How we BLESSed distributional semantic evaluation

Marco Baroni
University of Trento
Trento, Italy
marco.baroni@unitn.it

Alessandro Lenci
University of Pisa
Pisa, Italy
alessandro.lenci@ling.unipi.it

Abstract

We introduce BLESS, a data set specifically designed for the evaluation of distributional semantic models. BLESS contains a set of tuples instantiating different, explicitly typed semantic relations, plus a number of controlled random tuples. It is thus possible to assess the ability of a model to detect truly related word pairs, as well as to perform in-depth analyses of the types of semantic relations that a model favors. We discuss the motivations for BLESS, describe its construction and structure, and present examples of its usage in the evaluation of distributional semantic models.

1 Introduction

In NLP, it is customary to distinguish between *intrinsic evaluations*, testing a system in itself, and *extrinsic evaluations*, measuring its performance in some task or application (Sparck Jones and Galliers, 1996). For instance, the intrinsic evaluation of a dependency parser will measure its accuracy in identifying specific syntactic relations, while its extrinsic evaluation will focus on the impact of the parser on tasks such as question answering or machine translation. Current approaches to the evaluation of **Distributional Semantic Models** (DSMs, also known as semantic spaces, vector-space models, etc.; see Turney and Pantel (2010) for a survey) are task-oriented. Model performance is evaluated in "semantic tasks", such as detecting synonyms, recognizing analogies, modeling verb selectional preferences, ranking paraphrases, etc. Measuring the performance of DSMs on such tasks represents an *indirect test* of their ability to capture lexical meaning. The task-oriented benchmarks adopted in distributional semantics have not specifically been designed to evaluate DSMs. For instance, the widely used TOEFL synonym detection task was designed to test the learners' proficiency in English as a second language, and not to investigate the structure of their semantic representations (cf. Section 2).

To gain a real insight into the abilities of DSMs to address lexical semantics, existing benchmarks must be complemented with a more intrinsically oriented approach, to perform *direct tests* on the specific aspects of lexical knowledge captured by the models. In order to achieve this goal, three conditions must be met: (i) to single out the particular aspects of meaning that we want to focus on in the evaluation of DSMs; (ii) to design a data set that is able to explicitly and reliably encode the target semantic information; (iii) to specify the evaluation criteria of the system performance on the data set, in order to get an estimate of the intrinsic ability of DSMs to cope with the selected semantic aspects. In this paper, we address these three conditions by presenting **BLESS** (**B**aroni and **L**enci **E**valuation of **S**emantic **S**paces), a new data set specifically geared towards the intrinsic evaluation of DSMs, downloadable from: http://clic.cimec.unitn.it/distsem.

2 Distributional semantics benchmarks

There are several benchmarks that have been widely adopted for the evaluation of DSMs, all of them capturing interesting challenges a DSM should meet. We briefly review here some commonly used and representative benchmarks, and discuss why we felt

Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011, pages 1–10,
Edinburgh, Scotland, UK, July 31, 2011.

the need to add BLESS to the set. We notice at the outset of this discussion that we want to carve out a space for BLESS, and not to detract from the importance and usefulness of other data sets. We further remark that we focus on data sets that, like BLESS, are monolingual English and, while task-oriented, not aimed at a specific application setting (such as machine translation or ontology population).

Probably the most commonly used benchmark in distributional semantics is the TOEFL **synonym detection task** introduced to computational linguistics by Landauer and Dumais (1997). It consists of 80 multiple-choice questions, each made of a target word (a noun, verb, adjective or adverb) and 4 response words, 1 of them a synonym of the target. For example, given the target *levied*, the matched words are *imposed*, *believed*, *requested*, *correlated*, the first one being the correct choice. The task for a system is then to pick the true synonym among the responses. The TOEFL task focuses on a single semantic relation, namely synonymy. Synonymy is actually not a common semantic relation and one of the hardest to define, to the point that many lexical semanticists have concluded that true synonymy does not exist (Cruse, 1986). Just looking at a few examples of synonym pairs from the TOEFL set will illustrate the problem: *discrepancy/difference*, *prolific/productive*, *percentage/proportion*, *to market/to sell*, *color/hue*. Moreover, the criteria adopted to choose the distractors (probably motivated by the language proficiency testing purposes of TOEFL) are not known. By looking at the set, it is hard to discern a coherent pattern. In certain cases, the distractors are semantically close to the target word (*volume*, *sample* and *profit* for *percentage*), whereas in other cases they are not (*home*, *trail*, and *song* for *annals*). It it thus not clear whether we are asking the models to distinguish a semantically related word (the synonym) from random elements, or a more tightly related word (the synonym, again) from other related words. The TOEFL task, finally, is based on a discrete choice (either you get the right word, or you don't), with the result that evaluation is "quantized", leading to large accuracy gains for small actual differences (one model that guesses one more synonym right than another gets 1.25% more points in percentage accuracy).

The WordSim 353 data set (Finkelstein et al., 2002) is a widely used example of **semantic similarity rating** set (see also Rubenstein and Goodenough (1965) and Miller and Charles (1991)). Subjects were asked to rate a set of 353 word pairs on a "similarity" scale and average ratings for each pair were computed. Models are then evaluated in terms of correlation of their similarity scores with average ratings across pairs. From the point of view of assessing the performance of a DSM, the WordSim (and related) similarity ratings are a mixed bag, in two senses. First, the data set contains a variety of different semantic relations. In a recent semantic annotation of the WordSim performed by Agirre et al. (2009) we find that, among the 174 pairs with above-median score (and thus presumably related), there is 1 identical pair, 17 synonym pairs, 28 hyper-/hyponym pairs, 30 coordinate pairs, 6 holo-/meronym pairs and 92 (more than half) pairs that are "topically related, but none of the above". Second, the scores are a mixture of intuitions about which of these relations are more semantically tight and intuitions about more or less connected pairs *within* each of the relations. For example, among the top-rated scores we find synonyms such as *journey/voyage* and coordinate concepts (*king/queen*). If we look at the relations characterizing pairs around the median rating, we find both less "perfect" synonyms (*monk/brother*, that are synonymous only under an unusual sense of *brother*) and less close coordinates (*skin/eye*), as well as pairs instantiating other, less taxonomically tight relations, such as many syntagmatically connected items (*family/planning*, *disaster/area*, *bread/butter*). Apparently, a single scale is merging intuitions about semantic similarity of specific pairs and semantic similarity of different relations.

A perhaps more principled way to evaluate DSMs that has recently gained some popularity is the **concept categorization task**, where a DSM has to cluster a set of nouns expressing basic-level concepts into gold standard categories. A particularly carefully constructed example is the Almuhareb-Poesio (AP) set of 402 concepts introduced in Almuhareb (2006). Concept categorization sets also include the Battig (Baroni et al., 2010) and ESSLLI 2008 (Baroni et al., 2008) lists. The AP concepts must be clustered into 21 classes, each represented by between 13 and 21 nouns. Examples include the *ve-*

hicle class (*helicopter*, *motorcycle*...), the *motivation* class (*ethics*, *incitement*, ...), and the *social unit* class (*platoon*, *branch*). The concepts are balanced in terms of frequency and ambiguity, so that, e.g., the *tree* class contains a common concept such as *pine* but also the *casuarina* tree, as well as the *samba* tree, that is not only an ambiguous term, but one where the non-arboreal sense dominates.

Concept categorization data sets, while interesting to simulate one of the basic aspects of human cognition, are limited to one kind of semantic relation (discovering coordinates). More importantly, the quality of the results will depend not only on the underlying DSMs, but also on the clustering algorithm being used (and on how this interacts with the overall structure of the DSM), thus making it hard to interpret the performance of DSMs. The forced "hard" category choice is also problematic, and exaggerates performance differences between models especially in the presence of ambiguous terms (a model that puts *samba* in the *occasion* class with *dance* and *ball* might be penalized as much as a model that puts it in the *monetary currency* class).

A more general issue with all benchmarks is that tasks are based on comparing a single quality score for each considered model (accuracy for TOEFL, correlation for WordSim, a clustering quality measure for AP, etc.). This gives little insight into *how* and *why* the models differ. Moreover, there is no well-established statistical procedure to assess significance of differences for most commonly used measures. Finally, either because the data sets were not originally intended as standard benchmarks, or even on purpose, they all are likely to cause coverage problems even for DSMs trained on very large corpora. Think of the presence of extremely rare nouns like *casuarina* in AP, of proper nouns in WordSim (it is not clear to us that DSMs are adequate semantic models for referring expressions – at the very least they should not be mixed up lightly with common nouns), or multi-word expressions in other data sets.

3 How we intend to BLESS distributional semantic evaluation

DSMs measure the distributional similarity between words, under the assumption that proximity in distributional space models semantic relatedness, including, as a special case, semantic similarity (Budanitsky and Hirst, 2006). However, semantically related words in turn differ for the type of relation holding between them: e.g., *dog* is strongly related to both *animal* and *tail*, but with different types of relations. Therefore, evaluating the intrinsic ability of DSMs to represent the semantic space of a word entails both (i) determining to what extent words close in semantic space are actually semantically related, and (ii) analyzing, among related words, which type of semantic relation they tend to instantiate. Two models can be equally very good in identifying semantically related words, while greatly differing for the type of related pairs they favor.

The BLESS data set complies with both these constraints. The set is populated with tuples expressing a **relation** between a target concept (henceforth referred to as **concept**) and a relatum concept (henceforth referred to as **relatum**). For instance, in the BLESS tuple *coyote-hyper-animal*, the concept *coyote* is linked to the relatum *animal* via the hypernymy relation (the relatum is a hypernym of the concept). BLESS focuses on a coherent set of basic-level nominal concrete concepts and a small but explicit set of semantic relations, each instantiated by multiple relata. Depending on the type of relation, relata can be nouns, verbs or adjectives. Moreover, BLESS also contains, for each concept, a number of *random* "relatum" words that are not semantically related to the concept. Thus, it also allows to evaluate a model in terms of its ability to harvest related words given a concept (by comparing true and random relata), and to identify specific types of relata, both in terms of semantic relation and part of speech.

A data set intending to represent a gold standard for evaluation should include tests items that are as little controversial as possible. The choice of restricting BLESS to concrete concepts is motivated by the fact that they are by far the most studied ones, and there is better agreement about the relations that characterize them (Murphy, 2002; Rogers and McClelland, 2004).

As for the types of relation to include, we are faced with a dilemma. On the one hand, there is wide evidence that taxonomic relations, the best understood type, only represent a tiny portion of the rich spectrum covered by semantic relatedness. On the other hand, most of these wider semantic rela-

tions are also highly controversial, and may easily lead to questionable classifications. For instance, concepts are related to events, but often it is not clear how to distinguish the events expressing a typical function of nominal concepts (e.g., *car* and *transport*), from those events that are also strongly related to them but without representing their typical function *sensu stricto* (e.g., *car* and *fix*). As will be shown in Section 4, the BLESS data set tries to overcome this dilemma by attempting a difficult compromise: Semantic relations are not limited to taxonomic types and also include attributes and events strongly related to a concept, but in these cases we have resorted to underspecification, rather than committing ourselves to questionable granular relations.

BLESS strives to capture those differences and similarities among DSMs that do not depend on coverage, processing choices or lexical preferences. BLESS has been constructed using a publicly available collection of corpora for reference (see Section 4.4 below), which means that anybody can train a DSM on the same data and be sure to have perfect coverage (but this is not strictly necessary). For each concept and relation, we pick a variety of relata (see next section) in order to abstract away from incidental gaps of models or different lexical/topical preferences. For example, the concept *robin* has 7 hypernyms including the very general and non-technical *animal* and *bird* and the more specific and technical *passerine*. A model more geared toward technical terminology might assign a high similarity score to the latter, whereas a commonsense-knowledge-oriented DSM might pick *bird*. Both models have captured similarity with a hypernym, and we have no reason, in general semantic terms, to penalize one or the other. To maximize coverage, we also make sure that, for each concept and relation, a reasonable number of relata are frequently attested in our reference corpora (see statistics below), we only include single-word relata and, where appropriate, we include multiple forms for the same relatum (both *sock* and *socks* as coordinates of *scarf* – as discussed in Section 4.1, we avoided similar ambiguous items as target concepts).

Currently, distributional models for attributional similarity and relational similarity (Turney, 2006) are tested on different data sets, e.g., TOEFL and SAT respectively (briefly, attributional similarity pertains to similarity between a pair of concepts in terms of shared properties, whereas relational similarity measures the similarity of the relations instantiated by couples of concept *pairs*). Conversely, BLESS is not biased towards any particular type of semantic similarity and thus allows both families of models to be evaluated on the same data set. Given a concept, we can analyze the types of relata that are selected by a model as more attributionally similar to the target. Alternatively, given a concept-relatum pair instantiating a specific semantic relation (e.g., hypernymy) we can evaluate a model ability to identify analogically similar pairs, i.e., others concept-relatum pairs instantiating the same relation (we do not illustrate this possibility here).

Finally, by collecting distributions of 200 similarity values for each relation, BLESS allows reliable statistical testing of the significance of differences in similarity within a DSM (for example, using the procedure we present in Section 5 below), as well as across DSMs (for example, via a linear/ANOVA model with relations and DSMs as factors – not illustrated here).

4 Construction

4.1 Concepts

BLESS includes 200 distinct English concrete nouns as target concepts, equally divided between living and non-living entities. Concepts have been grouped into 17 broader classes: AMPHIBIAN_REPTILE (including amphibians and reptiles: *alligator*), APPLIANCE (*toaster*), BIRD (*crow*), BUILDING (*cottage*), CLOTHING (*sweater*), CONTAINER (*bottle*), FRUIT (*banana*), FURNITURE (*chair*), GROUND_MAMMAL (*beaver*), INSECT (*cockroach*), MUSICAL_INSTRUMENT (*violin*), TOOL (i.e., manipulable tools or devices: *hammer*), TREE (*birch*), VEGETABLE (*cabbage*), VEHICLE (*bus*), WATER_ANIMAL (including fish and sea mammals: *herring*), WEAPON (*dagger*).

All 200 BLESS concepts are single-word nouns in the singular form (we avoided concepts such as *socks* whose surface form might change depending on lemmatization choices). The major source we used to select the concepts were the McRae Norms (McRae et al., 2005), a collection of living and non-living basic-level concepts described by 725 sub-

jects with semantic features, each tagged with its property type. As further constraints guiding our selection, we wanted concepts with a reasonably high frequency (cf. Section 4.4), we avoided ambiguous or highly polysemous concepts and we balanced inter- and intra-class composition. Classes include both prototypical and atypical instances (e.g., *robin* and *penguin* for BIRD), and have a wide spectrum of internal variation (e.g., the class VEHICLE contains wheeled, air and sea vehicles). 175 BLESS concepts are attested in the McRae Norms, while the remnants were selected by the authors according to the above constraints. The average number of concepts per class is 11.76 (median 11; min. 5 AMPHIBIAN_REPTILE; max. 21 GROUND_MAMMAL).

4.2 Relations

For each concept noun, BLESS includes several relatum words, linked to the concept by one of the following 5 relations. COORD: the relatum is a noun that is a co-hyponym (coordinate) of the concept, i.e., they belong to the same (narrowly or broadly defined) semantic class: *alligator-coord-lizard*; HYPER: the relatum is a noun that is a hypernym of the concept: *alligator-hyper-animal*; MERO: the relatum is a noun referring to a part/component/organ/member of the concept, or something that the concept contains or is made of: *alligator-mero-mouth*; ATTRI: the relatum is an adjective expressing an attribute of the concept: *alligator-attri-aquatic*; EVENT: the relatum is a verb referring to an action/activity/happening/event the concept is involved in or is performed by/with the concept: *alligator-event-swim*. BLESS also includes the relations RAN.N, RAN.J and RAN.V, which relate the target concepts to control tuples with random noun, adjective and verb relata, respectively.

The BLESS relations cover a wide spectrum of information useful to describe a target concept and to qualify the notion of semantic relatedness: taxonomically related entities (*hyper* and *coord*), typical attributes (*attri*), components (*mero*), and associated events (*event*). However, except for *hyper* and *coord* (corresponding to the standard relations of class inclusion and co-hyponymy respectively), the other BLESS relations are highly underspecified. For instance, *mero* corresponds to a very broad notion of meronymy, including not only parts (*dog-tail*), but also the material (*table-wood*) as well as the members (*hospital-patient*) of the entity the target concept refers to (Winston et al., 1987); *event* is used to represent the behaviors of animals (*dog-bark*), typical functions of instruments (*violin-play*), and events that are simply associated with the target concept (*car-park*); *attri* captures a large range of attributes, from physical (*elephant-big*) to evaluative ones (*car-expensive*). As we said in section 3, we did not attempt to further specify these relations to avoid any commitment to controversial ontologies of property types. Note that we exclude synonymy both because of the inherent problems in this very notion (Cruse, 1986), and because it is impossible to find convincing synonyms for 200 concrete concepts.

In BLESS, we have adopted the simplifying assumption that each relation type has relata belonging to the same part of speech: nouns for *hyper*, *coord* and *mero*, verbs for *event*, and adjectives for *attri*. Therefore, we abstract away from the fact that the same semantic relation can be realized with different parts of speech, e.g., a related event can be expressed by a verb (*transport*) or by a noun (*transportation*).

4.3 Relata

The relata of the non-random relations are English nouns, verbs and adjectives selected and validated by both authors using two types of sources: *semantic sources* (the McRae Norms (McRae et al., 2005), WordNet (Fellbaum, 1998) and ConceptNet (Liu and Singh, 2004)) and *text sources* (Wikipedia and the Web-derived ukWaC corpus, see Section 4.4 below). These resources greatly differ in dimension, origin and content and therefore provide complementary views on relata. Their relative contribution to BLESS also depends on the type of relation and the target concept. For instance, the rich taxonomic structure of WordNet has been the main source of information for many technical hypernyms (e.g. *gymnosperm*, *oscine*), which instead are missing from more commonsense-oriented resources such as the McRae Norms and ConceptNet. Meronyms are rarer in WordNet, and were collected mainly from the latter two resources, with many technical terms (e.g., parts of ships, weapons) harvested from the Wikipedia entries for the target concepts.

Attributes and events were collected from McRae

Norms, ConceptNet and ukWaC. In the McRae Norms, the number of features per concept is fairly limited, but they correspond to highly distinctive, prototypical and cognitively salient properties. ConceptNet instead provides a much wider array of associated events and attributes that are part of our commonsense knowledge about the target concepts (e.g., the events *park*, *steal* and *break*, etc. for *car*). ConceptNet relations such as *Created_by*, *Used_for*, *Capable_of* etc. have been analyzed to identify potential event relata, while the *Has_property* relation has been inspected to look for attributes. The most salient adjectival and verbal collocates of the target nouns in the ukWaC corpus were also used to identify associated attributes and events. For instance, the target concept *elephant* is not attested in the McRae Norms and has few properties in ConceptNet. Thus, many of its related events have been harvested from ukWaC. They include verbs such as *hunt*, *kill*, etc. which are quite salient and frequent with respect to elephants, although they can hardly be defined as prototypical properties of this animal. As a result of the combined use of such different types of sources, the BLESS relata are representative of a wide spectrum of semantic information about the target concepts: they include domain-specific terms side by side to commonsense ones, very distinctive features of a concept (e.g., *hoot* for *owl*) together with attributes and events that are instead shared by a whole class of concepts (e.g., all animals have relata such as *eat*, *feed*, and *live*), prototypical features as well as events and attributes that are statistically salient for the target, etc.

In many cases, the concept properties contained in semantic sources are expressed with phrases, e.g., *lay eggs*, *eat grass*, *live in Africa*, etc. We decided, however, to keep only single-word relata in BLESS, because DSMs are typically populated with single words, and, when they are not, they differ in the kinds of multi-word elements they store. Therefore, phrasal relata have always been reduced to their head: a verb for properties expressed by a verb phrase, and a noun for properties expressed by a noun phrase. For instance, from the property *lay eggs*, we derived the event relatum *lay*.

To extract the random relata, we adopted the following procedure. For each relatum that instantiates a true relation with the concept, we also randomly picked from our combined corpus (cf. Section 4.4) another lemma with the same part of speech, and frequency within 1 absolute logarithmic unit from the frequency of the corresponding true relatum. Since picking a random term does not guarantee that it will not be related to the concept, we filtered the extracted list by crowdsourcing, using the Amazon Mechanical Turk via the CrowdFlower interface (CF).[1] We presented CF workers with the list of about 15K concept+random-term pairs selected with the procedure we just described, plus a manually checked validation set (a "gold set" in CF terminology) comprised of 500 concept+true-relatum pairs and 500 concept+random-term pairs (these elements are used by CF to determine the reliability of workers, and discard the ratings of unreliable ones), plus a further set of 1.5K manually checked concept+true-relatum pairs to make the random-true distribution less skewed. The workers' task was, for each pair, to check a YES radio button if they thought there is a relation between the words, NO otherwise. The words were annotated with their part of speech, and workers were instructed to pay attention to this information when making their choices. Extensive commented examples of both related pairs and unrelated ones were also provided in the instruction page. A minimum of 2 CF workers rated each pair, and, conservatively, we preserved only those items (about 12K) that were unanimously rated as unrelated to their concept by the judges. See Table 1 for summary statistics about the preserved random sets (nouns: RAND.N, adjectives: RAN.J, verbs:RAN.V).

4.4 BLESS statistics

For frequency information, we rely on the combination of the freely available ukWaC and Wackypedia corpora (size: 1.915B and 820M tokens, respectively).[2] The data set contains 200 concepts that have a mean corpus frequency of 53K occurrences (min. 1416 *chisel*, max. 793K *car*). The relata of these concepts (26,554 in total) are distributed as reported in Table 1.

Note that the distributions reflect certain "natural" differences between relations (hypernyms tend to be more frequent words than coordinates, but there are

[1] http://crowdflower.com/

[2] http://wacky.sslmit.unibo.it/

	frequency			*cardinality*		
relation	*min*	*avg*	*max*	*min*	*avg*	*max*
COORD	0	37K	1.7M	6	17.1	35
HYPER	31	138K	1.9M	2	6.7	15
MERO	0	133K	2M	2	14.7	53
ATTRI	0	501K	3.7M	4	13.6	27
EVENT	0	517K	5.4M	6	19.1	40
RAN.N	0	92K	2.4M	16	32.9	67
RAN.J	1	472K	4.5M	3	10.9	24
RAN.V	1	508K	7.7M	4	16.3	34

Table 1: Distribution (minimum, mean and maximum) of the relata of all BLESS concepts: the *frequency* columns report summary statistics for corpus counts across relata instantiating a relation; the *cardinality* columns report summary statistics for number of relata instantiating a relation across the 200 concepts, only considering relata with corpus frequency ≥ 100.

more coordinates than hypernyms, etc.). Instead of trying to artificially control for these differences, we assess their impact in Section 5 by looking at the behavior of baselines that exploit the frequency and cardinality of relations as proxies to semantic similarity (such factors could also be entered as regressors in a linear model).

5 Evaluation

This section illustrates one possible way to use BLESS to explore and evaluate DSMs. Given the similarity scores provided by a model for a concept with all its relata across all relations, we pick the relatum with the highest score (nearest neighbour) for each relation (see discussion in Section 3 above on why we allow models to pick their favorite from a set of relata instantiating the same relation). In this way, for each of the 200 BLESS concepts, we obtain 8 similarity scores, one per relation. In order to factor out concept-specific effects that might add to the overall score variance (for example, a frequent concept might have a denser neighborhood than a rarer one, and consequently the nearest relatum scores of the former are trivially higher than those of the latter), we transform the 8 similarity scores of each concept onto standardized z scores (mean: 0; s.d: 1) by subtracting from each their mean, and dividing by their standard deviation. After this transformation, we produce a **boxplot** summarizing the distribution of scores per relation across the 200 concepts (i.e., each box of the plot summarizes the distribution of the 200 standardized scores picked for each relation). Our boxplots (see examples in Fig. 1 below) display the median of a distribution as a thick horizontal line within a box extending from the first to the third quartile, with whiskers covering 1.5 of the interquartile range in each direction from the box, and values outside this extended range – extreme outliers – plotted as circles (these are the default boxplotting option of the R statistical package).[3] While the boxplots are extremely informative about the relation types that are best captured by models, we expect some degree of overlap among the distributions of different relations, and in such cases we might want to ask whether a certain model assigns significantly higher scores to one relation rather than another (for example, to *coordinates* rather than *random nouns*). It is difficult to decide *a priori* which pairwise statistical comparisons will be interesting. We thus take a conservative approach in which we perform *all* pairwise comparisons using the **Tukey Honestly Significant Difference** test, that is similar to the standard *t* test, but accounts for the greater likelihood of Type I errors when multiple comparisons are performed (Abdi and Williams, 2010). We only report the Tukey test results for those comparisons that are of interest in the analysis of the boxplots, using the standard $\alpha = 0.05$ significance threshold.

5.1 Models

Occurrence and co-occurrence statistics for all models are extracted from the combined ukWaC and Wackypedia corpora (see Section 4.4 above). We exploit the automated morphosyntactic annotation of the corpora by building our DSMs out of lemmas (instead of inflected words), and relying on part of speech information.

Baselines. The **RelatumFrequency** baseline uses the frequency of occurrence of a relatum as a surrogate of its cosine with the concept. With this approach, we want to verify that the unequal frequency distribution across relations (see Table 1 above) is not trivially sufficient to differentiate relation classes in a semantically interesting way. For our second baseline, we assign a random number as cosine sur-

[3] `http://www.r-project.org/`

rogate to each relatum (to smooth these random values, we generate them by first sampling, for each relatum, 10K random variates from a uniform distribution, and then averaging them). If the set of relata instantiating a certain relation is larger, it is more likely that it will contain the highest random value. Thus, this **RelationCardinality** baseline will favor relations that tend to have large relata set across concepts, controlling for effects due to different cardinalities across semantic relations (again, see Table 1 above).

DSMs. We choose a few ways to construct DSMs for illustrative purposes only. All the models contain vector representations for the same words, namely, approximately, the top 20K most frequent nouns, 5K most frequent adjectives and 5K most frequent verbs in the combined corpora. All the models use Local Mutual Information (Evert, 2005; Baroni and Lenci, 2010) to weight raw co-occurrence counts (this association measure is obtained by multiplying the raw count by Pointwise Mutual Information, and it is a close approximation to the Log-Likelihood Ratio). Three DSMs are based on counting co-occurrences with collocates within a window of fixed width, in the tradition of HAL (Lund and Burgess, 1996) and many later models. The **ContentWindow2** model records sentence-internal co-occurrence with the nearest 2 content words to the left and right of each target concept (the same 30K target nouns, verbs and adjectives are also employed as context content words). **ContentWindow20** is like ContentWindow2, but considers a larger window of 20 words to the left and right of the target. **AllWindow2** adopts the same window of ContentWindow2, but considers all co-occurrences, not only those with content words. The **Document** model, finally, is based on a (Local-Mutual-Information transformed) word-by-document matrix, recording the distribution of the 30K target words across the documents in the concatenated corpus. This DSM is thus akin to traditional Latent Semantic Analysis (Landauer and Dumais, 1997), without dimensionality reduction. The content-window-based models have, by construction, about 30K dimensions. The other models are much larger, and for practical reasons we only keep 1 million dimensions (those that account, cumulatively, for the largest proportion of the overall Local Mutual Information mass).

5.2 Results

The concept-by-concept z-normalized distributions of cosines of relata instantiating each of our relations are presented, for each of the example models, in Fig. 1. The RelatumFrequency baseline shows a preference for adjectives and verbs in general, independently of whether they are meaningful (attributes, events) or not (random adjectives and verbs), reflecting the higher frequencies of adjectives and verbs in BLESS (Table 1). The RelationCardinality baseline produces even less interesting results, with a strong preference for random nouns, followed by coordinates, events and random verbs (as predicted by the distribution in Table 1). We can conclude that the semantically meaningful patterns produced by the other models cannot be explained by trivial differences in relatum frequency or relation cardinality in the BLESS data set.

Moving then to the real DSMs, ContentWindow2 essentially partitions the relations into 3 groups: coordinates are the closest relata, which makes sense since they are, taxonomically, the most similar entities to target concepts. They are followed by (but significantly closer to the concept than) events, hypernyms and meronyms (events and hypernyms significantly above meronyms). Next come the attributes (significantly lower cosines than all relation types above). All the meaningful relata are significantly closer to the concepts than the random relata. Similar patterns can be observed in the ContentWindow20 distribution, however in this case the events, while still significantly below the coordinates, are significantly above the (statistically indistinguishable) hypernym, meronym and attribute set. Again, all meaningful relata are above the random ones. Both content-window-based models provide reasonable results, with ContentWindow2 being probably closer to our "ontological" intuitions. The high ranking of events is probably explained by the fact that a nominal concept will often appear as subject or object of verbs expressing associated events (*dog barks*, *fishing tuna*), and thus the corresponding verbs will share even relatively narrow context windows with the concept noun. The AllWindow2 distribution probably reflects the fact that many contexts picked by this DSM are function

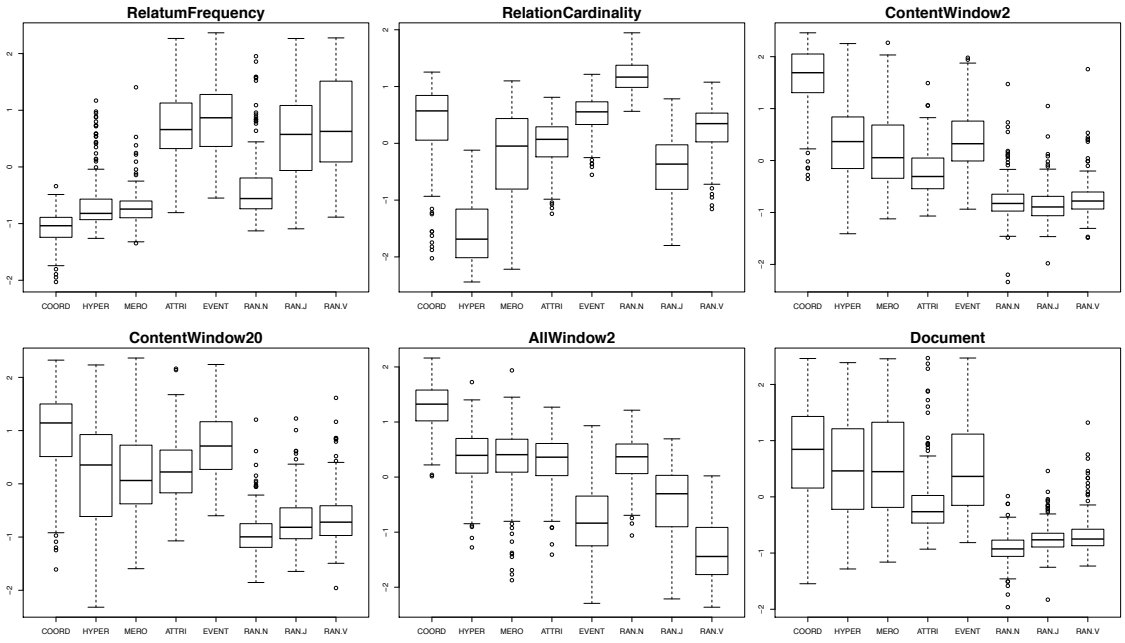

Figure 1: Distribution of relata cosines across concepts (values on ordinate are cosines after concept-by-concept z-normalization).

words, and thus they capture syntactic, rather than semantic distributional properties. As a result, random nouns are as high (statistically indistinguishable from) hypernyms and meronyms. Interestingly, attributes also belong to this subset of relations – probably due to the effect of determiners, quantifiers and other DP-initial function words, that will often occur both before nouns and before adjectives. Indeed, even random adjectives, although significantly below the other relations we discussed, are significantly above both random and meaningful verbs (i.e., events). For the Document model, all meaningful relations are significantly above the random ones. However, coordinates, while still the nearest neighbours (significantly closer than all other relations) are much less distinct than in the window-based models. Note that we cannot say *a priori* that ContentWindow2 is better than Document because it favors coordinates. However, while they are both able to sort out true and random relata, the latter shows a weaker ability to discriminate among different types of semantic relations (co-occurring within a document is indeed a much looser cue to similarity than specifically co-occurring within a narrow window). Traditional DSM tests, based on a single quality measure, would not have given us this broad view of how models are behaving.

6 Conclusion

We introduced BLESS, the first data set specifically designed for the intrinsic evaluation of DSMs. The data set contains tuples instantiating different, explicitly typed semantic relations, plus a number of controlled random tuples. Thus, BLESS can be used to evaluate both the ability of DSMs to discriminate truly related word pairs, and to perform in-depth analyses of the types of semantic relata that different models tend to favor among the nearest neighbors of a target concept. Even a simple comparison of the performance of a few DSMs on BLESS - like the one we have shown here - is able to highlight interesting differences in the semantic spaces produced by the various models. The success of BLESS will obviously depend on whether it will become a reference model for the evaluation of DSMs, something that can not be foreseen *a priori*. Whatever its destiny, we believe that the BLESS approach can boost and innovate evaluation in distributional semantics, as a key condition to get at a deeper understanding of its potentialities as a viable model for meaning.

References

Herv Abdi and Lynne Williams. 2010. Newman-Keuls and Tukey test. In N.J. Salkind, D.M. Dougherty, and B. Frey, editors, *Encyclopedia of Research Design*. Sage, Thousand Oaks, CA.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasça, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of HLT-NAACL*, pages 19–27, Boulder, CO.

Abdulrahman Almuhareb. 2006. *Attributes in Lexical Acquisition*. Phd thesis, University of Essex.

Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni, Stefan Evert, and Alessandro Lenci, editors. 2008. *Bridging the Gap between Semantic Theory and Computational Simulations: Proceedings of the ESSLLI Workshop on Distributional Lexical Semantic*. FOLLI, Hamburg.

Marco Baroni, Eduard Barbu, Brian Murphy, and Massimo Poesio. 2010. Strudel: A distributional semantic model based on properties and types. *Cognitive Science*, 34(2):222–254.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32:13–47.

D. A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences*. Dissertation, Stuttgart University.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

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*, 20(1):116–131.

Thomas Landauer and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Hugo Liu and Push Singh. 2004. ConceptNet: A practical commonsense reasoning toolkit. *BT Technology Journal*, pages 211–226.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28:203–208.

Ken McRae, George Cree, Mark Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

George Miller and Walter Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Gregory Murphy. 2002. *The Big Book of Concepts*. MIT Press, Cambridge, MA.

Timothy Rogers and James McClelland. 2004. *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press, Cambridge, MA.

Herbert Rubenstein and John Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Karen Sparck Jones and Julia R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer Verlag, Berlin.

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

Peter Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Morton E. Winston, Roger Chaffin, and Douglas Herrmann. 1987. A taxonomy of part-whole relations. *Cognitive Science*, 11:417–444.