SemEval-2013 Task 5: Evaluating Phrasal Semantics

Ioannis Korkontzelos
National Centre for Text Mining
School of Computer Science
University of Manchester, UK
ioannis.korkontzelos@man.ac.uk

Torsten Zesch
UKP Lab, CompSci Dept.
Technische Universität Darmstadt
Germany
zesch@ukp.informatik.tu-darmstadt.de

Fabio Massimo Zanzotto
Department of Enterprise Engineering
University of Rome "Tor Vergata"
Italy
zanzotto@info.uniroma2.it

Chris Biemann
FG Language Technology, CompSci Dept.
Technische Universität Darmstadt
Germany
biem@cs.tu-darmstadt.de

Abstract

This paper describes the SemEval-2013 Task 5: "Evaluating Phrasal Semantics". Its first subtask is about computing the semantic similarity of words and compositional phrases of minimal length. The second one addresses deciding the compositionality of phrases in a given context. The paper discusses the importance and background of these subtasks and their structure. In succession, it introduces the systems that participated and discusses evaluation results.

1 Introduction

Numerous past tasks have focused on leveraging the meaning of word types or words in context. Examples of the former are noun categorization and the TOEFL test, examples of the latter are word sense disambiguation, metonymy resolution, and lexical substitution. As these tasks have enjoyed a lot success, a natural progression is the pursuit of models that can perform similar tasks taking into account multiword expressions and complex compositional structure. In this paper, we present two subtasks designed to evaluate such phrasal models:

a. Semantic similarity of words and compositional phrases

b. Evaluating the compositionality of phrases in context

The aim of these subtasks is two-fold. Firstly, considering that there is a spread interest lately in phrasal semantics in its various guises, they provide an opportunity to draw together approaches to numerous related problems under a common evaluation set. It is intended that after the competition, the evaluation setting and the datasets will comprise an ongoing benchmark for the evaluation of these phrasal models.

Secondly, the subtasks attempt to bridge the gap between established lexical semantics and full-blown linguistic inference. Thus, we anticipate that they will stimulate an increased interest around the general issue of phrasal semantics. We use the notion of phrasal semantics here as opposed to lexical compounds or compositional semantics. Bridging the gap between lexical semantics and linguistic inference could provoke novel approaches to certain established tasks such as lexical entailment and paraphrase identification, and ultimately lead to improvements in a wide range of applications in natural language processing, such as document retrieval, clustering and classification, question answering, query expansion, synonym extraction, relation extraction, automatic translation, or textual advertisement matching in search engines, all of which de-

pend on phrasal semantics.

Data Sources & Methodology Data instances of both subtasks are drawn from the large-scale, freely available WaCky corpora (Baroni et al., 2009). We ensured that data instances occur frequently enough in the WaCky corpora, so that participating systems could gather statistics for building distributional vectors or other uses. As the evaluation data only contains very small annotated samples from freely available web documents, and the original source is provided, we could provide them without violating copyrights.

The size of the WaCky corpora is suitable for training reliable distributional models. Sentences are already lemmatized and part-of-speech tagged. Participating approaches making use of distributional methods, part-of-speech tags or lemmas, were strongly encouraged to use these corpora and their shared preprocessing, to ensure the highest possible comparability of results. Additionally, this had the potential to considerably reduce the work-load of participants. For the first subtask, data were provided in English, German and Italian and for the second subtask in English and German.

The range of methods applicable to both subtasks was deliberately not limited to any specific branch of methods, such as distributional or vector models of semantic compositionality. We believe that the subtasks can be tackled from different directions and we expect a great deal of the scientific benefit to lie in the comparison of very different approaches, as well as how these approaches can be combined. An exception to this rule is the fact that participants in the first subtask were not allowed to use dictionaries or lexicons. Since the subtask is considered fundamental and its data were created from online knowledge resources, systems using similar tools to address it would be of limited use.

Participating systems were allowed to attempt one or both subtasks, in one or all of the languages supported. However, it was expected that systems performing well at the first basic subtask would provide a good starting point for dealing with the second subtask, which is considered harder. Moreover, language-independent models were of special interest.

The remainder of this paper is structured as follows: Section 2 discusses the first subtask, which is about semantic similarity of words and compositional phrases. In subsection 2.1 the subtask is described in detail together with some information about its background. Subsection 2.2 discusses the data creation process and subsection 2.3 discusses the participating systems and their results. Section 3 introduces the second subtask, which is about evaluating the compositionality of phrases in context. Subsection 3.1 explains the data creation process for this subtask. In subsection 3.2 the evaluation statistics of participating systems are presented. Section 4 is a discussion about the conclusions of the entire task. Finally, in section 5 we summarize this presentation and discuss briefly our vision about challenges in distributional semantics.

2 Subtask 5a: Semantic Similarity of Words and Compositional Phrases

The aim of this subtask is to evaluate the component of a semantic model that computes the similarity between word sequences of different length. Participating systems are asked to estimate the semantic similarity of a word and a short sequence of two words. For example, they should be able to figure out that *contact* and *close interaction* are similar whereas *megalomania* and *great madness* are not.

This subtask addresses a core problem, since satisfactory performance in computing the similarity of full sentences depends on similarity computations on shorter sequences.

2.1 Background and Description

This subtask is based on the assumption that we first need a basic set of functions to compose the meaning of two words, in order to construct more complex models that compositionally determine the meaning of sentences, as a second step. For compositional distributional semantics, the need for these basic functions is discussed in Mitchell and Lapata (2008). Since then, many models have been proposed for addressing the task (Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Guevara, 2010), but still comparative analysis is in general based on comparing sequences that consist of two words.

As in Zanzotto et al. (2010), this subtask proposes

contact/[kon-takt]

1. the act or state of touching; a touching or meeting, as of two things or people.
2. close interaction
3. an acquaintance, colleague, or relative through whom a person can gain access to information, favors, influential people, and the like.

Figure 1: The definition of *contact* in a sample dictionary

to compare the similarity of a 2-word sequence and a single word. This is important as it is the basic step to analyze models that can compare any word sequences of different length.

The development and testing set for this subtask were built based on the idea described in Zanzotto et al. (2010). Dictionaries were used as sources of positive training examples. Dictionaries are natural repositories of equivalences between words under definition and sequences of words used for defining them. Figure 1 presents the definition of the word *contact*, from which the pair *(contact, close interaction)* can be extracted. Such equivalences extracted from dictionaries can be seen as natural and unbiased data instances. This idea opens numerous opportunities:

- Since definitions in dictionaries are syntactically rich, we are able to create examples for different syntactic relations.
- We have the opportunity to extract positive examples for every language for which dictionaries with a sufficient number of entries are available.

Negative examples were generated by matching words under definition with randomly chosen defining sequences. In the following subsection, we provide more details about the application of this idea to build the development and testing set for subtask 5a.

Language	Train set	Test set	Total
English	5,861	3,907	9,768
German	1,516	1,010	2,526
Italian	1,275	850	2,125
German - no names	1,101	733	1,834

Table 1: Quantitative characteristics of the datasets

2.2 Data Creation

Data for this subtask were provided in English, German and Italian. Pairs of words under definitions and defining sequences were extracted from the English, German and Italian part of Wiktionary, respectively. In particular, for each language, all Wiktionary entries were downloaded and part-of-speech tagged using the Genia tagger (Tsuruoka et al., 2005). In succession, definitions that start with noun phrases were kept, only. For the purpose of extracting word and sequence pairs for this subtask, we consider as noun phrases, sequence that consist of adjectives or noun and end with a noun. In cases where the extracted noun phrase was longer than two words, the right-most two sequences were kept, since in most cases noun phrases are governed by their right-most component. Subsequently, we discarded instances whose words occur too infrequently in the WaCky corpora (Baroni et al., 2009) of each language. WaCky corpora are available freely and are large enough for participating systems to extract distributional statistics. Taking the numbers of extracted instances into account, we set the frequency thresholds at 10 occurrences for English and 5 for German and Italian.

Data instances extracted following this process were then checked by a computational linguist. Candidate pairs in which the definition sequence was not judged to be a precise and adequate definition of the word under definition were discarded. The final data sets were divided into training and held-out testing sets, according to a 60% and 40% ratio, respectively. The first three rows of table 1 present the numbers of the train and test sets for the three languages chosen. It was identified that a fair percentage of the German instances (approximately 27%) refer to the definitions of first names or family names. These instances were discarded from the German data set to

produce the data set described in the last row of table 1.

The training set was released approximately 3 months earlier than the test data. Each instance in the former set was annotated as positive or negative, while the instances of the latter were unannotated.

2.3 Results

Participating systems were evaluating on their ability to predict correctly whether the components of each test instance, i.e. word-sequence pair, are semantically similar or distinct. Participants were allowed to use or ignore the training data, i.e. the systems could be supervised or unsupervised. Unsupervised systems were allowed to use the training data for development and parameter tuning. Since this is a core task, participating systems were not be able to use dictionaries or other prefabricated lists. Instead, they were allowed to use distributional similarity models, selectional preferences, measures of semantic similarity etc.

Participating system responses were scored in terms of standard information retrieval measures: accuracy (A), precision (P), recall (R) and F_1 score (Radev et al., 2003). Systems were encouraged to submit at most 3 solutions for each language, but submissions for fewer languages were accepted.

Five research teams participated. Ten system runs were submitted for English, one for German (on data set: German - no names) and one for Italian. Table 2 illustrates the results of the evaluation process. The teams of Hochschule Hannover - University of Applied Sciences and Arts *(HsH)*, the CLaC Laboratory - Concordia University, the University of Matanzas "Camilo Cienfuegos" and DLSI - University of Alicante *(UMCC_DLSI-(EPS))*, and Harbin Institute of Technology *(ITNLP)* approached the task in a supervised way, while IRIT & CNRS (MELODI) participated with two unsupervised approaches. Interestingly, these approaches performed better than some supervised ones for this experiment. Below, we summarise the properties of participating systems.

The system of HsH used distributed similarity and especially random indexing to compute similarities between words and possible definitions, under the hypothesis that a word and its definition are distributionally more similar than a word and an arbitrary definition. Considering all open-class words, context vectors over the entire WaCky corpus were computed for the word under definition, the defining sequence, its component words separately, the addition and multiplication of the vectors of the component words and a general context vector. Then, various similarity measures were computed on the vectors, including an innovative length-normalised version of Jensen-Shannon divergence. The similarity values are used to train a Support Vector Machine (SVM) classifier (Cortes and Vapnik, 1995).

The first approach (run 1) developed in the CLaC Laboratory is based on a weighted semantic network to measure semantic relatedness between the word and the components of the phrase. A PART classifier is used to generate a partial decision trained on the semantic relatedness information of the labelled training set. The second approach uses a supervised distributional method based on words frequently occurring in the Web1TB corpus to calculate relatedness. A JRip classifier is used to generate rules trained on the semantic relatedness information of the training set. This approach was used in conjunction with the first one as a backup method (run 2). In addition, features generated by both approaches were used to train the JRIP classifier collectively (run 3).

The first approach developed by MELODI at IRIT & CNRS, called *lvw*, uses a dependency-based vector space model computed over the ukWaC corpus, in combination with Latent Vector Weighting (Van de Cruys et al., 2011). The system computes the similarity between the first noun and the head noun of the second phrase, which was weighted according to the semantics of the modifier. The second approach, called *dm*, used a dependency-based vector space model, but, unlike the first approach, disregarded the modifier in the defining sequence. Since both systems are unsupervised, the training data was used to train a similarity threshold parameter, only.

The system of UMCC_DLSI-(EPS) locates the synsets of words in data instances and computes the semantic distances between each synset of the word under definition and each synsets of the defining sequence words. In succession, a classifier is trained using features based on distance and WordNet relations.

The first attempt of ITNLP (run 1) consisted of an

Language	Rank	Participant Id	run Id	A	R	P	rej. R	rej. P	F_1
English	1	HsH	1	.803	.752	.837	.854	.775	.792
	3	CLaC	3	.794	.707	.856	.881	.750	.774
	2	CLaC	2	.794	.695	.867	.893	.745	.771
	4	CLaC	1	.788	.638	.910	.937	.721	.750
	5	MELODI	lvw	.748	.614	.838	.882	.695	.709
	6	UMCC_DLSI-(EPS)	1	.724	.613	.787	.834	.683	.689
	7	ITNLP	3	.703	.501	.840	.904	.645	.628
	8	MELODI	dm	.689	.481	.825	.898	.634	.608
	9	ITNLP	1	.663	.392	.857	.934	.606	.538
	10	ITNLP	2	.659	.427	.797	.891	.609	.556
German	1	HsH	1	.825	.765	.870	.885	.790	.814
Italian	1	UMCC_DLSI-(EPS)	1	.675	.576	.718	.774	.646	.640

Table 2: Task 5a: Evaluation results. A, P, R, rej. and F_1 stand for accuracy, precision, recall, rejection and F_1 score, respectively.

SVM classifier trained on semantic similarity computations between the word under definition and the defining sequence in each instance. Their second attempt also uses an SVM, however trained on WordNet-based similarities. The third attempt of ITNLP is a combination of the previous two; it combines their features to train an SVM classifier.

3 Subtask 5b: Semantic Compositionality in Context

An interesting sub-problem of semantic compositionality is to decide whether a target phrase is used in its literal or figurative meaning in a given context. For example "big picture" might be used literally as in *Click here for a bigger picture* or figuratively as in *To solve this problem, you have to look at the bigger picture*. Another example is "old school" which can also be used literally or figuratively: *He will go down in history as one of the old school, a true gentlemen.* vs. *During the 1970's the hall of the old school was converted into the library.*

Being able to detect whether a phrase is used literally or figuratively is e.g. especially important for information retrieval, where figuratively used words should be treated separately to avoid false positives. For example, the example sentence *He will go down in history as one of the old school, a true gentlemen.* should probably not be retrieved for the query "school". Rather, the insights generated from subtask 5a could be utilized to retrieve sentences using a similar phrase such as "gentleman-like behavior". The task may also be of interest to the related research fields of metaphor detection and idiom identification.

There were no restrictions regarding the array of methods, and the kind of resources that could be employed for this task. In particular, participants were allowed to make use of pre-fabricated lists of phrases annotated with their probability of being used figuratively from publicly available sources, or to produce these lists from corpora. Assessing how well the phrase suits its context might be tackled using e.g. measures of semantic relatedness as well as distributional models learned from the underlying corpus.

Participants of this subtask were provided with real usage examples of target phrases. For each usage example, the task is to make a binary decision whether the target phrase is used literally or figuratively in this context. Systems were tested in two different disciplines: a *known phrases* task where all target phrases in the test set were contained in the training, and an *unknown phrases* setting, where all target phrases in the test set were unseen.

3.1 Data Creation

The first step in creating the corpus was to compile a list of phrases that can be used either literally or

Task	Dataset	# Phrases	# Items	Items per phrase	# Liter.	# Figur.	# Both
known	train	10	1,424	68–188	702	719	3
	dev	10	358	17–47	176	181	1
	test	10	594	28–78	294	299	1
unseen	train	31	1,114	4–75	458	653	3
	dev	9	342	4–74	141	200	1
	test	15	518	8–73	198	319	1

Table 3: Quantitative characteristics of the datasets

metaphorically. Thus, we created an initial list of several thousand English idioms from Wiktionary by listing all entries under the category ENGLISH IDIOMS using the JWKTL Wiktionary API (Zesch et al., 2008). We manually filtered the list removing most idioms that are very unlikely to be ever used literally (anymore), e.g. *to knock on heaven's door*. For each of the resulting list of phrases, we extracted usage contexts from the ukWaC corpus (Baroni et al., 2009). Each usage context contains 5 sentences, where the sentence with the target phrase appears in a randomized position. Due to segmentation errors, some usage contexts actually might contain less than 5 sentences, but we manually filtered all usage contexts where the remaining context was insufficient. This was done in the final cleaning step where we also manually removed (near) duplicates, obvious spam, encoding problems etc.

The target phrases in context were annotated for *figurative*, *literal*, *both* or *impossible to tell* usage, using the CrowdFlower[1] crowdsourcing annotation platform. We used about 8% of items as "gold" items for quality assurance, and had each example annotated by three crowdworkers. The task was comparably easy for crowdworkers, who reached 90%-94% pairwise agreement, and 95% success on the gold items. About 5% of items with low agreement and marked as impossible were removed. Table 3 summarizes the quantitative characteristics of all datasets resulting from this process. We took care in sampling the data as to keep similar distributions across the training, development and testing parts.

[1] www.crowdflower.com

3.2 Results

Training and development datasets were made available in advance, test data was provided during the evaluation period without labels. System performance was measured in accuracy. Since all participants provided classifications for all test items, the accuracy score is equivalent to precision/recall/F1. Participants were allowed to enter up to three different runs for evaluation. We also provide baseline accuracy scores, which are obtained by always assigning the most frequent class (figurative).

Table 4 provides the evaluation results for the known phrases task, while Table 5 ranks participants for the unseen phrases task. As expected, the *unseen phrases* setting is much harder than the *known phrases* setting, as for unseen phrases it is not possible to learn lexicalised contextual clues. In both settings, the winning entries were able to beat the MFC baseline. While performance in the *known phrases* setting is close to 80% and thus acceptable, the general task of recognizing the literal or figurative use of unseen phrases remains very challenging, with only a small improvement over the baseline. We refer to the system descriptions for more details on the techniques used for this subtask.

4 Task Conclusions

In this section, we further discuss the findings and conclusion of the evaluation challenge in the task of "Phrasal Semantics".

Looking at the results of both subtasks, one observes that the maximum performance achieved is higher for the first than the second subtask. For this comparison to be fair, trivial baselines should be taken into account. A system randomly assigning an output value would be on average 50% correct in

Rank	System	Run	Accuracy
1	IIRG	3	.779
2	UNAL	2	.754
3	UNAL	1	.722
5	IIRG	1	.530
4	*Baseline MFC*	-	.503
6	IIRG	2	.502

Table 4: Task 5b: Evaluation results for the known phrases setting

Rank	System	Run	Accuracy
1	UNAL	1	.668
2	UNAL	2	.645
3	*Baseline MFC*	-	.616
4	CLaC	1	.550

Table 5: Task 5b: Evaluation results for the unseen phrases setting

the first subtask, since the numbers of positive and negative instances in the testing set are equal. Similarly, a system assigning the most frequent class, i.e. the figurative use of any phrase, would be 50.3% and 61.6% accurate in the second subtask for seen and unseen test instances, respectively. It should also be noted that the testing instances in the first subtask are unseen in the respective training set. As a result, in terms of baselines, the second subtask on unseen data should be considered easier than the first subtask. However, the best performing systems achieved much higher accuracy in the first than in the second subtask. This contradiction confirms our conception that the first subtask is less complex than the second.

In the first subtask, it is evident that no method performs much better or much worse than the others. Although the participating systems have employed a wide variety of approaches and tools, the difference between the best and worst accuracy achieved is relatively limited, in particular approximately 14%. Even more interestingly, unsupervised approaches performed better than some supervised ones. This observation suggests that no "golden recipe" has been identified so far for this task. It is a matter of future research to investigate which components the participating systems take advantage of different sources of information and then probably proceed to the development of hybrid methods aiming at improved performance.

In the second subtask, the results of evaluation on known phrases are much higher than on unseen phrases. This was expected, as for unseen phrases it is not possible to learn lexicalised contextual clues. Thus, the second subtask has succeeded in identifying the complexity threshold upto which the current state-of-the-art can address the computational problem. Further than this threshold, i.e. for unseen phrases, current systems have not yet succeeded in addressing it. In conclusion, the difficulty in evaluating the compositionality of previously unseen phrases in context highlights the overall complexity of the second subtask.

5 Summary and Future Work

In this paper we have presented the 5^{th} task of SemEval 2013, "Evaluating Phrasal Semantics", which consists of two subtasks: (1) semantic similarity of words and compositional phrases, and (2) compositionality of phrases in context. The former subtask, which focussed on the first step of composing the meaning of phrases of any length, is less complex than the latter subtask, which considers the effect of context to the semantics of a phrase. The paper presents details about the background and importance of these subtasks, the data creation process, the systems that took part in the evaluation and their results.

In the future, we expect evaluation challenges on phrasal semantics to progress towards two directions: (a) the synthesis of semantics of sequences longer than two words, and (b) aiming to improve the performance of systems that determine the compositionality of previously unseen phrases in context. The evaluation results of the first task suggest that state-of-the-art systems can compose the semantics of two word sequences with a promising level of success. However, this task should be seen as the first step towards composing the semantics of sentence-long sequences. As far as subtask 5b is concerned, the accuracy achieved by the participating systems on unseen testing data was low, only slightly better than the most frequent class baseline, which assigns the figurative use to all test phrases.

Thus, the subtask cannot be considered well addressed by the state-of-the-art and further progress should be sought.

Acknowledgements

The work relevant to subtask 5a described in this paper is funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 318736 (OSSMETER).

We would like to thank Tristan Miller for helping with the subtleties of English idiomatic expressions, and Eugenie Giesbrecht for support in the organization of subtask 5b. This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the Hessian research excellence program Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz (LOEWE) as part of the research center *Digital Humanities*.

References

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37, Uppsala, Sweden. Association for Computational Linguistics.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Çelebi, Danyu Liu, and Elliott Drabek. 2003. Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 375–382, Morristown, NJ, USA. Association for Computational Linguistics.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a robust Part-of-Speech tagger for biomedical text. In Panayiotis Bozanis and Elias N. Houstis, editors, *Advances in Informatics*, volume 3746, chapter 36, pages 382–392. Springer Berlin Heidelberg, Berlin, Heidelberg.

Tim Van de Cruys, Thierry Poibeau, and Anna Korhonen. 2011. Latent vector weighting for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1012–1022, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary. *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, 15:60.