings of the ACL Workshop on Distributional Semantics and Compositionality*, pages 1–9, Portland, OR.