# Distributional semantics in the real world: building word vector representations from a truth-theoretic model

Elizaveta Kuzmenko & Aurélie Herbelot
University of Trento
{firstname}.{lastname}@unitn.it

**Abstract**

Distributional semantics models (DSMs) are known to produce excellent representations of word meaning, which correlate with a range of behavioural data. As lexical representations, they have been said to be fundamentally different from truth-theoretic models of semantics, where meaning is defined as a correspondence relation to the world. There are two main aspects to this difference: a) DSMs are built over corpus data which may or may not reflect 'what is in the world'; b) they are built from word co-occurrences, that is, from lexical types rather than entities and sets. In this paper, we inspect the properties of a distributional model built over a set-theoretic approximation of 'the real world'. To achieve this, we take the annotation a large database of images marked with objects, attributes and relations, convert the data into a representation akin to first-order logic and build several distributional models using various combinations of features. We evaluate those models over both relatedness and similarity datasets, demonstrating their effectiveness in standard evaluations. This allows us to conclude that, despite prior claims, truth-theoretic models are good candidates for building graded lexical representations of meaning.

## 1 Introduction

In recent years, distributional semantics models (DSMs) (Erk, 2012; Clark, 2012; Turney and Pantel, 2010) have received close attention from the linguistic community. One reason for this is that they are known to produce excellent representations of lexical meaning, which account for similarity and polysemy and correlate well with a range of behavioural data (Lenci, 2008; Mandera et al., 2017). DSMs are built on the basis of word co-occurrences in large corpora, stemming from the hypothesis that words co-occurring in similar contexts tend to share their meaning (Firth, 1957). As such, they are fundamentally different from truth-theoretic models of semantics, where meaning is defined as a correspondence relation between predicates and the world. This difference can be explicated further by noting two features of DSMs. First, they are built over corpus data which may or may not reflect 'what is in the world' (Herbelot, 2013) – and consequently does not reflect human experience gained from real world data (Andrews et al., 2009). Second, they are built from *word* co-occurrences, that is, from lexical types rather than entities and sets. In contrast, formal models account for denotation and set-theoretic aspects of language, but they are often said to lack the ability to account for lexical similarity and gradedness. This has been the basis for wanting to combine formal and distributional semantics in the past (Boleda and Herbelot, 2016): the role of DSMs, it is claimed, is to bring the lexicon to denotational approaches to meaning.

In the present paper, we build a large set-theoretic model as an approximation of "the real world", and show that quality vector representations can in fact be extracted from it. To obtain our model, we take the annotation of the Visual Genome (henceforth VG), a large database of images annotated with objects, attributes and relations (Krishna et al., 2017), and regard this data as an informative, although incomplete, description of the world. We convert the annotated data into a representation akin to some underspecified first-order logic. From this representation, we build several DSMs from various aspects of the representation and inspect the properties of the created spaces. We evaluate our models with both relatedness and similarity datasets (MEN, Bruni et al., 2012, and SimLex-999, Hill et al., 2015).

## 2 Related Work

Our work fits into attempts to bridge the gap between distributional and formal semantics. The subfield of Formal Distributional Semantics (FDS, Boleda and Herbelot, 2016) includes efforts to a) investigate the mapping from distributional models to formal semantic models (Herbelot and Vecchi, 2015; Erk, 2016; Wang et al., 2017); b) enrich formal semantics with distributional data (Garrette et al., 2011; Beltagy et al., 2013); and c) account for particular logical phenomena in vector spaces, including composition (Coecke et al., 2011; Boleda et al., 2013; Baroni et al., 2012; Bernardi et al., 2013; Asher et al., 2016 amongst many others). We also note the relevance of the work on constructing distributional spaces from syntactically or semantically parsed data (e.g. Padó and Lapata, 2007; Grefenstette et al., 2014; Hermann and Blunsom, 2013), which echoes the way we construct vector spaces from various types of predicative contexts. In contrast to those efforts, however, our data is not a standard corpus reflecting word usage but a collection of logical forms expressing true sentences with respect to a model of the world.

Most similar to our endeavour is the work by Young et al. (2014), who also take multimodal datasets as a basis to learn denotations. Their model is however created for the task of semantic inference and takes the extension of a word to be the set of situations it applies to. We introduce notions of entities and properties in our own model.

## 3 Building a truth-theoretic space

In order to build a "real world" space, we require a representation akin to a set-theoretic model. We take the annotation of the Visual Genome (VG) dataset (Krishna et al., 2017) as a proxy for such model, under the assumption that it provides a set of 'true' sentences about the world. VG contains 108,077 images associated with structured annotations. There are three types of annotation in the dataset: a) entities, or **objects** (e.g. *'window'*, *'elephant'*) – the individuals present in a given image; b) **attributes** (e.g. *'red'*, *'made of bricks'*) which describe the properties of objects; c) **relationships** (e.g. *'on'*, *'has'*, *'wearing'*) which correspond to relations between objects. The dataset also features situations, or **scene graphs**, which correspond to a single image and describe all the objects that co-occur in that image. Thus, a situation might contain a tree, a car, a woman, a dog, a sidewalk and a shade (from the tree), associated with bounding boxes. We do not use the image itself but solely the annotation data from the graph.

Every object in VG is assigned a WordNet synset and a unique id. This allows us to pre-process the data into shallow logical forms corresponding to predicate / entity pairs, ordered by situation and implicitly coordinated by a $\wedge$ within that situation. For instance, the following toy example indicates that situation 1 contains a tall brick building, identified by variable 1058505 in VG, on which we find a black sign, identified by variable 1058507. Note that the identifiers are 'real-world' variables, which pick out particular objects in the world.

$$\begin{array}{ll} S1 & building.n.01(1058508), tall(1058508), brick(1058508) \\ & sign.n.02(1058507), black(1058507), on(1058507{,}1058508) \end{array}$$

Intuitively, this representation allows us to capture all the distinct objects annotated with e.g. the synset *'building.n.01'* to generate the set of buildings (*building'*) in our universe.[1] To avoid data sparsity, we convert all relations into one-place predicates, by replacing each argument in turn with its corresponding synset. So in the example above, $on(1058507{,}1058508)$ becomes $on(1058507{,}building.n.01)$, $on(sign.n.02{,}1058508)$, which formalises that 1058507 is in the set of things that are on buildings, while 1058508 is in the set of things that signs are on.

Formally, the VG data can then be considered a set-theoretic model $M =< U{,}I >$ where $U$ is the universe (the set of all objects in the model, as identified by 'real-world' variables), and $I$ is an interpretation function mapping from a set of $n$-place predicates $P$ to $n$-tuples of objects in $U$ (with $n = 1$ given our pre-processing of relations). $P$ is the union of synsets ($Syn$), attributes ($Att$) and

[1] Note that we are not making use of the sense information provided by the synset in this work. Most words in VG are anyway used in a unique sense.

relations ($Rel$) in VG. We then build a distributional space $S =< U,P,D,F,A,C >$ where $U$ and $P$ are the universe and the predicates as above; $D$ are the dimensions of the space so that $D \subseteq P$ (that is, any combination of $Syn$, $Att$ and $Rel$); and $F$ some extraction function over our corpus of shallow logical forms $C$. $F$ is of the form $U \times D \rightarrow \{0,1\}$, i.e. it returns whether a particular dimension is predicated of an entity, giving us boolean entity vectors for all objects in VG. Finally, an aggregation function $A : (U \times D \rightarrow \{0,1\}) \rightarrow (P \times D \rightarrow \mathbb{N}_0)$ returns the final space by summing the entity vectors corresponding to each predicate in $P$: $\mathbb{N}_0$ is a natural number expressing how many times dimension $D$ is predicated of entities of type $P$. The summing operation follows the model-theoretic intuition that a predicate $p$ denotes a set which is the *union* of all things that are $p$: for instance, all dog entity vectors are summed to produce a vector for the predicate dog$'$.

In addition, we consider two ways to augment this original setup. One is by adding situational information to the mix: while relations give us a handle on what type of things a particular entity associates with via a particular predicate, this information does not include the type of things the entity simply *co-occurs* with. For instance, we may have a situation where a dog interacts with a ball (encoded by some relation *dog - chew - ball*), but VG relations do not directly tell us that the dog entity co-occurs with a park entity or a cloud entity. Another way to augment the data is by adding encyclopedic information to the VG data, which could be part of a more 'complete' model including some generalizations over the encoded sets. To do this, we extract hypernyms from WordNet (Miller et al., 1990) using the *nltk* package.[2] Only one level – the immediate parents of the concept – is taken into account. We note that hypernyms are different from the other VG features in that they don't come from natural utterances (no one would say *"domestic animal"* in place of *"dog"* in a natural context).

In what follows, we build variations of the model $M$ by counting co-occurrences between our basic entity set (aggregated into predicates with function $A$) and the following features $D \subseteq P$: attributes (*Att*), relations (*Rel*), situations (*Sit*), hypernyms (*Hyp*), and all combinations thereof.[3]

# 4 Evaluation

To measure the quality of constructed models, we evaluate them on two standard datasets: MEN and SimLex-999. The MEN dataset is supposed to capture the relatedness notion, which is defined as the relation between pairs of entities that are associated but not actually similar. SimLex-999 accounts for similarity, which is defined as the relation between words which share physical or functional features, as well as categorical information (Hill et al., 2015). Both datasets are structured in the same way: they consist of word pairs human-coded for their level of association. They respectively include 3000 (MEN) and 999 (SimLex) word pairs. To evaluate our DSMs, we follow standard practice and compute the Spearman $\rho$ correlation between the cosine similarity scores given by the model and the gold annotation. Results are shown in Table 1. To maximise comparability between different spaces and with text corpora, scores are given for raw co-occurrence matrices, and no dimensionality reduction or other optimization of the space is conducted. Note that due to the size of VG, we cannot evaluate on all pairs in the datasets. We show actual coverage in brackets next to the correlation scores.

Trends are similar both for MEN and SimLex-999. We get overall best results (highlighted in **bold**) for the models built using relations, situational information, and relations together with situations. Other models have significantly lower quality, both for single features and for their combinations. It should be noted that taking all the features together does not improve the quality of the space.

In the last column of Table 1 we report the total number of co-occurrences in each variation of the world-based model. They are included in order to make sure that we do not observe solely the effect of increasing the amount of data. Indeed, models with the greatest number of co-occurrences show medium quality, and for some combinations of features the score even decreases with more data (e.g., compare the *Hyp* and *Hyp + Sit* models, where the MEN score stays more or less the same and the SimLex score

[2] `http://www.nltk.org/`

[3] The code to pre-process the Visual Genome and the data to reproduce the experiments can be found at `https://github.com/lizaku/dsm-from-vg`.

| Setting | MEN | SimLex-999 | Num. co-occurrences |
|---|---|---|---|
| Attributes (*Att*) | 0.1801 (871) | 0.1119 (217) | 1 854 033 |
| Relations (*Rel*) | **0.5499** (847) | **0.2861** (216) | 6 481 872 |
| Situations (*Sit*) | **0.5294** (847) | **0.2480** (216) | 22 894 730 |
| Hypernyms (*Hyp*) | 0.3399 (956) | 0.2128 (244) | 1 989 576 |
| *Att + Rel* | 0.346 (871) | 0.1840 (217) | 10 720 260 |
| *Att + Sit* | 0.4492 (871) | 0.2042 (217) | 25 988 265 |
| *Rel + Sit* | **0.5326** (847) | **0.2463** (216) | 32 170 563 |
| *Att + Hyp* | 0.2385 (975) | 0.2055 (244) | 5 114 997 |
| *Rel + Hyp* | **0.5193** (956) | **0.2979** (244) | 10 878 274 |
| *Hyp + Sit* | 0.3860 (956) | 0.1731 (244) | 26 882 218 |
| *Att + Rel + Hyp* | 0.3430 (975) | 0.2367 (244) | 16 391 743 |
| *Att + Rel + Sit* | 0.4503 (871) | 0.2018 (217) | 37 652 176 |
| *Att + Sit + Hyp* | 0.3260 (975) | 0.1319 (244) | 31 252 206 |
| *Rel + Hyp + Sit* | 0.3900 (956) | 0.1760 (244) | 38 571 325 |
| *Att + Rel + Hyp + Sit* | 0.3283 (975) | 0.1337 (244) | 45 329 361 |

Table 1: Spearman $\rho$ correlation for various models on MEN and SimLex-999.

| Count-based | | Predictive (word2vec) | | Co-occurrences |
|---|---|---|---|---|
| MEN | SimLex-999 | MEN | SimLex-999 | |
| 0.081 (749) | 0.050 (462) | 0.024 (749) | 0.003 (462) | 2 000 000 |
| 0.158 (995) | 0.010 (546) | 0.043 (995) | 0.019 (546) | 5 000 000 |
| 0.225 (1226) | 0.038 (610) | 0.049 (1226) | 0.020 (610) | 15 000 000 |
| 0.226 (1455) | 0.037 (688) | 0.031 (1455) | 0.046 (688) | 30 000 000 |
| 0.253 (1554) | 0.056 (696) | 0.031 (1554) | 0.044 (696) | 40 000 000 |

Table 2: Spearman correlation on MEN and SimLex-999 datasets (Wikipedia spaces)

becomes lower). Moreover, the *Rel* model shows the highest score on a moderately small amount of data for the MEN dataset, and for the SimLex-999 dataset the score is a bit lower, whereas the *Rel + Hyp* model becomes the best (though hypernyms come from outside the model).

To compare performance of our truth-theoretic models with traditional DSMs built from text corpora, we create count-based models from the English Wikipedia using a window of $\pm$ 2 words around a target. We modulate corpus size to roughly match the number of co-occurrences extracted from VG.[4] Additionally, we train predictive models with Word2Vec (Mikolov et al., 2013) with the same number of co-occurrences as in the count-based variants. We use the same window size of 2, and the dimensionality of vectors is set to 300. The evaluation scores for different corpora sizes are shown in Table 2. We can see that, in contrast with the VG models, the score for count-based models is dependent on the amount of data provided to the DSM, and generally lower for similar numbers of co-occurrences (scores are consistent with results reported by Sahlgren and Lenci, 2016). Predictive models are simply not able to construct high-quality word representations from such amount of data.

When we try to improve the quality of our best world-based model (*Rel*) by applying normalisation, dimensionality reduction (to 300 dimensions) and PPMI weighting, we reach scores of **0.6539** on MEN (847 pairs are evaluated because not all of the pairs in the evaluation dataset are present in the VG space) and **0.3353** on SimLex-999 (216 pairs evaluated). Whilst results are not directly comparable, we nevertheless note that the MEN score is close to the figure of 0.68 reported for the inter-annotator

[4]Models are built using `https://github.com/akb89/entropix`.

correlation on the full 3000 pairs.[5] It is also only a few points lower than the best score of 0.72 obtained by Baroni et al. (2014) over 2.6B words (around 1600 times more data than in *Rel* on the basis of a $\pm 2$ word window size). The SimLex figure is also well above the figure of 0.233 reported by Hill et al. (2015) on an SVD model trained over 150M words ($\approx$ 100 times more data).

# 5 Discussion

Some interesting observations can be made with regard to the type of properties that seem to be relevant to modeling conceptual association. First, the relative results we are observing across the VG models are not artefactual of model size. Thus, a model based on situations, with 22M co-occurrences, performs worse than the model with relations, which comprises only 6M co-occurrences. This tells us that some aspects of the model-theoretic data are much more important than others and that some can even be detrimental. This finding echoes results in Emerson and Copestake (2016), which indicated that selecting particular relations from parsed data can improve performance on SimLex.

Second, the VG models outperform the standard spaces by a large margin on SimLex, even with small amounts of data. This confirms that SimLex encodes a notion of similarity that is better captured by looking at how things 'are' truth-theoretically rather than what we say about them. The fact that attributes perform badly on that dataset, however, contradicts the idea that SimLex encodes similarity of intrinsic features. Indeed, *relations* outperform any other combination of features, showing that how things associate with other things may be more important than how they intrinsically are.

Third, an additional point can be made about relations and situations. While both *Rel* and *Sit* models perform well on their own, the combined *Rel + Sit* model has lower quality (around two points are lost on MEN and four points on SimLex, compared to *Rel* alone), which means that situations take the score down. This can be explained by the fact that situations are a "noisy superset" of relations: some of the entities that co-occur in a situation will have an explicit relation associated with them (e.g., *cat* and *mouse* related by *chase'(x,y)*), while others may indeed solely co-occur (e.g., *cat* and *fork* in a scene with a pet sitting next to a dining table). So it seems that aspects of the world that entities are *actively* involved in are more important to define them than simple 'bystander' individuals.

Finally, using hypernyms improves the quality of models when evaluated on SimLex. This confirms previous results showing that using dictionaries and lexical databases helps getting better performance on SimLex (Faruqui and Dyer, 2015; Recski et al., 2016). It also indicates than when computing similarity, humans may indeed activate some 'meta-knowledge' which is not directly encoded in the basic level categories (Rosch et al., 1976) people use to describe a situation.

# 6 Conclusion

Both distributional semantics and formal semantics have their own advantages and disadvantages, but their unification provides a really powerful tool for studying the interaction between similarity and relatedness, as well as finding out which properties human tap into when making association judgments.

This paper has shown that we can study the distributional behaviour of concepts from a (large enough) truth-theoretic model. Thus, standard distributional semantics is not unique in accounting for conceptual distance. Further, the vector spaces we created have the advantages of formal models, by linking to a clear notion of entity and associated properties. Crucially, we have also demonstrated that by choosing the right properties, the truth-theoretic vector space achieves superiour performance compared to a usage-based DSM on considerably less data. While this point does not have practical application, we believe this result may have implications for understanding how humans themselves build concepts from the limited set of situations they are exposed to.

In the future, we will experiment with other image-annotated datasets or knowledge graphs to further understand which formal relations might be at the basis of human similarity judgments.

[5] See `https://staff.fnwi.uva.nl/e.bruni/MEN`.

## References


Andrews, M., G. Vigliocco, and D. Vinson (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological review 116*(3), 463.

Asher, N., T. Van de Cruys, A. Bride, and M. Abrusán (2016). Integrating type theory and distributional semantics: a case study on adjective–noun compositions. *Computational Linguistics 42*(4), 703–725.

Baroni, M., R. Bernardi, N.-Q. Do, and C.-c. Shan (2012). Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 23–32. Association for Computational Linguistics.

Baroni, M., G. Dinu, and G. Kruszewski (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pp. 238–247.

Beltagy, I., C. Chau, G. Boleda, D. Garrette, K. Erk, and R. Mooney (2013). Montague meets markov: Deep semantics with probabilistic logical form. In *Second Joint Conference on Lexical and Computational Semantics (*SEM2013)*, Atlanta, Georgia, USA, pp. 11–21.

Bernardi, R., G. Dinu, M. Marelli, and M. Baroni (2013). A relatedness benchmark to test the role of determiners in compositional distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Volume 2, pp. 53–57.

Boleda, G., M. Baroni, T. N. Pham, and L. McNally (2013). Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS2013)*, Potsdam, Germany, pp. 35–46.

Boleda, G. and A. Herbelot (2016). Formal distributional semantics: Introduction to the special issue. *Computational Linguistics 42*(4), 619–635.

Bruni, E., G. Boleda, M. Baroni, and N.-K. Tran (2012). Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 136–145. Association for Computational Linguistics.

Clark, S. (2012). Vector space models of lexical meaning. In S. Lappin and C. Fox (Eds.), *Handbook of Contemporary Semantics – second edition*. Wiley-Blackwell.

Coecke, B., M. Sadrzadeh, and S. Clark (2011). Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis: A Festschrift for Joachim Lambek 36*(1–4), 345–384.

Emerson, G. and A. Copestake (2016). Functional distributional semantics. *arXiv preprint arXiv:1606.08003*.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: a survey. *Language and Linguistics Compass 6*, 635–653.

Erk, K. (2016). What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics 9*, 17–1.

Faruqui, M. and C. Dyer (2015). Non-distributional word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL2015)*, Volume 2, pp. 464–469.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. studies in linguistic analysis. *Oxford: Philological Society. [Reprinted in Selected Papers of J.R. Firth 1952-1959, ed. Frank R. Palmer, 1968. London: Longman]*.

Garrette, D., K. Erk, and R. Mooney (2011). Integrating logical representations with probabilistic information using Markov logic. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS2011)*, pp. 105–114.

Grefenstette, E., M. Sadrzadeh, S. Clark, B. Coecke, and S. Pulman (2014). Concrete sentence spaces for compositional distributional models of meaning. In *Computing meaning*, pp. 71–86. Springer.

Herbelot, A. (2013). What is in a text, what isn't, and what this has to do with lexical semantics. In *Proceedings of the Tenth International Conference on Computational Semantics (IWCS 2013)*, Potsdam, Germany.

Herbelot, A. and E. M. Vecchi (2015). Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 22–32.

Hermann, K. M. and P. Blunsom (2013). The role of syntax in vector space models of compositional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Volume 1, pp. 894–904.

Hill, F., R. Reichart, and A. Korhonen (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics 41*(4), 665–695.

Krishna, R., Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision 123*(1), 32–73.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics 20*(1), 1–31.

Mandera, P., E. Keuleers, and M. Brysbaert (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language 92*, 57–78.

Mikolov, T., W.-t. Yih, and G. Zweig (2013). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pp. 746–751.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography 3*(4), 235–244.

Padó, S. and M. Lapata (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics 33*(2), 161–199.

Recski, G., E. Iklódi, K. Pajkossy, and A. Kornai (2016). Measuring semantic similarity of words using concept networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 193–200.

Rosch, E., C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem (1976). Basic objects in natural categories. *Cognitive psychology 8*(3), 382–439.

Sahlgren, M. and A. Lenci (2016). The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 975–980.

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

Wang, S., S. Roller, and K. Erk (2017). Distributional modeling on a diet: One-shot word learning from text only. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, Volume 1, pp. 204–213.

Young, P., A. Lai, M. Hodosh, and J. 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.