Visual Denotations for Recognizing Textual Entailment

Dan Han[1]
dan.han@aist.go.jp

Pascual Martínez-Gómez[1]
pascual.mg@aist.go.jp

Koji Mineshima[2]
mineshima.koji@ocha.ac.jp

[1]Artificial Intelligence Research Center, AIST
[2]Ochanomizu University
Tokyo, Japan

Abstract

In the logic approach to Recognizing Textual Entailment, identifying phrase-to-phrase semantic relations is still an unsolved problem. Resources such as the Paraphrase Database offer limited coverage despite their large size whereas unsupervised distributional models of meaning often fail to recognize phrasal entailments. We propose to map phrases to their visual denotations and compare their meaning in terms of their images. We show that our approach is effective in the task of Recognizing Textual Entailment when combined with specific linguistic and logic features.

1 Introduction and Related Work

Recognizing Textual Entailment (RTE) is a challenging task that was described as *the best way of testing an NLP system's semantic capacity* (Cooper et al., 1994). In this task, given a text T and a hypothesis H, the objective is to recognize whether T implies H (yes), whether T contradicts H (no) or otherwise (unk). For example, given:

(T) Some men walk in the tall and green grass.
(H) Some people walk in the field.

the system needs to recognize that T implies H (yes). Although humans can easily solve these problems, machines face great difficulties (Dagan et al., 2013). RTE has been approached from different perspectives, ranging from purely statistical systems (Lai and Hockenmaier, 2014; Zhao et al., 2014) to purely logical (Bos et al., 2004; Abzianidze, 2015; Mineshima et al., 2015) and hybrid systems (Beltagy et al., 2013).

We evaluate our idea on top of a logic system since they generally offer a high precision and interpretability, which is useful to our purposes. In this approach, there are two main challenges. The first challenge is to model the logical semantic composition of sentences guided by the syntax and logical words (e.g. *most*, *not*, *some*, *every*). The second challenge is to introduce lexical knowledge that describes the relationship between words or phrases (e.g. *men* → *people*, *tall and green grass* → *field*).

Whereas the relationship *men* → *people* can be found in high precision ontological resources such as WordNet (Miller, 1995), phrasal relations such as *tall and green grass* → *field* are not available in databases such as the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013) despite their large size. Moreover, although unsupervised distributional similarity models have an infinite domain (given a compositional function on words), they often fail to identify entailments (e.g. *guitar* has a high similarity to *piano* but they do not entail each other). To address these issues, Roller et al. (2014) investigated supervised methods to identify word-to-word hypernym relations given word vectors whereas Beltagy et al. (2016) proposed a mechanism to extract phrase pairs from T and H and train a classifier to identify paraphrases in unseen T-H problems. Our approach is largely inspired by their work and our intention is to increase the performance of these phrase and sentence level entailment classifiers using multimodal features.

Our assumption is that *the same concept expressed using different phrase forms is mapped to similar visual representations* since humans tend to ground the meaning of phrases into the same visual denotation. In a similar line, Kiela and Bottou (2014) proposed a simple yet effective concatenation of pre-trained distributed word representations and visual features, whereas Izadinia et al. (2015) suggests a tighter parametric integration using a set of hand annotated phrasal entail-

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2853–2859
Copenhagen, Denmark, September 7–11, 2017.

ment relations; however, their work was limited to recognizing word or phrase relations, ignoring the additional challenges that come in RTE which we show is critical. Young et al. (2014) and Lai and Hockenmaier (2014) did tackle sentence-level RTE using visual denotations. However, their approach is only applicable to those RTE problems whose words or phrases appear in the FLICKR30K corpus, which is a considerable limitation. Lai and Hockenmaier (2017) extended the approach to also recognize unseen phrasal semantic relations using a neural network augmented with conditional probabilities estimated from visual denotations. Instead, our approach is much simpler and similarly effective.

Our contribution is a method to judge phrase-to-phrase semantic relations using an asymmetric similarity scoring function between their sets of visual denotations. We identify the conditions in which this function contributes to sentence-level RTE and show empirically its benefit. Our approach is simpler than previous methods and it does not require annotated phrase relations. Moreover, this approach is not limited to specific corpora or evaluation datasets and it is potentially language independent.

2 Methodology

We formulate our framework in terms of a classifier $g_\theta : \mathcal{T} \times \mathcal{H} \rightarrow \{\text{yes}, \text{no}, \text{unk}\}$ that outputs an entailment judgment for any text $T \in \mathcal{T}$ and hypothesis $H \in \mathcal{H}$. There are three key issues in designing an effective classifier that uses visual denotations: i) to discern when it is appropriate to use visual denotations to recognize phrasal entailments, ii) to extract candidate phrase pairs and iii) to map those phrases into visual denotations[1] and measure their semantic similarity in terms of their associated images.

Textual and Logic Features The first issue is to understand the linguistic and logic limitations of visual denotations in recognizing phrasal entailments. From our observations, the linguistic phenomena that make visual denotations ineffective are word-to-word verb relations (e.g. *laughing* and *crying*) since their associated images may depict different actions with similar entities (e.g. pictures of *a baby crying* are similar to those of *a baby laughing*); antonym relations between any word in a phrase pair (e.g. similar images for *big car* and *small vehicle*); and words that denote people of different gender (e.g. *boy* versus *lady*, *man* versus *woman*) as they often display high visual similarity compared to other entities. The logic phenomena we identified signal sentences with small differences in critical words, phrases or structures, as in the presence of negations (e.g. images of *no cat* still display cats), passive-active constructions and subject-object case mismatches (e.g. images of *boy eats apple* and *apple eats boy* are similar) between T and H.

These logic phenomena can be easily detected from logic formulas with the aid of the variable unification during the theorem proving process. For instance, using event semantics (Davidson, 1967; Parsons, 1990), an active sentence *a boy eats an apple* and its corresponding passive sentence *an apple is eaten by a boy* can be compositionally mapped to the same logical formula, i.e., $\exists e \exists x \exists y (\textbf{boy}(x) \wedge \textbf{apple}(y) \wedge \textbf{eat}(e) \wedge (\textbf{subj}(e) = x) \wedge (\textbf{obj}(e) = y))$, while *a boy eats an apple* and *an apple eats a boy* are mapped to different formulas. When trying to prove the formula corresponding to H from the formula corresponding to T, one needs to unify the variables contained in these formulas, so that the non-logical predicates such as **boy**, **apple** and **eat** in T and H are aligned by taking into account logical signals.

Extract candidate phrase pairs between T and H The second issue is to find candidate phrase pairs between T and H for which we compare their visual denotations. In our running example (see Figure 1), a desirable candidate phrase pair would be *tall and green grass* and *field*. We use a tree mapping algorithm (Martínez-Gómez and Miyao, 2016) that finds node correspondences between the syntactic trees of T and H. The search is carried out bottom-up, guided by an ensemble of cost functions. This ensemble rewards word or phrase correspondences that are equal or if a linguistic relationship (i.e. synonymy, hypernymy, etc.) holds between them according to WordNet. This tree mapping implicitly defines hierarchical phrase pair correspondences between T and H. We only select those phrase pairs for which both phrases have less than 6 words. We believe that discerning the entailment relation between longer phrases should be left to the logic prover and the compositional mechanism of meaning.

[1]We approximate the visual denotations of a phrase by obtaining the images associated to that phrase.

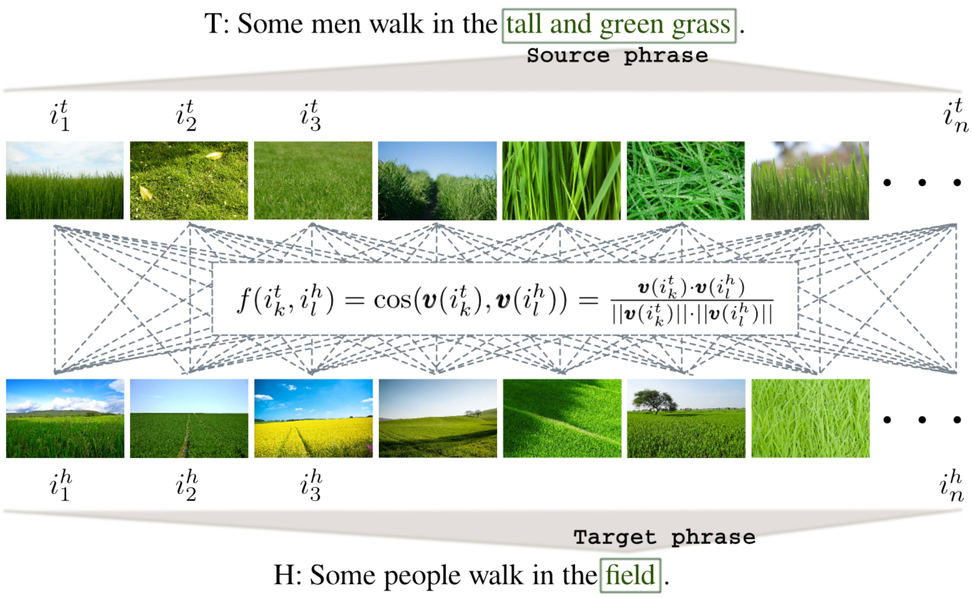

Figure 1: Phrase-image mappings for the phrase pair *tall and green grass* and *field* in one RTE problem.

Visual Features At this stage it remains to measure the semantic relation between the candidate phrase pairs (extracted with the tree mapping algorithm described above) using their visual denotations (see Figure 1 for a schematic diagram[2]). For this purpose, we select the phrase pairs (t, h) with highest and lowest *similarity score*. We define the similarity score as the average cosine similarity between the *best* image correspondences. That is:

$$\mathrm{score}(t,h) = \frac{1}{|I_h|} \sum_{i_l^h \in I_h} \max_{i_k^t \in I_t} f(i_k^t, i_l^h) \quad (1)$$

where $I_t = \{i_1^t, \dots, i_n^t\}$ are the n images associated with phrase t from T and $I_h = \{i_1^h, \dots, i_n^h\}$ are the n images for phrase h from H, for $1 \leq l, k \leq n$. Note the asymmetry in Eq. 1 which captures semantic subsumptions (a picture of *river* is among the pictures of *body of water*). The function f returns the cosine similarity between two images:

$$f(i_k^t, i_l^h) = \cos(\mathbf{v}(i_k^t), \mathbf{v}(i_l^h)) = \frac{\mathbf{v}(i_k^t)\cdot\mathbf{v}(i_l^h)}{||\mathbf{v}(i_k^t)|| \cdot ||\mathbf{v}(i_l^h)||} \quad (2)$$

where $\mathbf{v}(i)$ is the vector representation of an image i. We obtain these vector representations concatenating the activations of the first 7 layers of GoogLeNet (Szegedy et al., 2015) as it is common practice (Kiela and Bottou, 2014).

Given the phrases with the highest and lowest similarity score,[3] we extract four features from each pair. The first feature is the similarity score itself. The other three features capture statistics of the relationship $f(I_t \times I_h)$ between the two sets of visual denotations I_t and I_h. This relationship $f(I_t \times I_h)$ is defined as the the matrix of image cosine similarities:

$$f(I_t \times I_h) = \begin{bmatrix} f(i_1^t, i_1^h) & f(i_1^t, i_2^h) & \cdots & f(i_1^t, i_n^h) \\ f(i_2^t, i_1^h) & f(i_2^t, i_2^h) & \cdots & f(i_2^t, i_n^h) \\ \vdots & \vdots & \ddots & \vdots \\ f(i_n^t, i_1^h) & f(i_n^t, i_2^h) & \cdots & f(i_n^t, i_n^h) \end{bmatrix} \quad (3)$$

Specifically, these three features are:

- $\max f(I_t \times I_h)$ returns the cosine similarity between the two most similar images. This feature is robust against polysemic phrases (at least one image associated to *pupil* is similar to at least one image associated to *student*) and hypernymy.
- $\mathrm{average} f(I_t \times I_h)$ returns the average similarity across all image pairs and aims to measure the visual denotation overlap between both phrases in the pair.
- $\min f(I_t \times I_h)$ returns the similarity between the two most different images and gives a notion of how different the meanings of the two phrases can be.

[2] Due to copyright, images in this paper are a subset of Google Image Search results for which we have a publishing license. Nevertheless, they are faithful representatives.

[3] If there are no candidate phrase pairs, the T-H problem is ignored. If there is only one phrase pair, such a pair is used as the pair with highest and lowest score.

All features above are concatenated into a feature vector which is paired with the T-H entailment gold label to train the classifier.

3 Experiments

Our system is independent from the logic backend but we use `ccg2lambda` (Martínez-Gómez et al., 2016)[4] for its high precision and capabilities to solve word-to-word divergences using WordNet and VerbOcean (Chklovski and Pantel, 2004).

We evaluate our system on the SemEval-2014 version of the SICK dataset (Marelli et al., 2014) with train/trial/test splits of $4,500/500/4,927$ T-H pairs and a yes/no/unk label distribution of $.29/.15/.56$. We chose SICK for its relatively limited vocabulary ($2,409$ words) and short sentences. The average T and H sentence length was 10.6 where 3.6 to 3.8 words appeared in T and not in H or vice versa. We used scipy's Random Forests (Breiman, 2001) as our entailment classifier with 500 trees and feature value standardization, trained and evaluated on those T-H pairs for which `ccg2lambda` outputs *unknown* (around 71% of the problems).

Using the tree mapping algorithm,[5] we obtained an average of 9.8 phrase pairs per T-H problem. We obtained $n = 30$ images for every phrase using Google Image Search API which we consider as our visual denotations. The images and their vector representations were obtained between Sept. 2016 and Feb. 2017 using the image miner and the feature extractor of Kiela (2016).[6]

Our main baseline is `ccg2lambda` when using only WordNet and VerbOcean to account for word-to-word lexical divergences. `ccg2lambda` is augmented with a classifier `c` that uses either text and logic features `t` or image features from 10, 20, or 30 images (`10i`, `20i` or `30i`). On the training data (Table 1), `ccg2lambda` obtains an accuracy of 82.89%. Using our classifier with all features, we carried out 10 runs of a 10-fold cross-validation on the training data and we obtained an accuracy (standard deviation) of 84.14 (0.06), 84.30 (0.14) and 84.28 (0.11) when using 10, 20 and 30 images, respectively. Thus, no significant differences in accuracy were observed for different numbers of images. When using only text and logic features (`c-t`), the accuracy dropped to 76.60 (0.03); when using only image features (`c-20i`), the accuracy dropped to 82.85%. These results show that using visual denotations to recognize phrasal entailments contributes to improvements in accuracy and that the interaction with text and logic features produces further gains.

System	Accuracy	Std.
`ccg2lambda`	82.89	–
`ccg2lambda, c-t-10i`	84.14	0.06
`ccg2lambda, c-t-20i`	**84.30**	0.14
`ccg2lambda, c-t-30i`	84.28	0.11
`ccg2lambda, c-t`	76.60	0.03
`ccg2lambda, c-20i`	82.85	0.08

Table 1: Results (accuracy and standard deviation) of the classifier `c` in a cross-validation on the training split of SICK dataset using text and logic features `t` for `10i`, `20i` and `30i` images.

System	Prec.	Rec.	Acc.
`ccg2lambda` + images	90.24	71.08	**84.29**
`ccg2lambda`, only text	96.95	62.65	83.13
L&H, text + images	–	–	82.70
L&H, only text	–	–	81.50
Illinois-LH, 2014	81.56	81.87	84.57
Yin & Schütze, 2017	–	–	87.10
Baseline (majority)	–	–	56.69

Table 2: Results on the test split of SICK dataset using precision, recall and accuracy. The system "`ccg2lambda + images`" uses text and logics features and 20 images per phrase: `c-t-20i`.

On the test data, we obtained 1.1% higher accuracy (84.29 versus 83.13) over the `ccg2lambda` baseline with a standard deviation of 0.07% over 10 runs (see Table 2) when using the setting `c-t-20i`. As a comparison, Lai and Hockenmaier (2017) obtain a similar accuracy increase when using visual denotations (1.2%) with a substantially more complex approach that requires training on the SNLI dataset (Bowman et al., 2015), a much larger corpus.

The best SemEval-2014 system obtained an accuracy of 84.57 (Lai and Hockenmaier, 2014) and other heavily engineered, finely-tuned systems (Beltagy et al., 2016; Yin and Schütze, 2017) reported up to 3% points of accuracy improvement since then. Thus, our results are still below the state of the art.

[4] https://github.com/mynlp/ccg2lambda
[5] https://github.com/pasmargo/t2t-qa
[6] https://github.com/douwekiela/mmfeat

Figure 2: True positive, ID: 4012; gold: yes.

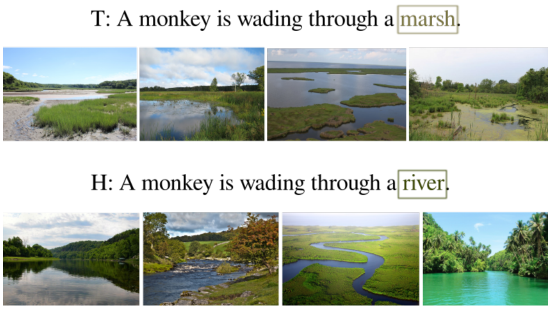

Figure 3: False positive, ID: 1215; gold: unk.

Figure 4: False negative, ID: 1318; gold: yes.

4 Error analysis

We had an average of 126 true positives (gold label yes, system label yes) and 81 false positives (gold label unk, system label yes) in our cross-validation over the training data. Figure 2 shows an example of a true positive where the tree mapping algorithm extracted the phrase pair *kangaroo that is little* and *baby kangaroo*. The image similarity features showed a high score causing the classifier to correctly produce the judgment yes. Figure 3 shows a false positive where the extracted phrase pair was *marsh* and *river* and for which the image similarity is unfortunately high. These cases are common when comparing people (*boy* and *man*) or scenery (such as *beach* and *desert*).

Figure 4 shows a false negative (gold label yes, system label unk) where the candidate phrase pair was *plastic sword* and *toy weapon*. In this case, there was only one image with a plastic sword within the images associated to *toy weapon* which may have caused the cosine similarities to be low.

5 Discussion and Conclusion

In this paper we have evaluated our method on the SICK dataset which was originally created from image captions. For that reason, the proportion of concepts with good visual denotations might be higher than in typically occurring RTE problems. Our future work is to assess the applicability of our approach into other RTE problems such as the RTE challenges, SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2017) datasets and further investigate what syntactic or semantic units can be best represented using visual denotations.

Another issue is the use of a commercial image search API as a black box to retrieve images. These search engines may include heuristics that map similar phrases or keywords into the same canonical form and that are difficult to control experimentally. However, we believe that our approach is still valid for a variety of image search mechanisms and it is generally useful to resolve lexical ambiguity at a high coverage.

We identified the conditions in which visual denotations are effective for sentence-level RTE and devised a simple scoring function to assess phrasal semantic subsumption, which may serve as the basis for more elaborated strategies. Our system is independent on the semantic parser but the entailment recognition mechanism requires a theorem prover that displays remaining sub-goals. The system and instructions are available at `https://github.com/mynlp/ccg2lambda`

Acknowledgments

This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This project is also supported by JSPS KAKENHI Grant Number 17K12747, partially funded by Microsoft Research Asia and JST CREST Grant Number JPMJCR1301, Japan. We thank Ola Vikholt and the anonymous reviewers for their valuable comments.

References

Lasha Abzianidze. 2015. A tableau prover for natural logic and language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2502, Lisbon, Portugal. Association for Computational Linguistics.

Islam Beltagy, Cuong Chau, Gemma Boleda, Dan Garrette, Katrin Erk, and Raymond Mooney. 2013. Montague meets markov: Deep semantics with probabilistic logical form. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 11–21, Atlanta, Georgia, USA. Association for Computational Linguistics.

Islam Beltagy, Stephen Roller, Pengxiang Cheng, Katrin Erk, and Raymond J. Mooney. 2016. Representing meaning with a combination of logical and distributional models. *Computational Linguistics*, 42(4):763–808.

Johan Bos, Stephen Clark, Mark Steedman, James R Curran, and Julia Hockenmaier. 2004. Wide-coverage semantic representations from a CCG parser. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proceedings of EMNLP 2004*, pages 33–40, Barcelona, Spain. Association for Computational Linguistics.

Robin Cooper, Richard Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspers, Hans Kamp, Manfred Pinkal, Massimo Poesio, Stephen Pulman, et al. 1994. FraCaS–a framework for computational semantics. *Deliverable*, D6.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing textual entailment: Models and applications*, volume 6. Morgan & Claypool Publishers.

Donald Davidson. 1967. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*. University of Pittsburgh Press.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Hamid Izadinia, Fereshteh Sadeghi, Santosh K. Divvala, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2015. Segment-phrase table for semantic segmentation, visual entailment and paraphrasing. In *The IEEE International Conference on Computer Vision (ICCV)*.

Douwe Kiela. 2016. Mmfeat: A toolkit for extracting multi-modal features. In *Proceedings of ACL-2016 System Demonstrations*, pages 55–60, Berlin, Germany. Association for Computational Linguistics.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 36–45, Doha, Qatar. Association for Computational Linguistics.

Alice Lai and Julia Hockenmaier. 2014. Illinois-LH: A denotational and distributional approach to semantics. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 329–334, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Alice Lai and Julia Hockenmaier. 2017. Learning to predict denotational probabilities for modeling entailment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 721–730, Valencia, Spain. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC2014*, pages 216–223.

Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2016. ccg2lambda: A compositional semantics system. In *Proceedings of ACL-2016 System Demonstrations*, pages 85–90, Berlin, Germany. Association for Computational Linguistics.

Pascual Martínez-Gómez and Yusuke Miyao. 2016. Rule extraction for tree-to-tree transducers by cost minimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 12–22, Austin, Texas. Association for Computational Linguistics.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. 2015. Higher-order logical inference with compositional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2061, Lisbon, Portugal. Association for Computational Linguistics.

Terence Parsons. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. The MIT Press.

Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1025–1036, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426.

Wenpeng Yin and Hinrich Schütze. 2017. Task-specific attentive pooling of phrase alignments contributes to sentence matching. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 699–709, Valencia, Spain. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Jiang Zhao, Man Lan, and Tiantian Zhu. 2014. ECNU: Expression- and message-level sentiment orientation classification in twitter using multiple effective features. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 259–264, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.