EVALution 1.0: an Evolving Semantic Dataset for Training and Evaluation of Distributional Semantic Models

Enrico Santus
The Hong Kong Polytechnic University
Hong Kong
esantus@gmail.com

Frances Yung
Nara Institute of Science and Technology
Nara, Japan
pikyufrances-y@is.naist.jp

Alessandro Lenci
Universita di Pisa
Pisa, Italy
alessandro.lenci@ling.unipi.it

Chu-Ren Huang
The Hong Kong Polytechnic University
Hong Kong
churen.huang@polyu.edu.hk

Abstract

In this paper, we introduce EVALution 1.0, a dataset designed for the training and the evaluation of Distributional Semantic Models (DSMs). This version consists of almost 7.5K tuples, instantiating several semantic relations between word pairs (including *hypernymy*, *synonymy*, *antonymy*, *meronymy*). The dataset is enriched with a large amount of additional information (i.e. relation domain, word frequency, word POS, word semantic field, etc.) that can be used for either filtering the pairs or performing an in-depth analysis of the results. The tuples were extracted from a combination of ConceptNet 5.0 and WordNet 4.0, and subsequently filtered through automatic methods and crowdsourcing in order to ensure their quality. The dataset is freely downloadable[1]. An extension in RDF format, including also scripts for data processing, is under development.

1 Introduction

Distributional Semantic Models (DSMs) represent lexical meaning in vector spaces by encoding corpora derived word co-occurrences in vectors (Sahlgren, 2006; Turney and Pantel, 2010; Lapesa and Evert, 2014). These models are based on the assumption that meaning can be inferred from the contexts in which terms occur. Such assumption is typically referred to as the *distributional hypothesis* (Harris, 1954).

DSMs are broadly used in Natural Language Processing (NLP) because they allow systems to automatically acquire lexical semantic knowledge in a fully unsupervised way and they have been proved to outperform other semantic models in a large number of tasks, such as the measurement of lexical semantic similarity and relatedness. Their geometric representation of semantic distance (Zesch and Gurevych, 2006) allows its calculation through mathematical measures, such as the *vector cosine*.

A related but more complex task is the identification of semantic relations. Words, in fact, can be similar in many ways. *Dog* and *animal* are similar because the former is a specific kind of the latter (*hyponym*), while *dog* and *cat* are similar because they are both specific kinds of *animal* (*coordinates*). DSMs do not provide by themselves a principled way to single out the items linked by a specific relation.

Several distributional approaches have tried to overcome such limitation in the last decades. Some of them use word pairs holding a specific relation as seeds, in order to discover patterns in which other pairs holding the same relation are likely to occur (Hearst, 1992; Pantel and Pennacchiotti, 2006; Cimiano and Völker, 2005; Berland and Charniak, 1999). Other approaches rely on linguistically grounded unsupervised measures, which adopt different types of distance measures by selectively weighting the vectors features (Santus et al., 2014a; Santus et al., 2014b; Lenci and Benotto, 2012; Kotlerman et al., 2010; Clarke,

[1]The resource is available at http://colinglab.humnet.unipi.it/resources/ and at https://github.com/esantus

Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015), pages 64–69,
Beijing, China, July 31, 2015.

2009; Weeds et al., 2004; Weeds and Weir, 2003). Both the abovementioned approaches need to rely on datasets containing semantic relations for training and/or evaluation.

EVALution is a dataset designed to support DSMs on both processes. This version consists of almost 7.5K tuples, instantiating several semantic relations between word pairs (including *hypernymy*, *synonymy*, *antonymy*, *meronymy*). The dataset is enriched with a large amount of additional information (i.e. relation domain, word frequency, word POS, word semantic field, etc.) that can be used for either filtering the pairs or performing an in-depth analysis of the results. The quality of the pairs is guaranteed by i.) their presence in previous resources, such as ConceptNet 5.0 (Liu and Singh, 2004) and WordNet 4.0 (Fellbaum, 1998), and ii.) a large agreement between native speakers (obtained in crowdsourcing tasks, performed with *Crowdflower*). In order to increase the homogeneity of the data and reduce its variability[2], the dataset only contains word pairs whose terms (henceforth *relata*) occur in more than one semantic relation. The additional information is provided for both *relata* and relations. Such information is based on both human judgments (e.g. relation domain, term generality, term abstractness, etc.) and on corpus data (e.g. frequency, POS, etc.).

2 Related Work

Up to now, DSMs performance has typically been evaluated against benchmarks developed for purposes other than DSMs evaluation. Except for BLESS (Baroni and Lenci, 2011), most of the adopted benchmarks include task-specific resources, such as the 80 multiple-choice synonym questions of the *Test of English as a Foreign Language* (*TOEFL*) (Landauer and Dumais, 1997), and general-purpose resources, such as WordNet (Fellbaum, 1998). None of them can be considered fully reliable for DSMs evaluation for several reasons: i.) general-purpose resources need to be inclusive and comprehensive, and therefore they either adopt broad definitions of semantic relations or leave them undefined, leading to inhomogeneous pairs; ii.) task-specific resources, on the other hand, adopt specific criteria for defining semantic relations, according to the scope of the resource (e.g. the word pairs may be more or less prototypical, according to the difficulty of the test); iii.) *relata* and relations are given without additional information, which is instead necessary for testing and analyze DSMs performance in a more detailed way (e.g. relation domain, word semantic field, word frequency, word POS, etc.).

Given its large size, in terms both of lexical items and coded relations, WordNet is potentially extremely relevant to evaluate DSMs. However, since it has been built by lexicographers without checking against human judgments, WordNet is not fully reliable as a gold standard. Moreover, the resource is also full with inconsistencies in the way semantic relations have been encoded. Simply looking at the hypernymy relation (Cruse, 1986), for example, we can see that it is used in both a taxonomical (i.e. *dog* is a hyponym of *animal*) and a vague and debatable way (i.e. *silly* is a hyponym of *child*). ConceptNet (Liu and Singh, 2004) may be considered even less homogeneous, given its size and the automatic way in which it was developed.

Landauer and Dumais (1997) introduces the 80 multiple-choice synonym questions of the *TOEFL* as a benchmark in the synonyms identification task. Although good results in such set (Rapp, 2003) may have a strong impact on the audience, its small size and the fact that it contains only synonyms cannot make it an accurate benchmark to evaluate DSMs.

For what concerns antonymy, based on similar principles to the *TOEFL*, Mohammed et al. (2008) proposes a dataset containing 950 closest-opposite questions, where five alternatives are provided for every target word. Their data are collected starting from 162 questions in the Graduate Record Examination (*GRE*).

BLESS (Baroni and Lenci, 2011) contains several relations, such as hypernymy, co-hyponymy, meronymy, event, attribute, etc. This dataset covers 200 concrete and unambiguous concepts divided in 17 categories (e.g. vehicle, ground mammal, etc.). Every concept is linked through the various semantic relations to several *relata* (which can be either nouns, adjectives or verbs). Unfortunately this dataset does not contain synonymy and antonymy related pairs.

With respect to entailment, Baroni et al.(2012)

[2]Reducing the variability should impact both on training and evaluation. In the former case, because it should help in identifying consistent patterns and discriminate them from the inconsistent ones. In the latter case, because it should allow meaningful comparisons of the results.

have built a dataset containing 1,385 positive (e.g. house-building) and negative (e.g. leader-rider) examples: the former are obtain by selecting particular hypernyms from WordNet, while the latter are obtained by randomly shuffling the hypernyms of the positive examples. The pairs are then manually double-checked.

Another resource for similarity is WordSim 353 (Finkelstein et al., 2002; Baroni and Lenci, 2011), which is built by asking subjects to rate the similarity in a set of 353 word pairs. While refining such dataset, Agirre (2009) found that several types of similarity are involved (i.e. he can recognize, among the others, hypernyms, coordinates, meronyms and topically related pairs).

Recently, Santus et al. (2014c; 2014b) use a subset of 2,232 English word pairs collected by Lenci/Benotto in 2012/13 through Amazon Mechanical Turk, following the method described by Scheible and Schulte im Walde (2014). Targets are balanced across word categories. Frequency and degree of ambiguity are also taken into consideration. The dataset includes hypernymy, antonymy and synonymy for nouns, adjectives and verbs.

The constant need for new resources has recently led Gheorghita and Pierrel (2012) to suggest an automatic method to build a hypernym dataset by extracting hypernyms from definitions in dictionaries. A precision of 72.35% is reported for their algorithm.

3 Design, Method and Statistics

As noted by Hendrickx et al. (2009), an ideal dataset for semantic relations should be exhaustive and mutually exclusive. That is, every word pair should be related by one, and only one, semantic relation. Unfortunately, such ideal case is very far from reality. Relations are ambiguous, hard to define and generally context-dependent (e.g. *hot* and *warm* may either be synonyms or antonyms, depending on the context).

EVALution is designed to reduce such issues by providing i.) consistent data, ii.) prototypical pairs and iii.) additional information. The first requirement is achieved by selecting only word pairs whose *relata* occur (independently) in more than one semantic relation, so that the variability in the data is drastically reduced. This should both improve the training process (being *relata* in more relations, the pairs can be used not only to find new patterns, but also to discriminate the ambiguous patterns from the safe ones) and the evaluation (allowing significant comparisons among the results). The second requirement is achieved by selecting only the pairs that obtain a large agreement between native speakers (judgments are collected in crowdsourcing tasks, performed with *Crowdflower*). Finally, the third requirement is achieved by providing additional information obtained through both human judgments (e.g. relation domain, term generality, term abstractness, etc.) and corpus-based analysis (e.g. frequency, POS, etc.).

3.1 Methodology

EVALution 1.0 is the result of a combination and filtering of ConceptNet 5.0 (Liu and Singh, 2004) and WordNet 4.0 (Fellbaum, 1998). Two kinds of filtering are applied: automatic filters and native speakers judgments. Automatic filtering is mainly intended to remove tuples including: i.) non-alphabetical terms; ii.) relations that are not relevant (see Table 1[3]); iii.) pairs that already appear in inverted order; iv.) pairs whose *relata* did not appear in at least 3 relations; v.) pairs that are already present in the BLESS and in the Lenci/Benotto datasets.

Relation	Pairs	*Relata*	Sentence template
IsA (hypernym)	1880	1296	X is a kind of Y
Antonym	1600	1144	X can be used as the opposite of Y
Synonym	1086	1019	X can be used with the same meaning of Y
Meronym	1003	978	X is ...
- PartOf	654	599	... part of Y
- MemberOf	32	52	... member of Y
- MadeOf	317	327	...made of Y
Entailment	82	132	If X is true, than also Y is true
HasA (possession)	544	460	X can have or can contain Y
HasProperty (attribute)	1297	770	Y is to specify X

Table 1: Relations, number of pairs, number of *relata* and sentence templates

Native speakers judgments are then collected

[3]For the definition of the semantic relations, visit: https://github.com/commonsense/conceptnet5/wiki/Relations

for the about 13K automatically filtered pairs. We create a task in *Crowdflower*, asking subjects to rate from 1 (Strongly disagree) to 5 (Strongly agree) the truth of sentences containing the target word pairs (e.g. dog *is a kind of* animal). We collect 5 judgments per sentence. Only pairs that obtain at least 3 positive judgments are included in the dataset. Table 1 summarizes the number of pairs per relation that passed this threshold and provides the sentence templates used to collect the judgments.

For the selected pairs and their *relata*, we perform two more crowdsourcing tasks, asking subjects to tag respectively the contexts/domains in which the sentences are true and the categories of the *relata*. Subjects are allowed to select one or more tags for each instance. For every *relatum*, we collect tags from 2 subjects, while for every pair we collect tags from 5 subjects. Table 2 contains the set of available tags for both relations and *relata*, and their distribution (only tags that were selected at least twice are reported).

3.2 Statistics

The dataset contains 7,429 word pairs, involving 1,829 *relata* (63 of which are multiword expressions). On average, every *relatum* occurs in 3.2 relations and every relation counts 644 *relata* (see Table 1).

For every *relatum*, the dataset contains four types of corpus-based metadata, including lemma frequency, POS distribution, inflection distribution and capitalization distribution. Such data is extracted from a combination of ukWaC and WaCkypedia (Santus et al., 2014a). Finally, for every relation and *relata*, descriptive tags collected through the crowdsourcing task described above are provided together with the number of subjects that have choosen them out of the total number of annotators. Table 2 describes the distribution of the tags.

4 Evaluation

In order to further evaluate the dataset, we built a 30K dimensions standard window-based matrix, recording co-occurrences with the nearest 2 content words to the left and the right of the target. Co-occurrences are extracted from a combination of the freely available ukWaC and WaCkypedia corpora (Santus et al., 2014a) and weighted with Local Mutual Information (LMI). We then calculate the *vector cosine* values for all the pairs in EVALution and for all those in BLESS (for comparison). Figure 1 shows the box-plots summarizing their distribution per relation.

Relation tag	**Distr.**	***Relata* tags**	**Distr.**
Event	2711	Basic/	382
		Subordinate/	163
		Superordinate	186
Time	266	General	565
		Specific	221
Space	962	Abstract/	430
		Concrete	531
Object	3011	Event	225
Nature	2372	Time	20
Culture	861	Space	115
Emotion	1005	Object	223
Relationship	1552	Animal	52
Communication	567	Plant	23
Food	404	Food	52
Color	269	Color	20
Business	245	People	100

Table 2: The distribution of tags for relations and *relata* (only tags that were selected at least twice are reported). Every relation and *relatum* can have more than one tag.

4.1 Discussion

As shown in Figure 1, the *vector cosine* values are higher for antonymy, possession (*HasA*), hypernymy (*IsA*), member-of, part-of and synonymy. This result is quite expected for synonyms, antonyms and hypernyms (Santus et al., 2014a; Santus et al., 2014b) and it is not surprising for member-of (e.g. star *MemberOf* constellation), part-of (e.g. word *PartOf* phrase) and possession (e.g. arm *HasA* hand). The *vector cosine* values are instead lower for entailment, attribute (*HasProperty*) and made-of, which generally involve *relata* that are semantically more distant.

In general, we can say that the variance between the distributions of *vector cosine* values per relation is low. This is however very similar to what happens with BLESS, where only coordinate and random pairs are significantly different, demonstrating once more that the *vector cosine* is not sufficient to discriminate semantic relations.

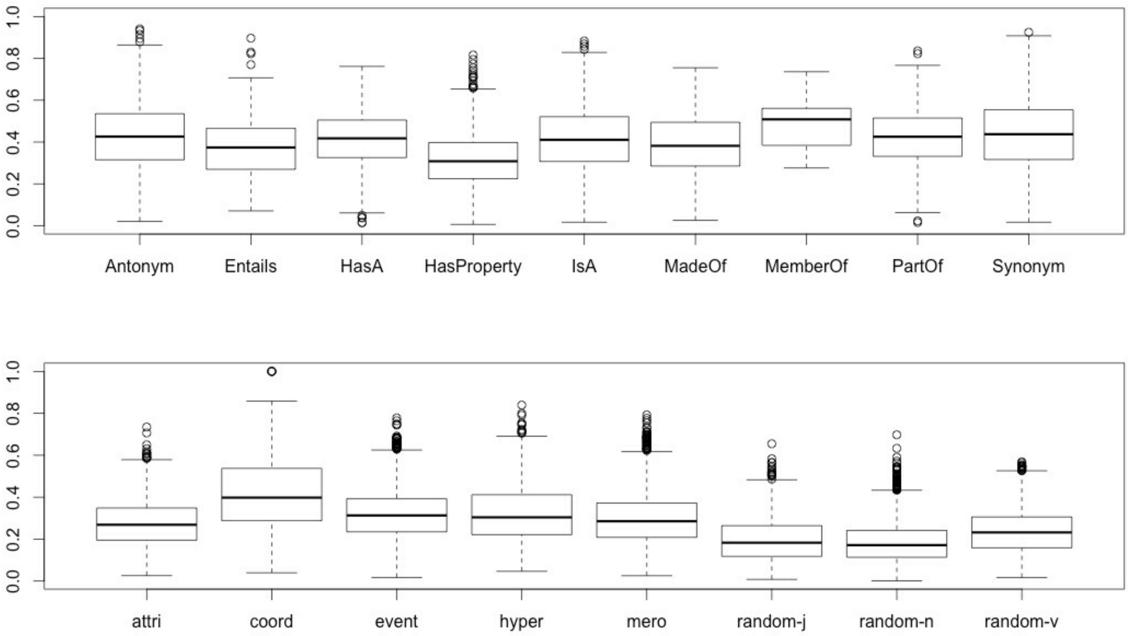

Figure 1: Distribution of vector cosine values in EVALution (above) and BLESS (below)

5 Conclusion and Future Work

EVALution is designed as an evolving dataset including tuples representing semantic relations between word pairs. Compared to previous resources, it is characterized by i.) internal consistency (i.e. few terms occurring in more relationships); ii.) prototypical pairs (i.e. high native speakers agreement, collected through crowdsourcing judgments); iii.) a large amount of additional information that can be used for further data filtering and analysis. Finally, it is freely available online at http://colinglab.humnet.unipi.it/resources/ and at https://github.com/esantus.

Further work is aiming to improve and extend the resource. This would require further quality-checks on data and metadata, the addition of new pairs and extra information, and the adoption of a format (such as RDF) that would turn our dataset into an interoperable linked open data. We are currently considering the *LEMON* model, which was previously used to encode BabelNet 2.0 (Ehrmann et al., 2014) and WordNet (McCrae et al., 2014). Some scripts will also be added for helping analyzing DSMs performance.

Acknowledgement

This work is partially supported by HK PhD Fellowship Scheme under PF12-13656.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung chieh Shan. 2012. Entailment above the word level in distributional semantics. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

Mathew Berland and Eugene Charniak. 1999. Finding parts in very large corpora. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Philipp Cimiano and Johanna Völker. 2005. text2onto. *Natural language processing and information systems*.

Daoud Clarke. 2009. Context-theoretic semantics for natural language: An overview. *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*.

D. Alan Cruse. 1986. *Lexical semantics*. Cambridge University Press.

Maud Ehrmann, Francesca Cecconi, Daniele Vannella, John P. McCrae, Philipp Cimiano, and Roberto Navigli. 2014. A multilingual semantic network as linked data: lemon-babelnet. *Proceedings of the Workshop on Linked Data in Linguistics*.

Christiane Fellbaum. 1998. *Wordnet*. Wiley Online Library.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems,*.

Inga Gheorghita and Jean-Marie Pierrel. 2012. Towards a methodology for automatic identification of hypernyms in the definitions of large-scale dictionary. *Proceedings of the International Conference on Language Resources and Evaluation*.

Zellig S. Harris. 1954. Distributional structure. *Word*.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. *Proceedings of the International Conference on Computational Linguistics*.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Multi-way classification of semantic relations between pairs of nominals. *Proceedings of the Workshop on Semantic Evaluations*.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*.

Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. *Proceedings of the First Joint Conference on Lexical and Computational Semantics*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT Technology Journal*.

John P. McCrae, Christiane D. Fellbaum, and Philipp Cimiano. 2014. Publishing and linking wordnet using lemon and rdf. *Proceedings of the Workshop on Linked Data in Linguistics*.

Saif Mohammad, Bonnie Dorr, and Graeme Hirst. 2008. Computing word-pair antonymy. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. *Proceedings of the International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics*.

Reinhard Rapp. 2003. Word sense discovery based on sense descriptor dissimilarity. *Proceedings of Machine Translation Summit*.

Magnus Sahlgren. 2006. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte Im Walde. 2014a. Chasing hypernyms in vector spaces with entropy. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

Enrico Santus, Qin Lu, Alessandro Lenci, and Chu-Ren Huang. 2014b. Taking antonymy mask off in vector space. *Proceedings of the Pacific Asia Conference on Language, Information and Computing*.

Enrico Santus, Qin Lu, Alessandro Lenci, and Chu-Ren Huang. 2014c. Unsupervised antonym-synonym discrimination in vector space. *Proceedings of the Italian Conference on Computational Linguistics*.

Silke Scheible and Sabine Schulte Im Walde. 2014. A database of paradigmatic semantic relation pairs for german nouns, verbs, and adjectives. *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Julie Weeds, David Weir, and Dianna McCarthy. 2004. Characterising measures of lexical distributional similarity. *Proceedings of the International Conference on Computational Linguistics*.

Torsten Zesch and Iryna Gurevych. 2006. Automatically creating datasets for measures of semantic relatedness. *Proceedings of the Workshop on Linguistic Distances*.