SemEval-2014 Task 3: Cross-Level Semantic Similarity

David Jurgens, Mohammad Taher Pilehvar, and **Roberto Navigli**
Department of Computer Science
Sapienza University of Rome
{jurgens,pilehvar,navigli}@di.uniroma1.it

Abstract

This paper introduces a new SemEval task on Cross-Level Semantic Similarity (CLSS), which measures the degree to which the meaning of a larger linguistic item, such as a paragraph, is captured by a smaller item, such as a sentence. High-quality data sets were constructed for four comparison types using multi-stage annotation procedures with a graded scale of similarity. Nineteen teams submitted 38 systems. Most systems surpassed the baseline performance, with several attaining high performance for multiple comparison types. Further, our results show that comparisons of semantic representation increase performance beyond what is possible with text alone.

1 Introduction

Given two linguistic items, semantic similarity measures the degree to which the two items have the same meaning. Semantic similarity is an essential component of many applications in Natural Language Processing (NLP), and similarity measurements between all types of text as well as between word senses lend themselves to a variety of NLP tasks such as information retrieval (Hliaoutakis et al., 2006) or paraphrasing (Glickman and Dagan, 2003).

Semantic similarity evaluations have largely focused on comparing similar types of lexical items. Most recently, tasks in SemEval (Agirre et al., 2012) and *SEM (Agirre et al., 2013) have introduced benchmarks for measuring Semantic Textual Similarity (STS) between similar-sized sentences and phrases. Other data sets such as that of Rubenstein and Goodenough (1965) measure similarity between word pairs, while the data sets of Navigli (2006) and Kilgarriff (2001) offer a binary similar-dissimilar distinction between senses. Notably, all of these evaluations have focused on comparisons between a single type, in contrast to application-based evaluations such as summarization and compositionality which incorporate textual items of different sizes, e.g., measuring the quality of a paragraph's sentence summarization.

Task 3 introduces a new evaluation where similarity is measured between items of different types: paragraphs, sentences, phrases, words and senses. Given an item of the lexically-larger type, a system measures the degree to which the meaning of the larger item is captured in the smaller type, e.g., comparing a paragraph to a sentence. We refer to this task as Cross-Level Semantic Similarity (CLSS). A major motivation of this task is to produce semantic similarity systems that are able to compare all types of text, thereby freeing downstream NLP applications from needing to consider the type of text being compared. Task 3 enables assessing the extent to which the meaning of the sentence "do u know where i can watch free older movies online without download?" is captured in the phrase "streaming vintage movies for free", or how similar is "circumscribe" to the phrase "beating around the bush." Furthermore, by incorporating comparisons of a variety of item sizes, Task 3 unifies in a single task multiple objectives from different areas of NLP such as paraphrasing, summarization, and compositionality.

Because CLSS generalizes STS to items of different types, successful CLSS systems can directly be applied to all STS-based applications. Furthermore, CLSS systems can be used in other similarity-based applications such as text simplification (Specia et al., 2012), keyphrase identification (Kim et al., 2010), lexical substitution (McCarthy and Navigli, 2009), summariza-

This work is licensed under a Creative Commons Attribution 4.0 International License. Page numbers and proceedings footer are added by the organizers. License details: http://creativecommons.org/licenses/by/4.0/

tion (Spärck Jones, 2007), gloss-to-sense mapping (Pilehvar and Navigli, 2014b), and modeling the semantics of multi-word expressions (Marelli et al., 2014) or polysemous words (Pilehvar and Navigli, 2014a).

Task 3 was designed with three main objectives. First, the task should include multiple types of comparison in order to assess each type's difficulty and whether specialized resources are needed for each. Second, the task should incorporate text from multiple domains and writing styles to ensure that system performance is robust across text types. Third, the similarity methods should be able to operate at the sense level, thereby potentially uniting text- and sense-based similarity methods within a single framework.

2 Task Description

2.1 Objective

Task 3 is intended to serve as an initial task for evaluating the capabilities of systems at measuring all types of semantic similarity, independently of the size of the text. To accomplish this objective, systems were presented with items from four comparison types: (1) paragraph to sentence, (2) sentence to phrase, (3) phrase to word, and (4) word to sense. Given a pair of items, a system must assess the degree to which the meaning of the larger item is captured in the smaller item. WordNet 3.0 was chosen as the sense inventory (Fellbaum, 1998).

2.2 Rating Scale

Following previous SemEval tasks (Agirre et al., 2012; Jurgens et al., 2012), Task 3 recognizes that two items' similarity may fall within a range of similarity values, rather than having a binary notion of similar or dissimilar. Initially a six-point (0–5) scale similar to that used in the STS tasks was considered (Agirre et al., 2012); however, annotators found difficulty in deciding between the lower-similarity options. After multiple revisions and feedback from a group of initial annotators, we developed a five-point Likert scale for rating a pair's similarity, shown in Table 1.[1]

The scale was designed to systematically order a broad range of semantic relations: synonymy, similarity, relatedness, topical association, and unrelatedness. Because items are of different sizes, the highest rating is defined as very similar rather than identical to allow for some small loss in the overall meaning. Furthermore, although the scale is designed as a Likert scale, annotators were given flexibility when rating items to use values between the defined points in the scale, indicating a blend of two relations. Table 2 provides examples of pairs for each scale rating for all four comparison type.

3 Task Data

Though several data sets exist for STS and comparing words and senses, no standard data set exists for CLSS. Therefore, we created a pilot data set designed to test the capabilities of systems in a variety of settings. The task data for all comparisons but word-to-sense was created using a three-phase process. First, items of all sizes were selected from publicly-available data sets. Second, the selected items were used to produce a second item of the next-smaller level (e.g., a sentence inspires a phrase). Third, the pairs of items were annotated for their similarity. Because of the expertise required for working with word senses, the word-to-sense data set was constructed by the organizers using a separate but similar process. In the training and test data, each comparison type had 500 annotated examples, for a total of 2000 pairs each for training and test. We first describe the corpora used by Task 3 followed by the annotation process. We then describe the construction of the word-to-sense data set.

3.1 Corpora

Test and training data were constructed by drawing from multiple publicly-available corpora and then manually generating a paired item for comparison. To achieve our second objective for the task, the data sets used to create item pairs included texts from specific domains, social media, and text with idiomatic or slang language. Table 3 summarizes the corpora and their distribution across the test and training sets for each comparison type, with a high-level description of the genre of the data. We briefly describe the corpora next.

The WikiNews, Reuters 21578, and Microsoft Research (MSR) Paraphrase corpora are all drawn from newswire text, with WikiNews being authored by volunteer writers and the latter two corpora written by professionals. Travel Guides was drawn from the Berlitz travel guides data in the Open American National Corpus (Ide and Suderman, 2004) and includes very verbose sentences

[1] Annotation materials along with all training and test data are available on the task website `http://alt.qcri.org/semeval2014/task3/`.

4 – Very Similar	The two items have very similar meanings and the most important ideas, concepts, or actions in the larger text are represented in the smaller text. Some less important information may be missing, but the smaller text is a very good summary of the larger text.
3 – Somewhat Similar	The two items share many of the same important ideas, concepts, or actions, but include slightly different details. The smaller text may use similar but not identical concepts (e.g., car vs. vehicle), or may omit a few of the more important ideas present in the larger text.
2 – Somewhat related but not similar	The two items have dissimilar meaning, but share concepts, ideas, and actions that are related. The smaller text may use related but not necessarily similar concepts (window vs. house) but should still share some overlapping concepts, ideas, or actions with the larger text.
1 – Slightly related	The two items describe dissimilar concepts, ideas and actions, but may share some small details or domain in common and might be likely to be found together in a longer document on the same topic.
0 – Unrelated	The two items do not mean the same thing and are not on the same topic.

Table 1: The five-point Likert scale used to rate the similarity of item pairs. See Table 2 for examples.

with many named entities. Wikipedia Science was drawn from articles tagged with the category *Science* on Wikipedia. Food reviews were drawn from the SNAP Amazon Fine Food Reviews data set (McAuley and Leskovec, 2013) and are customer-authored reviews for a variety of food items. Fables were taken from a collection of Aesop's Fables. The Yahoo! Answers corpus was derived from the Yahoo! Answers data set, which is a collection of questions and answers from the Community Question Answering (CQA) site; the data set is notable for having the highest degree of ungrammaticality in our test set. SMT Europarl is a collection of texts from the English-language proceedings of the European parliament (Koehn, 2005); Europarl data was also used in the PPDB corpus (Ganitkevitch et al., 2013), from which phrases were extracted. Wikipedia was used to generate two phrase data sets from (1) extracting the definitional portion of an article's initial sentence, e.g., "An [article name] is a [definition]," and (2) captions for an article's images. Web queries were gathered from online sources of real-world queries. Last, the first and second authors generated slang and idiomatic phrases based on expressions contained in Wiktionary.

For all comparison types, the test data included one genre that was not seen in the training data in order to test the generalizability of the systems on data from a novel domain. In addition, we included a new type of challenge genre with Fables; unlike other domains, the sentences paired with the fable paragraphs were potentially semantic interpretations of the intent of the fable, i.e., the moral of the story. These interpretations often have little textual overlap with the fable itself and require a deeper interpretation of the paragraph's meaning in order to make the correct similarity judgment.

Prior to the annotation process, all content was filtered to ensure its size and format matched the desired text type. By average, a paragraph in our dataset consists of 3.8 sentences. Typos and grammatical mistakes in the community-produced content were left unchanged.

3.2 Annotation Process

A two-phase process was used to produce the test and training data sets for all but word-to-sense. Phase 1 generates the item pairs from source texts and Phase 2 rates the pairs' similarity.

Phase 1 In this phase, annotators were shown the larger text of a comparison type and then asked to produce the smaller text of the pair at a specified similarity; for example an annotator may be shown a paragraph and asked to write a sentence that is a "3" rating. Annotators were instructed to leave the smaller text blank if they had difficulty understanding the larger text.

The requested similarity ratings were balanced to create a uniform distribution of similarity values. Annotators were asked only to generate ratings of 1–4; pairs with a "0" rating were automatically created by pairing the larger item with random selections of text of the appropriate size from the same corpus. The intent of Phase 1 is to produce varied item pairs with an expected uniform distribution of similarity values along the rating scale.

Four annotators participated in Phase 1 and were paid a bulk rate of €110 for completing the work. In addition to the four annotators, the first two organizers also assisted in Phase 1: Both completed items from the SCIENTIFIC genre and the first organizer produced 994 pairs, including all

PARAGRAPH TO SENTENCE

Paragraph: Teenagers take aerial shots of their neighbourhood using digital cameras sitting in old bottles which are launched via kites - a common toy for children living in the favelas. They then use GPS-enabled smartphones to take pictures of specific danger points - such as rubbish heaps, which can become a breeding ground for mosquitoes carrying dengue fever.

Rating	Sentence
4	Students use their GPS-enabled cellphones to take birdview photographs of a land in order to find specific danger points such as rubbish heaps.
3	Teenagers are enthusiastic about taking aerial photograph in order to study their neighbourhood.
2	Aerial photography is a great way to identify terrestrial features that aren't visible from the ground level, such as lake contours or river paths.
1	During the early days of digital SLRs, Canon was pretty much the undisputed leader in CMOS image sensor technology.
0	Syrian President Bashar al-Assad tells the US it will "pay the price" if it strikes against Syria.

SENTENCE TO PHRASE

Sentence: Schumacher was undoubtedly one of the very greatest racing drivers there has ever been, a man who was routinely, on every lap, able to dance on a limit accessible to almost no-one else.

Rating	Phrase
4	the unparalleled greatness of Schumacher's driving abilities
3	driving abilities
2	formula one racing
1	north-south highway
0	orthodontic insurance

PHRASE TO WORD

Phrase: loss of air pressure in a tire

Rating	Word
4	flat-tire
3	deflation
2	wheel
1	parking
0	butterfly

WORD TO SENSE

Word: automobile$_n$

Rating	Sense
4	car^1_n (a motor vehicle with four wheels; usually propelled by an internal combustion engine)
3	vehicle^1_n (a conveyance that transports people or objects)
2	bike^1_n (a motor vehicle with two wheels and a strong frame)
1	highway^1_n (a major road for any form of motor transport)
0	pen^1_n (a writing implement with a point from which ink flows)

Table 2: Example pairs and their ratings.

those for the METAPHORIC genre, and those that the other annotators left blank.

Phase 2 Here, the item pairs produced in Phase 1 were rated for their similarity according to the scale described in Section 2.2. An initial pilot study showed that crowdsourcing was only moderately effective for producing these ratings with high agreement. Furthermore, the texts used in Task 3 came from a variety of genres, such as scientific domains, which some workers had difficulty understanding. While we note that crowdsourcing has been used in prior STS tasks for generating similarity scores (Agirre et al., 2012; Agirre et al., 2013), both tasks' efforts encountered lower worker score correlations on some portions of the dataset (Diab, 2013), suggesting that crowdsourcing may not be reliable for judging the similarity of certain types of text. See Section 3.5 for additional details.

Therefore, to ensure high quality, the first two organizers rated all items independently. Because the sentence-to-phrase and phrase-to-word comparisons contain slang and idiomatic language, a third American English mother tongue annotator was added for those data sets. The third annotator was compensated €250 for their assistance.

Annotators were allowed to make finer-grained distinctions in similarity using multiples of 0.25. For all items, when any two annotators disagreed by one or more scale points, we performed an adjudication to determine the item's rating in the gold standard. The adjudication process revealed that nearly all disagreements were due to annotator mistakes, e.g., where one annotator had overlooked a part of the text or had misunderstood the text's meaning. The final similarity rating for an unadjudicated item was the average of its ratings.

3.3 Word-to-Sense

Word-to-sense comparison items were generated in three phases. To increase the diversity and challenge of the data set, the word-to-sense was created for four types of words: (1) a word and its intended meaning are in WordNet, (2) a word was not in the WordNet vocabulary, e.g., the verb "zombify," (3) the word is in WordNet, but has a novel meaning that is not in WordNet, e.g., the adjective "red" referring to Communist, and (4) a set of challenge words where one of the word's senses and a second sense are directly connected by an edge in the WordNet network, but the two senses are not always highly similar.

		Paragraph-to-Sentence		Sentence-to-Phrase		Phrase-to-Word	
Corpus	Genre	Train	Test	Train	Test	Train	Test
WikiNews	Newswire	15.0	10.0	9.2	6.0		
Reuters 21578	Newswire	20.2	15.0			5.0	
Travel Guides	Travel	15.2	10.0	15.0	9.8		
Wikipedia Science	Scientific	–	25.6	–	14.8		
Food Reviews	Review	19.6	20.0				
Fables	Metaphoric	9.0	5.2				
Yahoo! Answers	CQA	21.0	14.2	17.6	17.4		
SMT Europarl	Newswire			35.4	14.4		
MSR Paraphrase	Newswire			10.0	10.0	8.8	6.0
Idioms	Idiomatic			12.8	12.6	20.0	20.0
Slang	Slang			–	15.0	–	25.0
PPDB	Newswire					10.0	10.0
Wikipedia Glosses	Lexicographic					28.2	17.0
Wikipedia Image Captions	Descriptive					23.0	17.0
Web Search Queries	Search					5.0	5.0

Table 3: Percentages of the training and test data per source corpus.

In Phase 1, to select the first type of word, lemmas in WordNet were ranked by frequency in Wikipedia; the ranking was divided into ten equally-sized groups, with words sampled evenly from groups in order to control for word frequency in the task data. For the second type, words not present in WordNet were drawn from two sources: examining words in Wikipedia, which we refer to as out-of-vocabulary (OOV), and slang words. For the third type, to identify words with a novel sense, we examined Wiktionary entries and chose novel, salient senses that were distinct from those in WordNet. We refer to words with a novel meaning as out-of-sense (OOS). Words of the fourth type were chosen by hand. The part-of-speech distributions for all four types of items were balanced as 50% noun, 25% verb, 25% adjective.

In Phase 2, each word was associated with a particular WordNet sense for its intended meaning, or the closest available sense in WordNet for OOV or OOS items. To select a comparison sense, we adopted a neighborhood search procedure: All synsets connected by at most three edges in the WordNet semantic network were shown. Given a word and its neighborhood, the corresponding sense for the item pair was selected by matching the sense with an intended similarity for the pair, much like how text items were generated in Phase 1. The reason behind using this neighborhood-based selection process was to minimize the potential bias of consistently selecting lower-similarity items from those further away in the WordNet semantic network.

In Phase 3, given all word-sense pairs, annotators were shown the definitions associated with the intended meaning of the word and of the sense. Definitions were drawn from WordNet or from Wiktionary, if the word was OOV or OOS. Annotators had access to the WordNet structure for the compared sense in order to take into account its parents and siblings.

3.4 Trial Data

The trial data set was created using a separate process. Source text was drawn from WikiNews; we selected the text for the larger item of each level and then generated the text or sense of the smaller. A total of 156 items were produced. After, four fluent annotators independently rated all items. Inter-annotator agreement rates varied in 0.734–0.882, using Krippendorff's α (Krippendorff, 2004) on the interval scale.

3.5 Data Set Discussion

The resulting annotation process produced a high-quality data set. First, Table 4 shows the inter-annotator agreement (IAA) statistics for each comparison type on both the full and unadjudicated portions of the data set. IAA was measured using Krippendorff's α for interval data. Because the disagreements that led to lower α in the full data were resolved via adjudication, the quality of the full data set is expected to be on par with that of the unadjudicated data. The annotation quality for Task 3 was further improved by manually adjudicating all significant disagreements.

In contrast, the data sets of current STS tasks aggregated data from annotators with moderate correlation with each other (Diab, 2013); STS-2012 (Agirre et al., 2012) saw inter-annotator Pearson correlations of 0.530–0.874 per data set and STS-2013 (Agirre et al., 2013) had average

	Training		Test	
Data	All	Unadj.	All	Unadj.
Para.-to-Sent.	0.856	0.916	0.904	0.971
Sent.-to-Phr.	0.773	0.913	0.766	0.980
Phr.-to-Word	0.735	0.895	0.730	0.988
Word-to-Sense	0.681	0.895	0.655	0.952

Table 4: IAA rates for the task data.

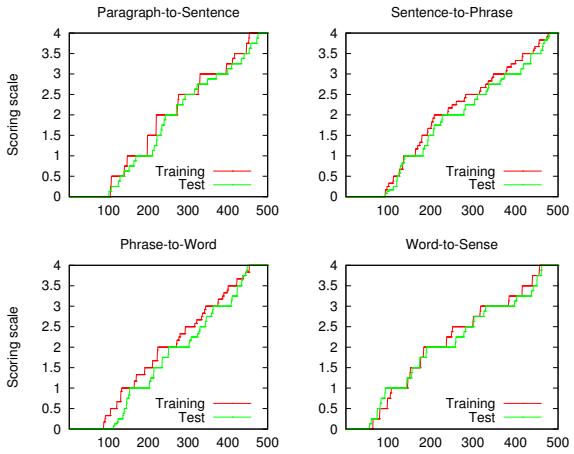

Figure 1: Similarity ratings distributions.

inter-annotator correlations of 0.377–0.832. However, we note that Pearson correlation and Krippendorff's α are not directly comparable (Artstein and Poesio, 2008), as annotators' scores may be correlated, but completely disagree.

Second, the two-phase construction process produced values that were evenly distributed across the rating scale, shown in Figure 1 as the distribution of the values for all data sets. However, we note that this creation procedure was very resource intensive and, therefore, semi-automated or crowdsourcing-based approaches for producing high-quality data will be needed to expand the size of the data in future CLSS-based evaluations. Nevertheless, as a pilot task, the manual effort was essential for ensuring a rigorously-constructed data set for the initial evaluation.

4 Evaluation

Participation The ultimate goal of Task 3 is to produce systems that can measure similarity for multiple types of items. Therefore, we strongly encouraged participating teams to submit systems that were capable of generating similarity judgments for multiple comparison types. However, to further the analysis, participants were also permitted to submit systems specialized to a single domain. Teams were allowed at most three system submissions, regardless of the number of comparison types supported.

Scoring Systems were required to provide similarity values for all items within a comparison type. Following prior STS evaluations, systems were scored for each comparison type using Pearson correlation. Additionally, we include a second score using Spearman's rank correlation, which is only affected by differences in the ranking of items by similarity, rather than differences in the similarity values. Pearson correlation was chosen as the official evaluation metric since the goal of the task is to produce similar scores. However, Spearman's rank correlation provides an important metric for assessing systems whose scores do not match human scores but whose rankings might, e.g., string-similarity measures. Ultimately, a global ranking was produced by ordering systems by the sum of their Pearson correlation values for each of the four comparison levels.

Baselines The official baseline system was based on the Longest Common Substring (LCS), normalized by the length of items using the method of Clough and Stevenson (2011). Given a pair, the similarity is reported as the normalized length of the LCS. In the case of word-to-sense, the LCS for a word-sense pair is measured between the sense's definition in WordNet and the definitions of each sense of the pair's word, reporting the maximal LCS. Because OOV and slang words are not in WordNet, the baseline reports the average similarity value of non-OOV items. Baseline scores were made public after the evaluation period ended.

Because LCS is a simple procedure, a second baseline based on Greedy String Tiling (GST) (Wise, 1996) was added after the evaluation period concluded. Unlike LCS, GST better handles the transpositions of tokens across the two texts and can still report high similarity when encountering reordered text. The minimum match length for GST was set to 6.

5 Results

Nineteen teams submitted 38 systems. Of those systems, 34 produced values for paragraph-to-sentence and sentence-to-phrase comparisons, 22 for phrase-to-word, and 20 for word-to-sense. Two teams submitted revised scores for their systems after the deadline but before the test set had

Team	System	Para-2-Sent	Sent-2-Phr	Phr-2-Word	Word-2-Sense	Official Rank	Spearman Rank
Meerkat Mafia	pairingWords†	0.794	0.704	**0.457**	**0.389**		
SimCompass	run1	0.811	0.742	0.415	0.356	1	1
ECNU	run1	0.834	0.771	0.315	0.269	2	2
UNAL-NLP	run2	0.837	0.738	0.274	0.256	3	6
SemantiKLUE	run1	0.817	0.754	0.215	0.314	4	4
UNAL-NLP	run1	0.817	0.739	0.252	0.249	5	7
UNIBA	run2	0.784	0.734	0.255	0.180	6	8
RTM-DCU	run1†	**0.845**	0.750	0.305			
UNIBA	run1	0.769	0.729	0.229	0.165	7	10
UNIBA	run3	0.769	0.729	0.229	0.165	8	11
BUAP	run1	0.805	0.714	0.162	0.201	9	13
BUAP	run2	0.805	0.714	0.142	0.194	10	9
Meerkat Mafia	pairingWords	0.794	0.704	-0.044	0.389	11	12
HULTECH	run1	0.693	0.665	0.254	0.150	12	16
GST Baseline		0.728	0.662	0.146	0.185		
HULTECH	run3	0.669	0.671	0.232	0.137	13	15
RTM-DCU	run2†	0.785	0.698	0.221			
RTM-DCU	run3	0.780	0.677	0.208		14	17
HULTECH	run2	0.667	0.633	0.180	0.169	15	14
RTM-DCU	run1	0.786	0.666	0.171		16	18
RTM-DCU	run3†	0.786	0.663	0.171			
Meerkat Mafia	SuperSaiyan	0.834	**0.777**			17	19
Meerkat Mafia	Hulk2	0.826	0.705			18	20
RTM-DCU	run2	0.747	0.588	0.164		19	22
FBK-TR	run3	0.759	0.702			20	23
FBK-TR	run1	0.751	0.685			21	24
FBK-TR	run2	0.770	0.648			22	25
Duluth	Duluth2	0.501	0.450	0.241	0.219	23	21
AI-KU	run1	0.732	0.680			24	26
LCS Baseline		0.527	0.562	0.165	0.109		
UNAL-NLP	run3	0.708	0.620			25	27
AI-KU	run2	0.698	0.617			26	28
TCDSCSS	run2	0.607	0.552			27	29
JU-Evora	run1	0.536	0.442	0.090	0.091	28	31
TCDSCSS	run1	0.575	0.541			29	30
Duluth	Duluth1	0.458	0.440	0.075	0.076	30	5
Duluth	Duluth3	0.455	0.426	0.075	0.079	31	3
OPI	run1		0.433	0.213	0.152	32	36
SSMT	run1	0.789				33	34
DIT	run1	0.785				34	32
DIT	run2	0.784				35	33
UMCC DLSI SelSim	run1		0.760			36	35
UMCC DLSI SelSim	run2		0.698			37	37
UMCC DLSI Prob	run1				0.023	38	38

Table 5: Task results. Systems marked with a † were submitted after the deadline but are positioned where they would have ranked.

been released. These systems were scored and noted in the results but were not included in the official ranking.

Table 5 shows the performance of the participating systems across all the four comparison types in terms of Pearson correlation. The two right-most columns show system rankings by Pearson (Official Rank) and Spearman's ranks correlation.

The SimCompass system attained first place, partially due to its superior performance on phrase-to-word comparisons, providing an improvement of 0.10 over the second-best system. The late-submitted version of the Meerkat Mafia pairingWords† system corrected a bug in the phrase-to-word comparison, which ultimately would have attained first place due to large performance improvements over SimCompass on phrase-to-word and word-to-sense. ENCU and UNAL-NLP systems rank respectively second and third while the former being always in top-4 and the latter being among the top-7 systems across the four comparison types. Most systems were able to surpass the naive LCS baseline; however, the more sophisticated GST baseline (which accounts for text transposition) outperforms two-thirds of the systems. Importantly, both baselines perform

poorly on smaller text, highlighting the importance of performing a *semantic* comparison, as opposed to a string-based one.

Within the individual comparison types, specialized systems performed well for the larger text sizes. In the paragraph-to-sentence type, the run1 system of UNAL-NLP provides the best official result, with the late RTM-DCU run1† system surpassing its performance slightly. Meerkat Mafia provides the best performance in sentence-to-phrase with its SuperSaiyan system and the best performances in phrase-to-word and word-to-sense with its late pairingWords† system.

Comparison-Type Analysis Performance across the comparison types varied considerably, with systems performing best on comparisons between longer textual items. As a general trend, both the baselines' and systems' performances tend to decrease with the size of lexical items in the comparison types. A main contributing factor to this is the reliance on textual similarity measures (such as the baselines), which perform well when two items' may share content. However, as the items' content becomes smaller, e.g., a word or phrase, the textual similarity does not necessarily provide a meaningful indication of the *semantic* similarity between the two. This performance discrepancy suggests that, in order to perform well, CLSS systems must rely on comparisons between semantic representations rather than textual representations. The two top-performing systems on these smaller levels, Meerkat Mafia and SimCompass, used additional resources beyond WordNet to expand a word or sense to its definition or to represent words with distributional representations.

Per-genre results and discussions Task 3 includes multiple genres within the data set for each comparison type. Figure 2 shows the correlation of each system for each of these genres, with systems ordered left to right according to their official ranking in Table 5. An interesting observation is that a system's official rank does not always match the rank from aggregating its correlations for each genre individually. This difference suggests that some systems provided good similarity judgments on individual genres, but their range of similarity values was not consistent between genres leading to lower overall Pearson correlation. For instance, in the phrase-to-word comparison type, the aggregated per-genre performance of Duluth-1 and Duluth-3 are among the best whereas their overall Pearson performance puts these systems among the worst-performing ones in the comparison type.

Among the genres, CQA, SLANG, and IDIOMATIC prove to be the more difficult for systems to interpret and judge. These genres included misspelled, colloquial, or slang language which required converting the text into semantic form in order to meaningfully compare it. Furthermore, as expected, the METAPHORIC genre was the most difficult, with no system performing well; we view the METAPHORIC genre as an open challenge for future systems to address when interpreting larger text. On the other hand, SCIENTIFIC, TRAVEL, and NEWSWIRE tend to be the easiest genres for paragraph-to-sentence and sentence-to-phrase. All three genres tend to include many named entities or highly-specific language, which are likely to be more preserved in the more-similar paired items. Similarly, DESCRIPTIVE and SEARCH genres were easiest in phrase-to-word, which also often featured specific words that were preserved in highly-similar pairs. In the case of word-to-sense, REGULAR proves to be the least difficult genre. Interestingly, in word-to-sense, most systems attained moderate performance for comparisons with words not in WordNet (i.e., OOV) but had poor performance for slang words, which were also OOV. This difference suggests that systems could be improved with additional semantic resources for slang.

Spearman Rank Analysis Although the goal of Task 3 is to have systems produce similarity judgments, some applications may benefit from simply having a ranking of pairs, e.g., ranking summarizations by goodness. The Spearman rank correlation measures the ability of systems to perform such a ranking. Surprisingly, with the Spearman-based ranking, the Duluth1 and Duluth3 systems attain the third and fifth ranks – despite being among the lowest ranked with Pearson. Both systems were unsupervised and produced similarity values that did not correlate well with those of humans. However, their Spearman ranks demonstrate the systems ability to correctly identify relative similarity and suggests that such unsupervised systems could improve their Pearson correlation by using the training data to tune the range of similarity values to match those of humans.

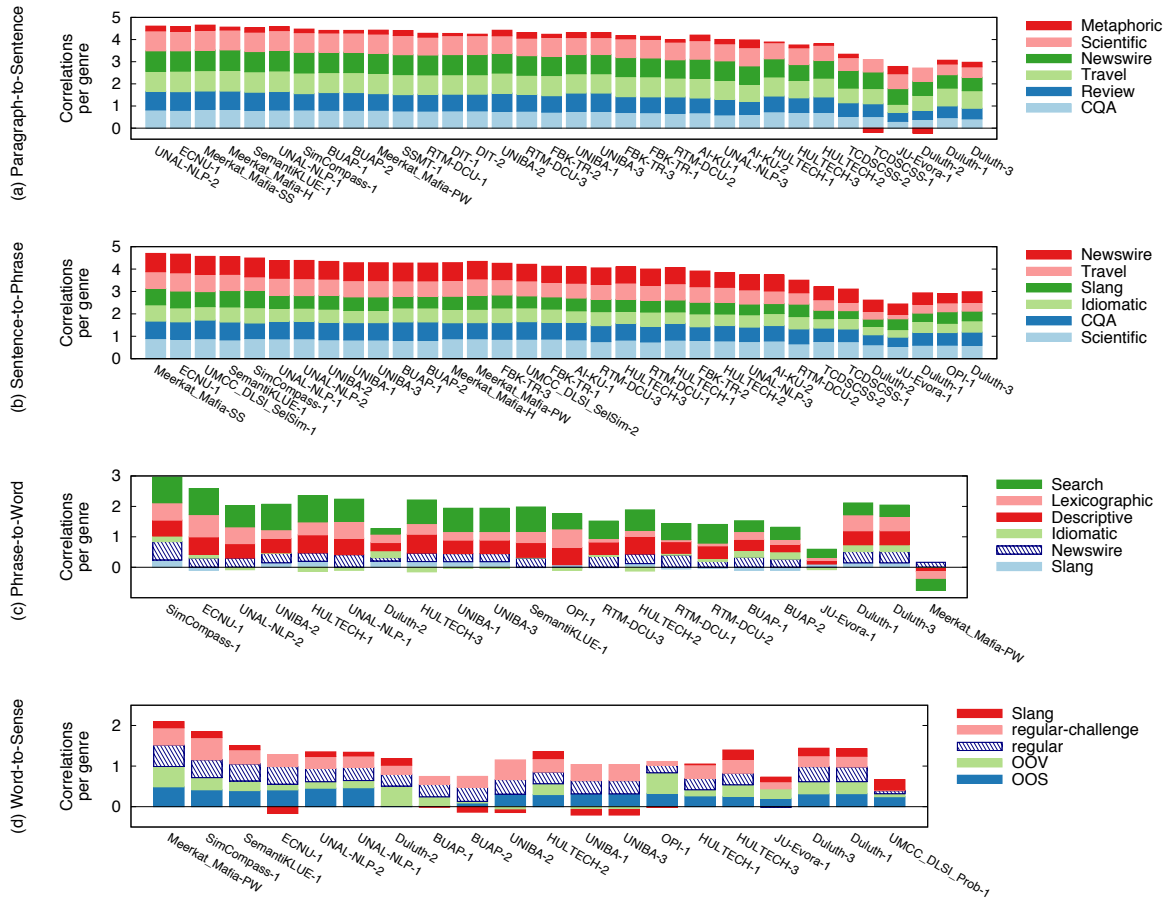

Figure 2: A stacked histogram for each system, showing its Pearson correlations for genre-specific portions of the gold-standard data, which may also be negative.

6 Conclusion

This paper introduces a new similarity task, Cross-Level Semantic Similarity, for measuring the semantic similarity of lexical items of different sizes. Using a multi-phase annotation procedure, we have produced a high-quality data set of 4000 items comprising of various genres, evenly-split between training and test with four types of comparison: paragraph-to-sentence, sentence-to-phrase, phrase-to-word, and word-to-sense. Nineteen teams submitted 38 systems, with most teams surpassing the baseline system and several systems achieving high performance in multiple types of comparison. However, a clear performance trend emerged where systems perform well only when the text itself is similar, rather than its underlying meaning. Nevertheless, the results of Task 3 are highly encouraging and point to clear future objectives for developing CLSS systems that operate on more semantic representations rather than text. In future work on CLSS evaluation, we first intend to develop scalable annotation methods to increase the data sets. Second, we plan to add new evaluations where systems are tested according to their performance in an application related to each comparison-type, such as measuring the quality of a paraphrase or summary.

Acknowledgments

We would like to thank Tiziano Flati, Marc Franco Salvador, Maud Erhmann, and Andrea Moro for their help in preparing the trial data; Gaby Ford, Chelsea Smith, and Eve Atkinson for their help in generating the training and test data; and Amy Templin for her help in generating and rating the training and test data.

The authors gratefully acknowledge the support of the ERC Starting Grant Multi-JEDI No. 259234.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval-2012)*, pages 385–393, Montréal, Canada.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 Shared Task: Semantic textual similarity, including a pilot on typed-similarity. In *Proceedings of the Second Joint Confer-*

*ence on Lexical and Computational Semantics (*SEM)*, Atlanta, Georgia.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Paul Clough and Mark Stevenson. 2011. Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, 45(1):5–24.

Mona Diab. 2013. Semantic textual similarity: past present and future. In *Joint Symposium on Semantic Processing*. Keynote address. `http://jssp2013.fbk.eu/sites/jssp2013.fbk.eu/files/Mona.pdf`.

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

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 758–764, Atlanta, Georgia.

Oren Glickman and Ido Dagan. 2003. Acquiring lexical paraphrases from a single corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 81–90, Borovets, Bulgaria.

Angelos Hliaoutakis, Giannis Varelas, Epimenidis Voutsakis, Euripides GM Petrakis, and Evangelos Milios. 2006. Information retrieval by semantic similarity. *International Journal on Semantic Web and Information Systems*, 2(3):55–73.

Nancy Ide and K. Suderman. 2004. The American National Corpus First Release. In *Proceedings of the* 4^{th} *Language Resources and Evaluation Conference (LREC)*, pages 1681–1684, Lisbon, Portugal.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 Task 2: Measuring Degrees of Relational Similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval-2012)*, pages 356–364, Montréal, Canada.

Adam Kilgarriff. 2001. English lexical sample task description. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, pages 17–20, Toulouse, France.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*, pages 21–26, Los Angeles, California.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, second edition.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908, Rio de Janeiro, Brazil.

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost Word Sense Disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 105–112, Sydney, Australia.

Mohammad Taher Pilehvar and Roberto Navigli. 2014a. A large-scale pseudoword-based evaluation framework for state-of-the-art Word Sense Disambiguation. *Computational Linguistics*, 40(4).

Mohammad Taher Pilehvar and Roberto Navigli. 2014b. A robust approach to aligning heterogeneous lexical resources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 468–478, Baltimore, USA.

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

Karen Spärck Jones. 2007. Automatic summarising: The state of the art. *Information Processing and Management*, 43(6):1449–1481.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. SemEval-2012 Task 1: English Lexical Simplification. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval-2012)*, pages 347–355. Association for Computational Linguistics.

Michael J. Wise. 1996. YAP3: Improved detection of similarities in computer program and other texts. In *Proceedings of the twenty-seventh SIGCSE technical symposium on Computer science education*, SIGCSE '96, pages 130–134, Philadelphia, Pennsylvania, USA.