The CogALex-V Shared Task on the Corpus-Based Identification of Semantic Relations

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

Anna Gladkova
The University of Tokyo, Japan
gladkova@phiz.c.u-tokyo.ac.jp

Stefan Evert
FAU Erlangen-Nürnberg, Germany
stefan.evert@fau.de

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

Abstract

The shared task of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V) aims at providing a common benchmark for testing current corpus-based methods for the identification of lexical semantic relations (*synonymy*, *antonymy*, *hypernymy*, *part-whole meronymy*) and at gaining a better understanding of their respective strengths and weaknesses. The shared task uses a challenging dataset extracted from EVALution 1.0 (Santus et al., 2015b), which contains word pairs holding the above-mentioned relations as well as semantically unrelated control items (*random*). The task is split into two subtasks: (i) identification of related word pairs vs. unrelated ones; (ii) classification of the word pairs according to their semantic relation. This paper describes the subtasks, the dataset, the evaluation metrics, the seven participating systems and their results. The best performing system in subtask 1 is GHHH ($F_1 = 0.790$), while the best system in subtask 2 is LexNet ($F_1 = 0.445$). The dataset and the task description are available at https://sites.google.com/site/cogalex2016/home/shared-task.

1 Introduction

Determining automatically if words are semantically related, and in what way, is important for Natural Language Processing (NLP) applications such as thesaurus generation (Grefenstette, 1994), ontology learning (Zouaq and Nkambou, 2008), paraphrase generation and identification (Madnani and Dorr, 2010), as well as for drawing inferences (Martinez-Gómez et al., 2016). Many NLP applications make use of handcrafted resources such as WordNet (Fellbaum, 1998). However, creating these resources is expensive and time-consuming; they are available for only a few languages, and their coverage inevitably lags behind the lexical and conceptual proliferation.

In the last decades, a number of corpus-based approaches have investigated the possibility of identifying lexical semantic relations by observing word usage. Even though these methods are still far from being able to provide a comprehensive model of how semantic relations work, pattern-based and distributional approaches (both supervised or unsupervised) have confirmed the existence of a strong connection between word meaning and word distribution.

The practical utility of this finding matches its theoretical significance. The connection between word meanings and their usage is gaining prominence in theories of the mental lexicon (Mikoajczak-Matyja, 2015) and language acquisition (Bybee and Beckner, 2015). The status of distributional semantics vis-à-vis linguistics and cognitive science (Lenci, 2008) depends on making progress in this area. To further assess and explore how much we can learn about semantic relations from word distribution, we propose a shared task as part of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V), co-located with COLING 2016 in Osaka, Japan.

The CogALex-V shared task is intended to provide a common benchmark for testing current corpus-based methods for the identification of lexical semantic relations in order to gain a better understanding of their respective strengths and weaknesses. It is articulated into two subtasks: (i) identification of semantically related word pairs vs. unrelated ones; (ii) classification of the word pairs according to their

Proceedings of the Workshop on Cognitive Aspects of the Lexicon,
pages 69–79, Osaka, Japan, December 11-17 2016.

semantic relation. Participants were provided with training and test datasets extracted from EVALution 1.0 (Santus et al., 2015b), as well as a scoring script for evaluating the output of their systems.

The shared task has been intended and designed as a "friendly competition": the goal was to identify strengths and weaknesses of various methods, rather than just "crowning" the best-performing model. In total, seven systems participated in the shared task. Most of them exploited Distributional Semantic Models (DSMs), either of the count-based or word-embedding type (Baroni et al., 2014). Most of them relied on distance or nearest neighbors in subtask 1, and on machine learning classifiers (e.g., Support Vector Machine (SVM), Convolutional Neural Network (CNN) and Random Forest (RF)) in subtask 2. Some systems enriched the DSM representation by adopting patterns (e.g., *LexNet*, the best system in subtask 2) or extracting distributional properties with unsupervised measures (e.g., ROOT18).

This paper reports the results achieved by the participating systems, providing insights about their respective strengths and weaknesses. It is organized as follows. Section 2 surveys similar shared tasks and provides an overview of existing methods for identifying lexical semantic relations. Section 3 introduces the task, the datasets, and the participating systems (each of them described in detail in a separate paper included in the workshop proceedings).[1] Section 4 lists the performance of the participating systems, analyzing it from several perspectives. Section 5 summarizes the findings, highlights the contribution of the shared task and suggests a few directions for future research.

2 Related Work

2.1 Shared Tasks on Semantic Relations Identification

The importance of efficient and accurate identification of different semantic relations for different NLP applications has already prompted several shared tasks, differing in the relations considered and the task definitions. These tasks are briefly surveyed in the current section.

SemEval-2007 shared task 4 (Girju et al., 2007) focused on seven "encyclopedic" semantic relations between nouns (*cause-effect, instrument-agency, product-producer, origin-entity, content-container, theme-tool, part-whole*). In order to disambiguate the senses, the participants could rely on WordNet synsets and/or on sentences in which the noun pairs were observed. The best system out of fifteen achieved 76.3% average accuracy.

SemEval-2010 shared task 8 (Hendrickx et al., 2010) considered the first five semantic relations of SemEval-2007 shared task 4, with the addition of *entity-destination, component-whole, member-collection, and message-topic*. These relations were annotated in sentence contexts. Given a sentence and two tagged nominals, the task was to predict the relation between those nominals and its direction. The best system out of twenty-eight achieved 82% accuracy. The participants were free to use various semantic, syntactic and morphological resources.

Related to the task of lexical semantic relation identification is the task of taxonomy construction, which essentially focuses on only one semantic relation: hypernymy (and its inverse, hyponymy). This task was explored in *SemEval-2015* (Bordea et al., 2015) and *SemEval-2016* (Buitelaar et al., 2016). The test data consisted of a list of domain terms that participants had to structure into a taxonomy (a list of pairs <term, hypernym>), possibly adding intermediate terms. The participating systems used lexical patterns, dictionary definitions, Wikipedia, knowledge bases, and vector space models. Also noteworthy is *SemEval-2016 Task 14* (Jurgens and Pilehvar, 2016), which asked participants to enrich WordNet taxonomy by determining, for a given new word, which synset it should be part of (thus combining detection of hypernyms with word sense disambiguation).

The present shared task differs from those listed above in the semantic relations it considers: *synonymy*, *antonymy*, *hypernymy*, *part-whole meronymy*, and *random* or "semantically unrelated". It also differs from SemEval-2010 task 8 and SemEval-2007 task 4 in the absence of sentence contexts for the pairs of target words. Most importantly, unlike the above tasks, the CogALex-V shared task forbids the use of any thesauri, knowledge bases, or semantic networks (particularly WordNet and ConceptNet), forcing the participating systems to rely exclusively on corpus data.

[1] Training and test data as well as further information about the shared task are available at `https://sites.google.com/site/cogalex2016/home/shared-task`.

2.2 Methods for the Identification of Semantic Relations

Up to this date, several corpus-based approaches to the identification of semantic relations have been proposed. Most of them, however, focus on a single semantic relation with the ambitious objective of isolating it from all the others. Dealing with multiple relations has been found particularly challenging, and few systems have attempted multi-class classifications. The exceptions include Turney (2008) and Pantel and Pennacchiotti (2006).

Early approaches rely on lexical-syntactic patterns (e.g. "tools *such as* hammers"). After the seminal work of Hearst (1992) who sketched methods for pattern discovery, Snow et al. (2004) adopted machine learning over dependency-paths-based features. While these approaches focused on hypernyms, Pantel and Pennacchiotti (2006) introduced *Espresso*, able to identify several semantic relations (i.e. *hypernymy*, *part-of*, *succession*, *reaction* and *production*) as well as to maximize recall by using the Web and precision by assessing the reliability of the patterns. Other pattern-based approaches to synonymy and antonymy are reported by Lin et al. (2003), Turney (2008), Wang et al. (2010) and Lobanova et al. (2010).

The major limitation of pattern-based approaches is that they require words to co-occur in the same sentence, strongly impacting the recall. Distributional approaches have therefore been adopted to reduce such limitations. They are based on the *Distributional Hypothesis* (Harris, 1954; Firth, 1957) that words occurring in similar contexts also bear similar meaning. Distributional approaches can be (i) unsupervised, generally consisting of mathematical functions that implement linguistic hypotheses about how and which contexts are relevant to identify specific relations (Kotlerman et al., 2010; Lenci and Benotto, 2012; Santus et al., 2014); or (ii) supervised, generally consisting of algorithms that automatically learn some distributional information about the words holding a specific relation (Weeds et al., 2014; Roller et al., 2014; Roller and Erk, 2016; Santus et al., 2016; Nguyen et al., 2016; Shwartz et al., 2016). While unsupervised approaches are commonly outperformed by supervised ones, the latter – which rely on distributional word vectors, either concatenated or combined through algebraic functions – seem to learn specific lexical properties of the words in the pairs rather than the general semantic relation existing between them (Weeds et al., 2014; Levy et al., 2015b). This has a negative impact on their performance on previously unseen words, lexically split datasets and unseen switched pairs (Santus et al., 2016).

One of the ongoing disputes in the NLP community concerns the relative merits and demerits of count-based distributional models and word embeddings (which are obtained by training neural networks rather than counting co-occurrence frequencies). While the latter seem to outperform the former in several tasks such as similarity estimation (Baroni et al., 2014), both types of models are subject to variation at the level of individual linguistic relations (Gladkova et al., 2016). Levy et al. (2015a) have also shown that optimization of hyperparameters can make a bigger difference than the choice between different models.

Finally, very recently, several scholars have investigated the possibility of integrating different kinds of information. Kiela et al. (2015) have used image generality for hypernymy detection, while Shwartz et al. (2016) have tried to identify the same relation by combining pattern-based and distributional information.

3 Shared task

3.1 Task description

The CogALex-V shared task was conducted as a "friendly competition" where participants had access to both training and testing datasets, released on the 8th and the 27th of September 2016, respectively. The participants were asked to evaluate the output of their system with the official evaluation script, released with the test set together with *random* and *majority* baselines. Each participant was furthermore requested to submit a description paper and the output of their system in the two subtasks by the 16th of October 2016. Two reviews for each paper were returned by the 25th of October 2016, and the camera-ready version was due on the 2nd of November. The shared task was split in two subtasks which are described below.

Subtask 1. For each word pair (e.g. *dog – fruit*), decide whether the terms are semantically related (TRUE) or not (FALSE). Given a TAB-separated input file with word pairs, participating systems

must add a third column specifying their prediction. This subtask was evaluated in terms of precision, recall and F_1-score for the identification of related word pairs. The unrelated word pairs were considered as noise.

Subtask 2. For each word pair (e.g. *cat – animal*), decide which semantic relation (if any) holds between the two words. The options are *synonymy* (SYN), *antonymy* (ANT), *hypernymy* (HYPER), *part-whole meronymy* (PART_OF) and *random* (RANDOM) for pairs where none of the four relations holds (see section 3.2). The input file was the same as for subtask 1. Participant systems were expected to return a TAB-separated file, where each word pair is annotated with exactly one relation label. This subtask was evaluated in terms of precision, recall and F_1-score for each of the four semantic relations. The unrelated word pairs (RANDOM) were considered as noise and therefore not considered in the final weighted average.

As mentioned above, the participating systems were supposed to be entirely corpus-based, without recourse to any existing dictionaries, knowledge bases or semantic networks. However, there was no restriction on the corpora that could be used. The participants were free to use the provided training data for supervised machine learning or for developing or tuning an unsupervised system. For example, they could use purely handwritten knowledge patterns for relation mining or to learn knowledge patterns from the CogALex-V training data, but they could not bootstrap knowledge patterns from a different set of seed terms, and no other training data was allowed.

Each participant was asked to submit the output of the system whose results are reported in the description paper. Further post-hoc experiments were encouraged at the authors' discretion.

3.2 Datasets

The training and test datasets were constructed on the basis of EVALution 1.0 (Santus et al., 2015b), a dataset for evaluating distributional semantic models that was derived from WordNet 4.0 (Fellbaum, 1998) and ConceptNet 5.0 (Liu and Singh, 2004), and then refined through automatic filters and crowdsourcing.

EVALution 1.0 includes various parts of speech, both single words and multi-word units (e.g., *grow_up*).[2] Words have been stemmed (e.g. *feeling* appears as *feel*). This increases ambiguity in the dataset, but it is also consistent with the fact that semantic relations between lexical items are typically independent from their morphosyntactic realization (e.g. the hypernymic pair *anger – feel* now represents morphological variants such as *anger – feeling* and *anger – to feel*).

After being extracted from WordNet or ConceptNet, the pairs (e.g. *sweet* SYN *candy*) were evaluated by CrowdFlower workers in order to obtain native speaker judgments, which can be used as a proxy for the prototypicality of the relations. The crowdsourcing task was to rate the truthfulness of sentences generated from the word pairs (according to the templates presented in table 1) on a scale from 1 to 5, where 1=*completely disagree* and 5=*completely agree*. Five judgments were collected for each sentence.

The CrowdFlower workers also tagged the general domains in which the relata were found more appropriate, such as "nature", "culture" or "emotion". Unfortunately the reliability of these tags is fairly low, as some workers applied them randomly. We can therefore consider trustworthy only tags that were selected by a high number of voters. In addition to domains, EVALution contains other metadata, either concerning the pairs (e.g., from which resource the pair is inherited) or the single words (e.g., word frequency, capitalization distribution, morphological distribution, part-of-speech distribution, etc.). This metadata can be used for subsequent analysis of the performance of the systems.[3]

For this shared task, we extracted a subset of EVALution 1.0 that covers 747 target words (318 in the training set and 429 in the test set) with at least one of the following relata: *synonym*, *antonym*, *hypernym* and *part-whole meronym*; only pairs with average rating ≥ 4 were considered. In order to increase the difficulty of the identification task, for every target word we generated several random pairs by switching

[2]Multi-word units were filtered out for the shared task.

[3]Metadata is not available for the random pairs, but it is available for the individual words in the random pairs because they were generated exclusively from words contained in EVALution 1.0.

Relation	Tag	Template	Example	Training	Testing
Synonymy	SYN	W2 can be used with the same meaning as W1	*candy-sweet, apartment-flat*	167	235
Antonymy	ANT	W2 can be used as the opposite of W1	*clean-dirty, add-take*	241	360
Hypernymy	HYPER	W1 is a kind of W2	*cannabis-plant, actress-human*	255	382
Part-whole meronymy	PART_OF	W1 is a part of W2	*calf-leg, aisle-store*	163	224
Random word	RANDOM	None of the above relations apply	*accident-fish, actor-mild*	2228	3059

Table 1: Semantic relations in the shared task dataset

the relata. These pairs – approximately three times as many as related pairs – are intended to act as noise for the models. They may contain associated words (e.g. *coffee – cup*, *brick – build*), but pairs accidentally holding any of the four semantic relations above were filtered out manually.[4]

The dataset is particularly challenging for several reasons. First, it does not provide part-of-speech information for the words in the pairs, leaving the participant systems with the burden of disambiguation (e.g. *fire – shoot* are synonyms only when both are interpreted as verbs). Second, several words were interpreted in a specific meaning that does not always correspond to the dominant sense (e.g. *compact – car*, where *compact* is a noun referring to a specific kind of car). Third, it combines relations inherited from a lexical resource like WordNet with relations that were obtained by crowdsourcing and pattern-based extraction (in ConceptNet), making their definitions less consistent. Fourth, the terms in EVALution are stemmed, thereby denying systems the possibility of using morphological clues as features for the classification. Finding semantic relations between morphologically heterogeneous words is an additional challenge, but it is very likely that NLP applications (e.g. those for paraphrase generation and entailment verification) would benefit from the ability to focus on semantics while ignoring morphological differences. These difficulties sometimes appear together, e.g. in the hypernymic pair *stable – build*, where *stable* is used in the sense of "a building with stalls where horses, cattle, etc., are kept and fed"[5] and *build* is the stemmed form of *building*.

Although the above-mentioned difficulties could impact the possible performance of the competing systems, they stem from the very nature of natural language semantics. This is confirmed by the fact that CrowdFlower workers were clearly able to identify those pairs as semantically related. During the analysis of the systems, EVALution 1.0 metadata can be used for pinpointing the sources of problems.

3.3 Participants

The CogALex-V shared task had 7 participating teams in subtask 1, and 6 of these teams also took part in subtask 2. The methods and corpora used by these teams are summarized in table 2.

4 Results

4.1 Evaluation procedure

The participants were provided with a Python script for the evaluation. Given the gold standard and a system output file as input, it calculated precision, recall and their harmonic mean F_1 for related pairs (in subtask 1) or semantic relations (in subtask 2), ignoring the unrelated pairs. In subtask 2, for example, scores were computed for *synonymy* (SYN), *antonymy* (ANT), *hypernymy* (HYPER) and *part-whole meronymy* (PART_OF); the overall ranking of the systems was based on their weighted average.

[4] As the filtering was carried out by only two annotators, it is possible that a few such accidentally related pairs may have been overlooked.

[5] `http://www.wordreference.com/definition/stable` (retrieved on 3rd of November 2016)

Team	Method(s)	Corpus size	Corpus
GHHH	Word analogies, linear regression and multi-task CNN	100B	Google News (pre-trained word2vec embeddings, 300 dim.);
		6B	Wikipedia + Gigaword 5 (pre-trained GloVe embeddings, 300 dim.),
		840B	Common Crawl (pre-trained GloVe embeddings, 300 dim.)
Mach5	angular distance in SVD-reduced count-based DSM for subtask 1 and linear SVM classifier based on 1200 SVD dimensions in subtask 2	9.5B	ENCOW 2014, traditional dependency-based DSM
LexNet	multi-layer perceptron classifying feature vectors that consist of embeddings for two words and all dependency paths that connect them in a corpus	6B	Wikipedia + Gigaword 5 (pre-trained GloVe embeddings, 50-dim.);
		100B	Google News (pre-trained word2vec embeddings, 300 dim.)
ROOT18	random forest classifier trained on 18 features representing unsupervised distributional properties of the investigated relations	2B	UkWac, count-based BOW DSM
LOPE	cosine similarity, nearest neighbor position indexing, assuming the order synonymy-antonymy-hypernymy-meronymy-random	100B	Google News (pre-trained word2vec embeddings, 300 dim.)
HsH-Supervised	cosine similarity, classification based on SVM	2B	ukWaC (sparse PPMI-weighted vectors, 17400 features)
CGSRC	CNN-based relation classification	100B	Google News (pre-trained word2vec embeddings, 50–300 dim.)

Table 2: Description of the participating systems

The script requires that the gold standard and the output file contain exactly the same pairs, in the same order, and using the same annotation labels.

4.2 Results and ranks

Most of the participating systems obtained fairly good results in subtask 1. Performance was however much worse for all of them (even the best systems) in subtask 2, demonstrating once more that the identification of semantic relations is a hard task that calls for more attention from the community.

Team	Subtask 1
GHHH	0.790
Mach5	0.778
LexNet	0.765
ROOT18	0.731
LOPE	0.713
HsH-Supervised	0.585
CGSRC	0.431

Team	Subtask 2
LexNet	0.445
GHHH	0.423
Mach5	0.295
ROOT18	0.262
CGSRC	0.252
LOPE	0.247

Table 3: Participating systems ranked by their F_1 scores in subtask 1 (left) and subtask 2 (right)

Table 3 ranks the participating systems according to their F_1-scores in subtask 1 and subtask 2. The best performing system in subtask 1 is GHHH ($F_1 = 0.790$), with the first 5 top systems being less than 10% behind, and Mach5 ($F_1 = 0.778$) and LexNet ($F_1 = 0.765$) less than 3%. This confirms that numerous corpus-based approaches are competitive in discriminating between related and unrelated word pairs. The situation is quite different for subtask 2, where the same three systems achieve the highest scores, but now LexNet comes first ($F_1 = 0.445$), GHHH second ($F_1 = 0.423$) with less than 3% difference, and Mach5 ($F_1 = 0.295$) lags behind much more than in subtask 1, achieving a score that is closer to the last three systems than to the first two.

As can be seen in Table 2, the top systems use very different approaches. GHHH investigates word analogies, linear regression and multi-task Convolutional Neural Networks (CNN) with 300-dimensional publicly available word embeddings trained on huge corpora (Google News, Common Crawl and Wikipedia + Gigaword 5). The authors found that linear regression works better in subtask 1 (i.e. binary

classification), while multi-task CNN performs better in subtask 2, which involves multi-class classification. Analogy was instead found less appropriate for semantic relation identification.

LexNet relies on Wikipedia + Gigaword 5 and Google News corpora, leveraging the combination of distributional and path-based information. The authors merged the 50-dimensional GloVe pre-trained embeddings (Pennington et al., 2014) for the words in the pairs with the average embedding vector – created using a LSTM (Hochreiter and Schmidhuber, 1997) – of all the dependency paths that connect them in the corpus. In subtask 1, LexNet is combined with vector cosine (calculated on word2vec embeddings trained on Google News) through weights that were learned on a validation set. In subtask 2, in order to avoid a bias towards the majority class RANDOM, a Multi-Layer Perceptron (MLP) is trained and applied only on pairs that were classified as related in subtask 1.

The third system, Mach5, investigates the structure and hyperparameters of two traditional dependency-filtered and dependency-structured DSMs trained on a Web corpus of 9.5 billion words. The author sets most parameters according to Lapesa and Evert (2014), focusing on feature selection and optimization of SVD dimensions. Distance information is used directly in subtask 1, while for subtask 2 a linear SVM classifier is applied to 1200-dimensional vectors representing partial Euclidean distance in the two SVD-reduced spaces. Given the competitive results in subtask 1 and the much lower performance achieved in subtask 2, it is evident that Mach5 was optimized for identifying non-random pairs rather than for recognizing and discriminating specific semantic relations.

The other systems include ROOT18, which relies on several unsupervised features extracted from ukWaC that aim at identifying specific semantic relations. Like Mach5, the system performs relatively well in subtask 1, but is much worse in subtask 2. LOPE achieves similar performance to ROOT18 in both subtasks. It uses word2vec embeddings trained on Google News to determine the most similar words for each target; it classifies as related only the words appearing in the top-N nearest neighbors (with $N = 600$). In subtask 2, LOPE classifies the semantic relations according to the rank of the words in the nearest neighbors list, assuming that they are ranked decreasingly as synonyms-antonyms-hypernyms-meronyms-randoms.

The other two systems, CGSRC and HsH-Supervised, perform worse in subtask 1. CGSRC, however, obtains results comparable to ROOT18 in subtask 2, while HsH-Supervised did not participate in this task. CGSRC relies on a CNN architecture with four layer types: an input layer, a convolution layer, a max pooling layer and a fully connected softmax layer for term-pair relation classification. The CNN works on word2vec embeddings trained on about 100 billion words of Google News corpus. Finally, HsH-Supervised is an SVM classifier trained on the multiplication of the distributional vectors of the two words in the pairs extracted from ukWaC (similar to the approach of Mach5 in subtask 2). This method was reported to perform worse than cosine similarity on the same vectors.

As a rough summary, all systems relied on DSMs, in either "count" (Mach5, ROOT18 and HSH-Supervised) or "predict" form (GHHH, LexNet, LOPE and CGSRC). These DSMs were trained on corpora whose size ranges from 2 billion to 840 billion words (with "count" models relying on the

W1	W2	Gold	Prediction
cold	bad	FALSE	TRUE
combine	create	FALSE	TRUE
come	fill	FALSE	TRUE
dark	narrow	FALSE	TRUE
democracy	peace	FALSE	TRUE
depress	injure	FALSE	TRUE
desert	darkness	FALSE	TRUE
desert	landscape	FALSE	TRUE
enjoyment	quality	FALSE	TRUE
eye	lens	FALSE	TRUE

Table 4: Sample of pairs that were misclassified by the top three systems

W1	W2	Gold	Prediction
club	weapon	TRUE	FALSE
cold	friendly	TRUE	FALSE
commerce	deal	TRUE	FALSE
contract	grow	TRUE	FALSE
cook	action	TRUE	FALSE
crowd	desert	TRUE	FALSE
crowd	one	TRUE	FALSE
crown	base	TRUE	FALSE
cube	die	TRUE	FALSE
dart	action	TRUE	FALSE

Table 5: Sample of pairs that were misclassified by the top three systems

smaller corpora between 2 and 9.5 billion words). There seems to be a correlation between corpus size and system performance, even though it is not linear. GHHH, for example, obtains its highest performance in subtask 2 with embeddings trained on 840 billion words, but when embeddings trained on 6 billion words are used the performance is only slightly behind. The impact is much bigger when comparing systems based on 2 billion words with systems based on 6 billion words of corpus data.

Another observation is that vector distance or nearest neighbor information seems to be sufficient to obtain competitive results in subtask 1, but subtask 2 proves to be much more complex. Several classifiers have been adopted (SVM, Linear Regression, Random Forest and CNN), but none of them seems to have a clear edge on the others: the best two systems rely on a CNN (GHHH) and on a MLP (LexNet), but the CNN is also used by CGSRC with much less convincing results.

Further information about the systems and their parameters can be found in the respective description papers in this volume.

4.3 Analysis of results

In order to provide some insights about what went wrong in the systems and whether the dataset might have to be blamed for their relatively low performance in subtask 2, we investigated how many and which pairs were misclassified by the top three systems, separately for each subtask.

Subtask 1. As many as 162 pairs out of 4,260 were misclassified by all the top three systems: 60 of them are unrelated pairs wrongly classified as related (see Table 4 for examples), while the remaining 102 are related pairs in the gold standard that were not recognized by the systems (see Table 5). As can be seen from Table 4, many of the false positives carry some kinds of association (e.g. *cold – bad*, *combine – create*, *eye – lens*, etc.), which in very few cases might be due to an accidental semantic relationship not filtered out by the annotators (e.g. *desert – landscape* as hypernymy). In Table 5, instead, we notice that most of the false negatives include highly ambiguous words, mostly used in rare senses (e.g. the hypernymic *club – weapon*, the antonymous *crown – base*, etc.) and/or very general hypernyms (e.g. *dart – action* and *cook – action*).

Subtask 2. As many as 513 pairs out of 4,260 were misclassified by all the top three systems. 237 of them received the same label. In Table 6 we summarize the number of pairs per relation that were misclassified, both with different labels (on the left) and with the same ones (on the right). Among the 237 misclassified pairs, the large majority (i.e. 172) were misclassified as RANDOM, while the others were misclassified between the various relations. With respect to these ones, hypernyms were most often confused with synonyms (even native speakers may have a hard time discriminating them: e.g. *dessert – sweet*) and antonyms (as they might share similar distributional properties, cf. Santus et al. (2015a)). Also, hypernyms were sometimes confused with part-whole meronyms. This is particularly likely to happen if one of the words is semantically ambiguous (e.g. *sugar – candy*). Further errors should probably be attributed to the stemmed form of the words (e.g. the hypernymic *pride – feel(ing)*), to their ambiguity (e.g. *duck – move*), and to a large difference in generality between the related words

Receiving any label		Receiving the same label	
143	ANT	62	ANT
140	HYPER	68	HYPER
85	PART_OF	50	PART_OF
22	RANDOM	9	RANDOM
123	SYN	48	SYN
513	total	237	total

Table 6: Pairs that were misclassified by the top three systems, organized by gold relation

(e.g. *cook – action*).

5 Conclusion

In this paper, we have described the shared task of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V), which aims at testing corpus-based methods for the identification of semantically related words on the same benchmark in order to gain a better understanding of how such methods can model the acquisition and manipulation of semantic relations.

A dataset extracted from EVALution 1.0 (Santus et al., 2015b), and split into a training and a test set, was provided at `https://sites.google.com/site/cogalex2016/home/shared-task` in September 2016, together with an evaluation script and two baselines (majority and random). Seven participants submitted their system output and their paper description in October 2016. The task was divided into two subtasks, respectively aiming at the binary classification of related vs. unrelated words and at the multi-class classification of synonyms, antonyms, hypernyms, meronyms and random pairs.

The systems achieved a reasonable F_1 score in the first subtask (GHHH was the best system with $F_1 = 0.790$), but a rather low performance in subtask 2 (LexNet was the best system with $F_1 = 0.445$). This is certainly due to the inherent difficulty of the multi-class setting, but compounded by a series of other difficulties rooted in the design of the dataset, which uses ambiguous and stemmed words without part-of-speech information. These results suggest that there is still need for improvement and we hope that this shared task has provided a challenging dataset and state-of-the-art baselines to support further investigation. We would also like to point out that our dataset includes metadata from EVALution 1.0 (i.e. semantic domain, word frequency, capitalization distribution, morphological distribution, part-of-speech distribution, etc.), which can be used to evaluate the performance of the system and to pinpoint the sources of problems.

As a general note to organizers of future shared tasks, we would suggest to keep the factors of variability in the participating systems as low as possible, or at least require explicit analyses of these factors. In fact, although we were able to draw some general conclusions about the participating systems (see section 4), it is hard to determine the precise impact of relevant factors such as corpus size, especially if these factors are not explicitly analyzed in all the system description papers.

References

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

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)A*, volume 452, pages 902–910. Association for Computational Linguistics.

Paul Buitelaar, Georgeta Bordea, and Els Lefever. 2016. SemEval-2016 Task 13: Taxonomy Extraction Evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091. Association for Computational Linguistics.

Joan L. Bybee and Clay Beckner. 2015. Usage-based theory. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*, pages 954–979. Oxford University Press.

Christiane Fellbaum, editor. 1998. *Wordnet: An Electronic Lexical Database*. Language, speech, and communication series. MIT Press Cambridge.

J. R. Firth. 1957. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 13–18. Association for Computational Linguistics.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesnt. In *Proceedings of naacl-hlt*, pages 8–15.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, pages 33–38. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

David Jurgens and Mohammad Taher Pilehvar. 2016. SemEval-2016 Task 14: Semantic Taxonomy Enrichment. In *Proceedings of SemEval-2016*, pages 1016–1026. Association for Computational Linguistics.

Douwe Kiela, Laura Rimell, Ivan Vulic, and Stephen Clark. 2015. Exploiting image generality for lexical entailment detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*. ACL.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(04):359–389.

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*, 2:531–545.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 75–79. Association for Computational Linguistics.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. 20(1):1–31.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015a. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015b. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics â AS Human Language Technologies (NAACL HLT 2015), Denver, CO*.

Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2003. Identifying synonyms among distributionally similar words. In *IJCAI*, volume 3, pages 1492–1493.

Hugo Liu and Push Singh. 2004. ConceptNeta practical commonsense reasoning tool-kit. 22(4):211–226.

Anna Lobanova, Gosse Bouma, and Erik Tjong Kim Sang. 2010. Using a treebank for finding opposites. In *Ninth International Workshop on Treebanks and Linguistic Theories*, page 139.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. 36(3):341–387.

Pascual Martinez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2016. ccg2lambda: A Compositional Semantics System. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics - System Demonstrations*, pages 85–90. Association for Computational Linguistics.

Nawoja Mikoajczak-Matyja. 2015. The Associative Structure of the Mental Lexicon: Hierarchical Semantic Relations in the Minds of Blind and Sighted Language Users. 19(1):1–18. [doi:10.1515/plc-2015-0001].

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. *CoRR*, abs/1605.07766.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 113–120. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43.

Stephen Roller and Katrin Erk. 2016. Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment. *arXiv preprint arXiv:1605.05433*.

Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *COLING*, pages 1025–1036.

Enrico Santus, Qin Lu, Alessandro Lenci, and Chu-Ren Huang. 2014. Unsupervised antonym-synonym discrimination in vector space. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014, 9-10 December 2014, Pisa*, volume 1, pages 328–333. Pisa University Press.

Enrico Santus, Alessandro Lenci, Qin Lu, and Chu-Ren Huang. 2015a. When similarity becomes opposition: Synonyms and antonyms discrimination in dsms. *Italian Journal on Computational Linguistics*, 1(1).

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015b. EVALution 1.0: An Evolving Semantic Dataset for Training and Evaluation of Distributional Semantic Models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015)*, pages 64–69.

Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. 2016. Nine features in a random forest to learn taxonomical semantic relations. *arXiv preprint arXiv:1603.08702*.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. *ACL 2016*.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Peter D. Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 905–912.

Wenbo Wang, Christopher Thomas, Amit Sheth, and Victor Chan. 2010. Pattern-based synonym and antonym extraction. In *Proceedings of the 48th Annual Southeast Regional Conference*, page 64. ACM.

Julie Weeds, Daoud Clarke, Jeremy Reffin, David J Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *COLING*, pages 2249–2259.

Amal Zouaq and Roger Nkambou. 2008. Building domain ontologies from text for educational purposes. 1(1):49–62.