Lang Resources & Evaluation
DOI 10.1007/s10579-015-9318-3

ORIGINAL PAPER

Cross level semantic similarity: an evaluation framework for universal measures of similarity

David Jurgens[1] · Mohammad Taher Pilehvar[2] · Roberto Navigli[2]

Abstract Semantic similarity has typically been measured across items of approximately similar sizes. As a result, similarity measures have largely ignored the fact that different types of linguistic item can potentially have similar or even identical meanings, and therefore are designed to compare only one type of linguistic item. Furthermore, nearly all current similarity benchmarks within NLP contain pairs of approximately the same size, such as word or sentence pairs, preventing the evaluation of methods that are capable of comparing different sized items. To address this, we introduce a new semantic evaluation called 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. Our pilot CLSS task was presented as part of SemEval-2014, which attracted 19 teams who submitted 38 systems. CLSS data contains a rich mixture of pairs, spanning from paragraphs to word senses to fully evaluate similarity measures that are capable of comparing items of any type. Furthermore, data sources were drawn from diverse corpora beyond just newswire, including domain-specific texts and social media. We describe the annotation process and its challenges, including a comparison with crowdsourcing, and identify the factors that make the dataset a rigorous assessment of a method's quality. Furthermore, we examine in detail the systems participating in the SemEval task to identify the common factors associated

✉ David Jurgens
jurgens@cs.mcgill.ca

Mohammad Taher Pilehvar
pilehvar@di.uniroma1.it

Roberto Navigli
navigli@di.uniroma1.it

[1] McGill University, Montreal, Canada

[2] Sapienza University of Rome, Rome, Italy

Published online: 12 October 2015

with high performance and which aspects proved difficult to all systems. Our findings demonstrate that CLSS poses a significant challenge for similarity methods and provides clear directions for future work on universal similarity methods that can compare any pair of items.

Keywords Similarity · Evaluation · Semantic textual similarity

1 Introduction

Semantic similarity measures the degree to which two linguistic items have the same meaning. Accurately measuring semantic similarity is an essential component of many applications in natural language processing (NLP), such as ontology learning, thesauri generation, and even machine translation evaluation. Therefore, multiple evaluations have been proposed for testing these computational approaches on their ability to accurately measure similarity, e.g., the RG-65 dataset (Rubenstein and Goodenough 1965) or the TOEFL synonym test (Landauer and Dumais 1997).

Semantic similarity evaluations have largely focused on comparing similar types of linguistic items. Most recently, a large amount of work has focused on semantic textual similarity (STS) (Agirre et al. 2012, 2013, 2014), which measures the similarity between similar-sized sentences and phrases. However, other widely-used semantic similarity evaluation datasets have been built around word similarity (Rubenstein and Goodenough 1965; Finkelstein et al. 2001) and associativity (Finkelstein et al. 2001); and, furthermore, a few works have proposed datasets based on identifying similar-dissimilar distinctions between a word's senses (Snow et al. 2007; Navigli 2006; Kilgarriff 2001). Notably, all of these evaluations have focused on comparisons between similar types of entity, e.g., comparing words, in contrast to the uses of semantic similarity in applications such as summarization and compositionality which compare entities of different sizes, e.g., measuring the similarity between a multi-word expression's meaning and a single word's meaning.[1]

To address this broader class of semantic similarity comparisons between textual items of different sizes, we introduce a new evaluation where similarity is measured between items of five different types: paragraphs, sentences, phrases, words and senses. Given an item of the lexically-larger type, a system is tasked with measuring 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). Our pilot CLSS task was presented as part of SemEval-2014 (Jurgens et al. 2014).

A major motivation of this task is to produce semantic similarity systems that report meaningful similarity scores for *all* types of input, thereby freeing downstream NLP applications from needing to consider the type of text being

[1] A notable exception are benchmarks in Information Retrieval where a relatively-short query is paired with full documents. Although these items are often compared using a common representation like a vector space, the interpretation of the comparison is not similarity, but rather relatedness.

compared. For example, CLSS measures 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 "circumscribe" is to the phrase "beating around the bush." Furthermore, by incorporating comparisons of a variety of item sizes, the evaluation 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 type, 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), summarization (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).

The proposed CLSS task 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 Related work

Effectively measuring semantic similarity is a long-standing objective of NLP and related fields, with most work focusing on datasets that measure the similarity of the same type of linguistic item (e.g., words or sentences). Most related to the CLSS objective are the works on Semantic Textual Similarity for phrases and sentences. Dolan et al. (2004) and Li et al. (2006) initially proposed two sentence similarity datasets focused on evaluating the quality of paraphrases, containing on the order of a hundred pairs for comparison. Most recently, STS tasks proposed as a part of SemEval have provided relatively large training and tests sets primarily for sentence similarity, with some phrases also included (Agirre et al. 2012, 2013, 2014). In these three tasks, sentence pairs were rated on a scale from zero (completely unrelated) to five (semantically identical). Sentence pairs were drawn from multiple corpora primarily consisting of paraphrases, newswire, and video descriptions. Notably, because of differences in source corpora, similarity rating distributions were not evenly distributed across the rating scales for each corpus.

The data used by these STS evaluations differs in two key ways from that used in the SemEval-2014 CLSS task (hereafter, CLSS-2014). First, with the exception of the 2014 STS task, which included social media text from Twitter, the source domains have focused on a small number of text genres that are widely supported by NLP tools. In contrast, our CLSS data includes data from a diverse selection of genres including (1) social media genres that are likely to contain numerous spelling

and grammatical mistakes, (2) genres that include idiomatic or metaphoric language, and (3) domain-specific genres that include phrases or words not present in many semantic resources. Although humans have little difficulty in rating similarity in these various genres (cf. Sect. 4.6), these genre differences create a more challenging evaluation setting for computational approaches and ensure that system performance is likely to generalize across a large number of domains.

The second key difference in the data of our CLSS task from the STS tasks is in the sizes of items being compared. Table 1 illustrates this by showing the average lengths of source and target items being compared, in terms of their number of content words. Nearly all pairs used in STS tasks have identical sizes (a ratio close to 1.0), with the exception of the FNWN subset from STS-2013 which is obtained from the definitions of manually mapped sense pairs of FrameNet 1.5 and WordNet 3.1 where the average gloss length is about three times larger in the former sense inventory.[2] In contrast, the average pair size ratio is about three in our datasets, ranging from 2.7 (sentence-to-phrase test set) to 4.1 (phrase-to-word training set). In addition to the size difference, the pairs in the STS datasets are different to those in the CLSS datasets as the two sides belong to the same lexical level in the former (usually a sentence) whereas in the latter they belong to two different lexical levels, e.g., sentence and phrase, which yield different syntactic structures.

Work on semantic similarity resources has also focused on comparing word meanings, which closely relates to the phrase-to-word similarity judgments of our CLSS task. Rubenstein and Goodenough (1965) propose a dataset of 65 word pairs rated by 51 human annotators on a scale from 0–4 for their similarity. Finkelstein et al. (2001) created the widely-used WordSim-353 dataset; however, the annotation process for this dataset conflated similarity and relatedness, leading to high similarity scores for pairs such as computer-keyboard or smart-stupid despite the dissimilarity in their meanings. Separating similarity from relatedness is essential for many semantic tasks, such as recognizing synonyms or appropriate paraphrases. To correct the conflation, Agirre et al. (2009) partitioned the WordSim-353 dataset into two subsets: one containing related pairs and a second containing similar pairs; however, as Hill et al. (2014) note, evaluation with these datasets is still problematic as (1) the annotation values were not gathered specific to similarity alone, so the rating values are difficult to interpret, and (2) the evaluation of a similarity-measuring system also required testing that the system does not highly rate the pairs in the relatedness-based dataset. Most recently, Hill et al. (2014) propose a new dataset, SimLex-999, which uses a revised annotation procedure to elicit only similarity judgments.

Our CLSS comparison differs in two key ways from these word comparison tasks. First, our rating scale (discussed later in Sect. 3) explicitly captures the differences between relatedness and synonymy and required annotators and systems to distinguish between the two. Second, our evaluation recognizes that multiword expressions and phrases can be similar or even synonymous with single words, e.g., "a very large, expensive house" and "mansion." Given that a semantic similarity

[2] We calculate an item's length in terms of the number of its content words, i.e., nouns, verbs, adjectives, and adverbs.

Table 1 Sizes of the items compared in semantic textual similarity and cross-level semantic similarity SemEval datasets

Task	Avg. source size (in words)	Avg. target size (in words)	Avg. size proportion
CLSS-2014 (Jurgens et al. 2014)			
Paragraph-to-sentence (training)	41.736	12.134	3.440
Paragraph-to-sentence (test)	43.158	11.978	3.603
Sentence-to-phrase (training)	12.194	3.582	3.582
Sentence-to-phrase (test)	10.754	4.002	2.687
Phrase-to-word (training)	4.222	1.032	4.091
Phrase-to-word (test)	3.438	1.068	3.219
STS-2014 (Agirre et al. 2014)			
Deft-forum	5.338	5.382	0.992
Deft-news	10.870	10.507	1.035
Headlines	5.792	5.933	0.976
Images	5.304	5.324	0.996
Tweet-news	8.508	5.971	1.425
OnWN	4.436	4.532	0.979
Average	6.708	6.275	1.067
STS-2013 (Agirre et al. 2013)			
FNWN	16.857	5.614	3.003
Headlines	5.719	5.683	1.006
OnWN	4.164	4.109	1.013
SMT	14.011	14.484	0.967
Average	10.188	7.472	1.497
STS-2012 (Agirre et al. 2012)			
OnWN	4.317	4.737	0.911
SMTnews	7.439	7.173	1.037
MSRpar (train)	11.446	11.408	1.003
MSRpar (test)	11.089	11.211	0.989
MSRvid (train)	4.373	4.337	1.008
MSRvid (test)	4.339	4.349	0.998
SMTeuroparl (train)	15.343	14.745	1.041
SMTeuroparl (test)	6.444	6.187	1.042
Average	8.099	8.018	1.004

measure is ideally expected to model a phrase as a whole, and not as a combination of the individual models of its constituent words, our framework provides an evaluation benchmark to bridge from semantic representation of individual words to that of phrases.

Word similarity datasets come with an implicit disambiguation where annotators must identify the concepts to which the words refer and compare those. In contrast, a small number of works have used sense-based data to explicitly mark the concepts

being compared. Most related to our CLSS work are that of Kilgarriff (2001) and Navigli (2006), whose datasets reflect sense similarity judgments on WordNet senses. Unlike word similarity judgments, these two datasets provide only a binary distinction between senses, indicating whether two senses are sufficiently similar that they can be considered identical or whether they are semantically distinct. In contrast to these works, our CLSS datatset compares a word with a sense along a graded similarity scale, capturing a wider range of semantic relationships between the word and sense such as synonymy or topical association. Also related are the works of Erk and McCarthy (2009), Erk et al. (2013) and Jurgens and Klapaftis (2013), who measure applicability of a word sense to a usage, which is analogous to measuring the similarity of the sense to a word in context. In contrast to these judgments, our CLSS datatset compares a word with a sense that is not necessarily a meaning of the word, capturing a broad range of semantic relationships between the two.

3 Task description

The SemEval-2014 task on CLSS 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).

Following previous SemEval tasks (Agirre et al. 2012; Jurgens et al. 2012), CLSS-2014 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 2.[3]

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 3 provides examples of pairs for each scale rating for all four comparison types. We use the sense notation of Navigli (2009) and show the nth sense of the *word* with part of speech p as $word_p^n$.

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

Table 2 The five-point Likert scale used to rate the similarity of item pairs in the CLSS task

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 meanings, 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

See Table 3 for examples

For each of the levels, the ability to distinguish between the rating scale's points supports multiple types of application, even when distinguishing lower-similarity pairs. For example, separating items rated as 0 from those as 1 can aid in improving topical coherence by removing linguistic items that are too dissimilar from a topic's current content. Further, at the phrase, word, and sense level, distinguishing between items rated 1 from 2 can potentially identify those items with highly-salient semantic relationships (e.g., meronymy) from those items that are just likely to appear in the same topic, thereby aiding taxonomy enrichment. At the paragraph-to-sentence level, distinguishing between a 2 and 3 can aid in multi-document summarization by identifying sentences that are novel and related to the current summary (rating 2) from those that are similar to the existing content (rating 3) and might be redundant.

4 Task data

The task's pilot dataset was designed to test the capabilities of systems in a variety of settings. Except for the word-to-sense setting, the task data for all comparison types was created using a three-phase procedure. First, items of all sizes were selected from publicly-available corpora. Second, each of the selected items was 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 dataset was constructed by the organizers using a separate but similar process. We generated 1000 pairs for each of the four comparison types which are equally distributed among training and test sets. In this Section we first describe the corpora used for the generation of CLSS datasets (Sect. 4.1) followed by the annotation process

Table 3 Example pairs and their ratings

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 ofa 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)

(Sect. 4.2). The construction procedure for the word-to-sense comparison type is detailed in Sect. 4.3.

4.1 Corpora

CLSS datasets 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, item pairs were created from different corpora that included texts from specific domains, social media, and text with idiomatic or slang language. Table 4 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 text was drawn from the Berlitz travel guides data in the Open American National Corpus (Ide and Suderman 2004) and includes very verbose sentences with many named entities. Wikipedia Science text was drawn from articles tagged with the category *Science* on Wikipedia. Food reviews were drawn from the SNAP Amazon Fine Food Reviews dataset (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 dataset, which is a collection of questions and answers from the Community Question Answering (CQA) site; the dataset 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 datasets 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.

In order to evaluate the ability of the participating systems at generalizing to data from a novel domain, the test dataset in each comparison type included one surprise genre that was not seen in the training data. 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. On average, a paragraph in our dataset consists of 3.8 sentences. Typos and grammatical mistakes in the community-produced content were left unchanged.

Table 4 Percentages of the training and test data per source corpus

Corpus	Genre	Paragraph-to-sentence		Sentence-to-phrase		Phrase-to-word	
		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

4.2 Annotation process

In order to ensure high-quality datasets, a two-phase procedure was used for the generation of all datasets but word-to-sense. Phase 1 deals with the generation of item pairs while ensuring a uniform distribution of items along the similarity scale. In Phase 2 the annotators rate the produced item pairs for their similarity.

Phase 1 The goal of this phase is to produce item pairs with an expected uniform distribution of similarity values along the rating scale. To this end, the larger texts that were drawn from different corpora were shown to annotators who were asked to produce the smaller text of the pair at a specified similarity. For instance, an annotator was given the phrase "drop a dime" and asked to write the paired word that is a "3" rating. The annotator provided "inform" for this phrase and the specified similarity value. 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. The procedure was only used for the generation of items with 1–4 ratings. Unrelated pairs (i.e., with 0 similarity score) were automatically generated by pairing the larger item with an item of appropriate size extracted randomly from the same genre. Annotators were not instructed to intentionally change the surface text to use synonyms or to paraphrase, though they were free to do. While requiring these changes would necessitate that systems use more semantic analysis for comparison instead of string similarity, we intended for performance on the dataset

to be representative of what would be expected in the real world; thus, string similarity-based approaches were not implicitly penalized through construction of the dataset.

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 authors also assisted in Phase 1: Both completed items from the SCIENTIFIC genre and the first author produced 994 pairs, including all those for the METAPHORIC genre, and those that the other annotators left blank.

Phase 2 Once item pairs were produced for different similarity ratings, they were stripped of their associated scores and were given to annotators for their similarity to be rated. An initial pilot annotation study showed that crowdsourcing did not produce high-quality annotations that agreed with the expert-based gold standard. Furthermore, the texts used in our datasets came from a variety of genres, such as scientific domains, which some workers had difficulty in understanding. While we note that crowdsourcing has been used in prior STS tasks for generating similarity scores (Agirre et al. 2012, 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.

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 annotator was added for those datasets. 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.

4.3 Word-to-sense

The word-to-sense dataset was produced in three phases. In Phase 1, we picked words in order to construct the word side of the dataset. Based on the type of their word side, the items in the word-to-sense dataset were put into five categories:

- *Regular* The word and its intended meaning are in WordNet. Words in this category were picked in a way to ensure a uniform distribution of words based on their importance as measured in terms of occurrence frequency. To this end, the lemmas in WordNet were ranked by their occurrence frequency in Wikipedia; the ranking was divided into ten equally-sized groups, with words sampled evenly from groups.
- *OOV* The word does not exist in the WordNet's vocabulary, e.g., the verb "increment." WordNet out-of-vocabulary (OOV) words were drawn from Wikipedia.

- *Slang* Slang words not already in WordNet were selected from slang terminologies such as Wiktionary. This category is separated from OOV in order to highlight any differences in performance due to the usage of the words and the resources available for slang compared to the technical or domain-specific terminology in the OOV category.
- *OOS* The word is in WordNet, but has a novel meaning that is not defined in the WordNet sense inventory, e.g., the noun "Barcelona" referring to the football team. In order 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).
- *Challenge* A set of challenge words where one of the word's senses and a second sense of another word are directly connected by an edge in the WordNet network, but the two senses are not always highly similar, e.g., the first sense of the noun "white_goods"[4] is directly linked to the first sense of the noun "plural"[5] but they do not possess a high similarity. The words in this category were chosen by hand. The part-of-speech distributions for all five types of items were balanced as 50 % noun, 25 % verb, 25 % adjective.

In Phase 2, we first associated each word 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, a neighborhood search procedure was adopted: All synsets connected by at most three edges in the WordNet semantic network were drawn. Given a word and its neighborhood, the corresponding sense for the item pair was selected by matching the sense with an intended similarity value 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, the annotators were provided with the definitions for the word's intended meaning and for the senses for all word-sense pairs and asked to rate each pair according to the rating scale. Definitions were drawn from WordNet or from Wiktionary, if the word was OOV or OOS. Annotators had access to WordNet for the compared sense in order to take into account its hypernyms and siblings.

4.4 Trial data

The generation procedure for the trial dataset was similar but on a smaller scale. For this dataset we generated pairs by sampling text from WikiNews and words from WordNet's vocabulary. The smaller side was then manually produced resulting in a total of 156 pairs for the four comparison types. Four fluent annotators were asked to independently rate all items. Inter-annotator agreement rates varied in 0.734–0.882, in terms of Krippendorff's α (Krippendorff 2004) on the interval scale.

[4] Defined as "large electrical home appliances (refrigerators or washing machines etc.) that are typically finished in white enamel".

[5] Defined as "the form of a word that is used to denote more than one".

4.5 Dataset OOV analysis

A major goal of the CLSS evaluation is to create robust semantic representations of arbitrary-sized texts that are meaningfully comparable. Given the use of WordNet in one comparison level (i.e., word to sense), we anticipated that WordNet might serve as either a common representation across levels or as a semantic resource for comparing items. However, WordNet is limited in the number of word forms it contains and often omits many technical or jargon terms. Therefore, we analyzed the percentages of words in each level that are not present in the vocabulary of WordNet 3.0. These OOV words present a significant challenge for WordNet-based similarity systems which must find alternate ways of representing and comparing such words.

Table 5 shows the percentage of content words[6] in the CLSS datasets that do not exist in WordNet 3.0's vocabulary, for different genres and for different comparison types in the training and test sets. Travel, CQA, and Newswire are the genres with most WordNet OOV percentage in the paragraph-to-sentence and sentence-to-phrase comparison types. These genres are characterized by their high number of named entities, words that are less likely to exist in the WordNet's vocabulary (cf. Sect. 4.1). The OOV percentage is relatively balanced across the two sides in different genres in these two comparison types (i.e., paragraph to sentence and sentence to phrase). Exceptions are Idiomatic and Slang in which the larger side tends to have a higher percentage of its content words not defined in WordNet.

Specifically, in the paragraph to sentence datasets, on average, 8.9 and 7.2 % of words are WordNet OOV in training and test sets, respectively. The mean OOV percentage value in this comparison type ranges from 1.2 for the Metaphoric genre to 11.48 for the Travel genre, both in the test set. Sentence to phrase datasets have 7.4 and 6.8 % of their words not defined in WordNet which is slightly lower in comparison to the paragraph to sentence type.

In the phrase-to-word datasets, about 6.5 % of words are not covered in WordNet in both training and test sets. It is notable that, in this comparison type, about a quarter of content words in the phrase side of the Descriptive genre are not defined in WordNet, denoting the high number of named entities in the image captions of Wikipedia, a resource from which these phrases have been obtained.

Finally, in the word-to-sense training set, 12.2 % of WordNet senses are paired with words that are not defined in the same sense inventory. This figure rises to more than 19 % in the test set. All these WordNet OOV words in the word-to-sense datasets belong to either OOV or slang type.

[6] We consider only words with one of the four parts of speech: noun, verb, adjective, and adverb.

Table 5 Percentage of WordNet OOV words per genre in the CLSS-2014 data

Genre	Training			Test		
	Source	Target	Mean	Source	Target	Mean
Paragraph to sentence						
Overall	8.52	9.28	8.90	7.58	6.84	7.21
Travel	11.28	10.21	10.74	11.31	11.64	11.48
Newswire	11.39	10.43	10.91	12.26	9.56	10.91
CQA	9.66	12.14	10.90	7.84	9.16	8.50
Metaphoric	2.04	2.55	2.29	1.36	1.07	1.21
Review	4.26	5.28	4.77	4.45	3.79	4.12
Scientific				5.13	4.56	4.85
Sentence to phrase						
Overall	7.51	7.31	7.41	7.42	6.15	6.78
Travel	10.35	10.97	10.66	10.33	7.08	8.70
Idiomatic	4.04	1.73	2.88	8.19	4.41	6.30
CQA	12.87	8.44	10.66	11.52	9.85	10.68
Newswire	5.94	6.76	6.35	5.78	6.46	6.12
Scientific				5.31	4.67	4.99
Slang				8.26	3.36	5.81
Phrase to word						
Overall	8.48	4.46	6.47	7.10	5.99	6.54
Newswire	2.52	3.31	2.91	3.05	4.94	3.99
Idiomatic	1.77	0.99	1.38	1.32	1.01	1.16
Descriptive	25.36	4.07	14.71	23.02	14.85	18.94
Lexicographic	2.35	6.90	4.63	0.99	2.20	1.59
Search	4.20	11.54	7.87	3.36	3.57	3.47
Slang				4.14	6.72	5.43
Word to sense						
Overall	12.2	–	–	19.6	–	–

4.6 Dataset discussion

Our multi-phase annotation procedure proved to result in high-quality datasets. Table 6 reports the inter-annotator agreement (IAA) statistics for each comparison type on both the full and unadjudicated portions of the dataset. 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 dataset is expected to be on par with that of the unadjudicated data.[7] The annotation quality for our datasets was further improved by manually adjudicating all significant disagreements.

[7] We note that the α for unadjudicated items is higher than that for all items, since the former set includes only those items for which annotators' scores differed by at most one point on the rating scale and thus the ratings had high agreement.

Table 6 IAA rates for the task data

Data	Training		Test	
	All	Unadjudicated	All	Unadjudicated
Paragraph-to-sentence	0.856	0.916	0.904	0.971
Sentence-to-phrase	0.773	0.913	0.766	0.980
Phrase-to-word	0.735	0.895	0.730	0.988
Word-to-sense	0.681	0.895	0.655	0.952

In contrast, the datasets of current STS tasks aggregated data with moderate inter-annotators correlation (Diab 2013); the inter-rater Pearson correlation varied between 0.530–0.874, 0.377–0.832, and 0.586–0.836 for different datasets in STS-2012 (Agirre et al. 2012), STS-2013 (Agirre et al. 2013), and STS-2014 (Agirre et al. 2014), respectively. 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.

In addition, thanks to the two-phase procedure used for the construction of CLSS datasets, the similarity scores in these datasets are evenly distributed across the rating scale, shown in Fig. 1 as the distribution of the values for all datasets. 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 in future CLSS-based evaluations. Nevertheless, as a pilot task, the manual effort was essential for ensuring a rigorously-constructed dataset.

4.7 Crowdsourcing replication

After our initial tests using crowdsourcing produced unsatisfactory similarity ratings, the annotation process for CLSS-2014 used trained experts, which resulted in a large bottleneck for scaling the annotation process to larger datasets. However, given crowdsourcing's use in creating other STS datasets (Agirre et al. 2012, 2013, 2014), after the task had concluded, we performed a partial replication study to compare the expert-based annotations from the test set with crowdsourced similarity ratings. The replication study's goal was to assess three main points: (1) what is the overall degree of worker rating bias on items, (2) how does rating bias vary by genre, and (3) how does rating bias vary by comparison level. Our aim is to identify specific portions of the annotation task that would be suitable to crowdsource.

To quantify the bias, CrowdFlower workers were recruited to annotate 15 items from each genre for all levels except word-to-sense.[8] To control for differences in the items' similarities, items were balanced across the similarity ratings seen within that genre. Workers were shown an identical set of instructions as the expert annotators, which included examples of pairs at each similarity level.

[8] We observed that working with WordNet senses in the crowdsourced setting proved too complex, e.g., due to the need to easily view a sense's hypernyms; hence, word-to-sense data was not replicated.

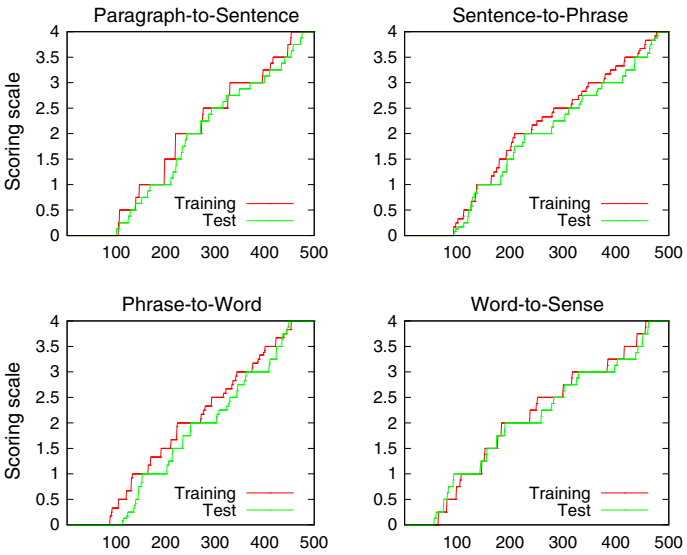

Fig. 1 Similarity rating distribution in the training and test datasets for different comparison types

Annotation tasks included pairs of only a single comparison type. In order to have tasks with similar completion times across levels, workers were shown four pairs per task for paragraph-to-sentence and nine pairs for the two other levels. In all cases, workers were paid $0.10 per task. Three worker judgments were collected per pair. Workers were required to pass a testing phase consisting of correctly rating five pairs within ± 1 of the gold standard rating; testing phase pairs consisted of a separate set of pairs, gathered in an identical way from the data. For all levels, this testing process removed approximately 40–50 % of the initial worker pool from participating further, underscoring the need to control for worker skill level and fluency.

Figure 2 shows the distribution of deviations from the expert ratings as a box-and-whisker plot, revealing that, on average, workers tend to assign higher similarity scores than experts. On the whole, workers more closely matched experts on paragraph-to-sentence comparisons (i.e., had the least scoring variance). Rating variance was highest for the METAPHORIC, IDIOMATIC, and SLANG genres. METAPHORIC texts required reading and understanding the text, which often resulted in workers missing the text's interpretation and rating the fable and its moral as unrelated. IDIOMATIC and SLANG required familiarity with the expressions; given the wide background of crowdsourced workers, we suspect that these deviations were due to lower fluency for colloquial language (Pavlick et al. 2014). Examining genres where workers had a significantly higher similarity rating (e.g., SCIENTIFIC and TRAVEL), we find workers were likely to rate two items with higher similarity if they shared named entities in common, regardless of how those entities functioned in the text, suggesting that workers were operating more on the basis of surface similarity than by reading comprehension.

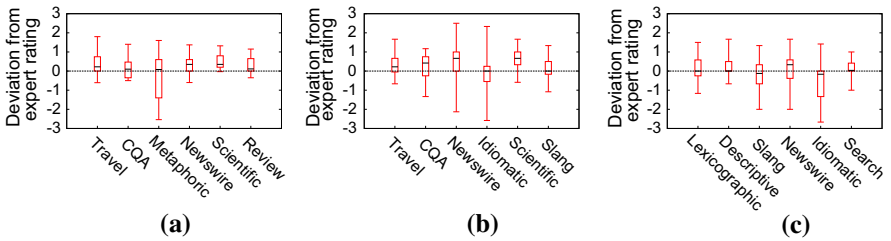

Fig. 2 The distribution of deviations in CrowdFlower ratings from expert ratings. *Lines* denote the median deviations, *boxes* denote the first and third quartiles of deviation, and *whiskers* the minimum and maximum. **a** Paragraph-to-sentence, **b** sentence-to-phrase, **c** phrase-to-word

As a follow-up analysis, worker agreement per genre was measured using Krippendorff's α for interval ratings. Table 7 reveals that the resulting agreement rates are largely below those considered acceptable for high-quality data (e.g., 0.8) and decrease as workers compare shorter pairs of text. However, one notable exception is the agreement for Newswire pairs at the paragraph-to-sentence level, which had an α of 0.874. While this agreement approaches that of the expert annotators, the worker's mean similarity rating per item for this data was 0.43 points higher on average than that of experts, as shown in Fig. 2. Thus, while workers have high agreement, they are agreeing on an answer that differs from the expert judgment. Furthermore, for the remaining comparison levels, the genres with highest agreement in the other two levels also have worker-based ratings that are higher than those of experts, indicating that increased worker agreement does not correspond to increased accuracy.

The results of the replication study suggest that directly crowdsourcing ratings would encounter four main challenges. First, crowdsourced workers did not have consistent ratings, with many genres seeing inflated similarity values. Second, workers encountered difficulties when rating texts such as idioms, which require native fluency to comprehend. Third, workers did not appear to spend sufficient effort to understand the text and instead relied on surface similarity, leading to higher variance in the ratings when text comprehension was required or when a pair's items had text in common but the two items had very different meanings. Fourth, even when workers do have high agreement on the similarity, the ratings did not match experts', indicating that examining IAA alone is insufficient for assessing dataset quality. Together, these findings suggest that crowdsourcing is not readily feasible as an alternative strategy for gathering CLSS similarity rating annotations for any level unless the process can be further adapted to control for worker ability and annotation quality.

5 Evaluation

5.1 Participation

The ultimate goal of the CLSS-2014 pilot task is to benchmark 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

Table 7 IAA rates per genre between CrowdFlower workers

Paragraph-to-sentence		Sentence-to-phrase		Phrase-to-word	
Genre	α	Genre	α	Genre	α
Travel	0.634	Travel	0.497	Descriptive	0.396
CQA	0.593	CQA	0.518	Lexicographic	0.761
Newswire	0.874	Newswire	0.407	Newswire	0.512
Scientific	0.618	Scientific	0.470	Search	0.546
Review	0.695	Idiomatic	0.445	Idiomatic	0.287
Metaphoric	0.386	Slang	0.697	Slang	0.236

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.

5.2 Scoring

Systems were required to provide similarity values for all items within a comparison type. However, systems were allowed to produce optional confidence values for each score, reflecting their certainty in the item's rating. In practice, few systems reported confidence scores, so we omit further discussion. 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 to those made by humans. 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.

5.3 Baselines

String similarity measures have provided competitive baselines for estimating the semantic similarity of two texts (Bär et al. 2012; Šarić et al. 2012). Therefore, the official baseline system was based on the longest common substring (LCS) measure, 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 also added. Unlike LCS, GST accounts for 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.

6 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 been released. These systems were scored and noted in the results but were not included in the official ranking. Table 8 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.

6.1 Systems analysis

Systems adopted a wide variety of approaches for measuring similarity, partly due to the different levels in which they participated. Table 9 summarizes the major resources and tools used by each system. Three main trends emerge. First, many systems benefited by combining the outputs of multiple similarity methods, with the corpus-based distributional similarity being the most common approach. Among the top-5 systems, three, i.e., SemantiKLUE, ECNU, SimCompass, used different classes of similarity measures such as distributional, knowledge-based, and string-based. Knowledge-based measures such as Lin (1998), that view WordNet as a

Table 8 The CLSS task results

Team	System	Para-to-sent	Sent-to-phr	Phr-to-word	Word-to-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

Table 8 continued

Team	System	Para-to-sent	Sent-to-phr	Phr-to-word	Word-to-sense	Official rank	Spearman rank
UMCC DLSI SelSim	Run1		0.760			36	35
UMCC DLSI SelSim	Run2		0.698			37	37
UMCC DLSI Prob	Run1				0.023	38	38

Bolded values indicate the best-performing system for each comparison type

Systems marked with a † were submitted after the deadline but are positioned where they would have ranked. The overall performance difference between the systems highlighted by ‡ is statistically significant at $p < 0.05$

semantic graph and measure similarity based on the structural properties of this graph, have been used in all the three systems. As for the distributional similarity measures, ECNU and SemantiKLUE used the conventional count-based models whereas SimCompass benefited from the more recent predictive model of word embeddings (Mikolov et al. 2013). The two systems of UNAL-NLP are the only ones that benefit from only one class of similarity measures, i.e., string similarity. The systems performed surprisingly well considering the fact that they only utilize a set of simple string-similarity features based on soft cardinality (Jimenez et al. 2010). Interestingly, the UNAL-NLP run1 system that ranked fifth does not also use any machine learning procedure for training, mirroring the potential for unsupervised semantic similarity measured seen in the recent work of Sultan et al. (2014, 2015).

Second, many systems modeled word senses using their textual definitions, rather than representing them using their structural properties in WordNet. In fact, in the word-to-sense comparison type, all the top-5 systems use WordNet sense inventory to transform a word sense to a textual item given by the corresponding synset's definition. This transformation permits the systems to model a word sense in the same manner they do for longer textual items such as phrases and sentences. Further, given the limited text in these definitions, many systems enriched the definition by including additional text from the neighboring senses or from other lexical resources. For example, the MeerkatMafia-pairingWords system used WordNik[9] in order to tackle the problem of uncovered WordNet OOV words. WordNik is a compilation of several dictionaries such as The American Heritage Dictionary and Wiktionary. As a result of this addition, the system attains the best performance in the word-to-sense comparison type whereas the second best performance of the system is in the paragraph-to-sentence type where it ranks no better than 10th. Given that many of the other top-performing systems have superior performance to MeerkatMafia-pairingWords for large texts but used only WordNet for glosses and synonyms, the addition of expanded semantic taxonomies for covering OOV words may provide a significant performance improvement. Alternatively, Jurgens and Pilehvar (2015) show that using Crown, an extension

[9] https://www.wordnik.com/.

Table 9 Resources and tools used by different systems in the CLSS 2014 task

Team	System	Rank					Main approach	Similarity measure				Resources				View a sense as its textual definition	Syntax	Unsupervised
		Official	P2S	S2P	P2W	W2S		String	Distributional	WN-based	Other	WordNet definitions	Wiktionary	Wikipedia	Corpora and other lexical resources			
AI-KU	run1	24	23	18	-	-	S-CODE clustering											
	run2	26	25	26	-	-					Represent items using distribution over lexical substitutes				ukWaC	-		✓
BUAP	run1	9	8	11	17	8	Heterogeneous features	✓	✓	✓								
	run2	10	9	12	18	9					Expand words in P2W with their related terms in Related Tags Service of Flickr	✓			Europarl, Project-Gutenberg, and Open Office Thesaurus, Word Reference	✓		
DIT	run1	34	13	-	-	-	Summarize text; expand text with WN synonyms											
	run2	35	15	-	-	-					TextRank, Text summarization					-	✓	
Duluth	Duluth1	30	33	32	21	19	First-order Lesk	✓										
	Duluth2	23	32	30	7	7	Second-order Vector		✓									
	Duluth3	31	34	34	20	18	Ensemble of Duluth1 and a variant with different stop words	✓				✓	✓	✓	Collaborative International Dictionary of English, Moby Thesaurus II, VERA, the Jargon File, the Free Online Dictionary of Computing, CIA World Factbook	✓		✓
ECNU	run1	2	2	2	2	4	Heterogeneous features	✓	✓	✓	MT-based features	✓		✓	New York Times Annotated Corpus (NYT)	✓	✓	
FBK-TR	run1	21	21	17	-	-	Heterogeneous features	✓	✓	✓	Topic modeling (LDA)	✓		✓	BNC	-		
	run2	22	17	23	-	-												
	run3	20	20	15	-	-												
HULTECH	run1	12	26	22	5	15	InfoSimba: document retrieval by Solr over Anchor ClueWeb12 dataset											
	run2	15	28	24	14	11					Document retrieval	✓			AnchorClueWeb12	✓		
	run3	13	27	20	8	16												
JU-Evora	run1	28	31	31	19	17	Discourse Representation Structure				Represent text as a graph (using discourse units) and then identify common supergraphs for similarity	✓				✓		
Meerkat_Mafia	pairingWords	11	10	14	22	1	Latent Semantic Analysis				Web search-based features							✓
	SuperSaiyan	17	3	1	-	-						✓	✓		2007 Web crawl (Stanford WebBase project), Wordnik, American Heritage Dictionary, DBpedia Lookup service	✓		
	Hulk2	18	4	13	-	-			✓									✓
OPI	run1	32	-	33	12	14	Word co-occurrence statistics in Wikipedia paragraphs		✓		Use Wikipedia co-occurrence statistics to cluster based on similarity levels and compute similarity using cluster centroids	✓		✓		✓		✓
RTM-DCU	run1	16	12	21	15	-	Similarity as a translation performance prediction using Referential Translation Machines	✓	✓		Machine Translation Performance Predictor (MTPP)				Corpus not specified	-	✓	
	run2	19	22	27	16	-												
	run3	14	16	19	13	-												
SemantiKLUE	run1	4	5	4	11	3	Heterogeneous features	✓	✓	✓	Alignment-based similarity measurement	✓			Wackypedia, ukWaC, UMBC WebBase, and UKCOW 2012, Google ngrams, SICK dataset	✓	✓	
SimCompass	run1	1	7	5	1	2	Deep learning word embeddings		✓	✓	Transform texts to a sets of topic centroids; then check for closest topics	✓			part of Google News dataset, Google Word2Vec	✓		
SSMT	run1	33	11	-	-	-	Compute MT metrics, use a regression to convert them to similarity predictions	✓			BLEU, ROGUE-L, ROGUE-S, METEOR				WMT12	-		
TCDSCSS	run1	29	30	29	-	-	Latent Semantic Analysis		✓						NIST TACs Knowledge Base Population	-		
	run2	27	29	28	-	-												
UMCC_DLSI_SemSim	run1	36	-	3	-	-	Heterogeneous features	✓	✓	✓	generate word vectors; compare items using vector alignment				UMBC WebBase Corpus	-	✓	
	run2	37	-	16	-	-												
UNAL-NLP	run1	5	6	6	3	6	Soft cardinality	✓				✓			AnchorClueWeb12	✓		✓
	run2	3	1	7	6	5									SICK dataset			
UNIBA	run1	7	18	9	4	12	WSD and Distributional Semantic Model	✓	✓		Model text fragments by applying PPR on subgraphs of BabelNet	✓		✓	BNC, BabelNet	✓		
	run2	6	14	8	9	10												
	run3	8	19	10	10	13												

of WordNet with Wiktionary content, results in a large CLSS performance benefit to off-the-shelf WordNet-based techniques on the word-to-sense subtask simply due to the presence of OOV terms in the resource.

Third, high performance on the phrase-to-word and word-to-sense comparisons requires moving beyond textual comparison, which is most clearly seen in the purely string-based UNAL-NLP systems, which both perform in the top six systems for the larger levels but whose ranks drop significantly when comparing smaller items. Indeed, systems that included additional lexical resources and those using distributional models at the word level tended to do better on average, though no clear trend emerges in which resources to use.

6.2 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 linguistic 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 (see also Pilehvar and Navigli 2015 for further analysis). 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.

6.3 Per-genre results and discussion

The CLSS-2014 task includes multiple genres within the dataset for each comparison type. Figure 3 shows the correlation of each system for each of these genres, with systems ordered left to right according to their official ranking in Table 8. 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 a 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

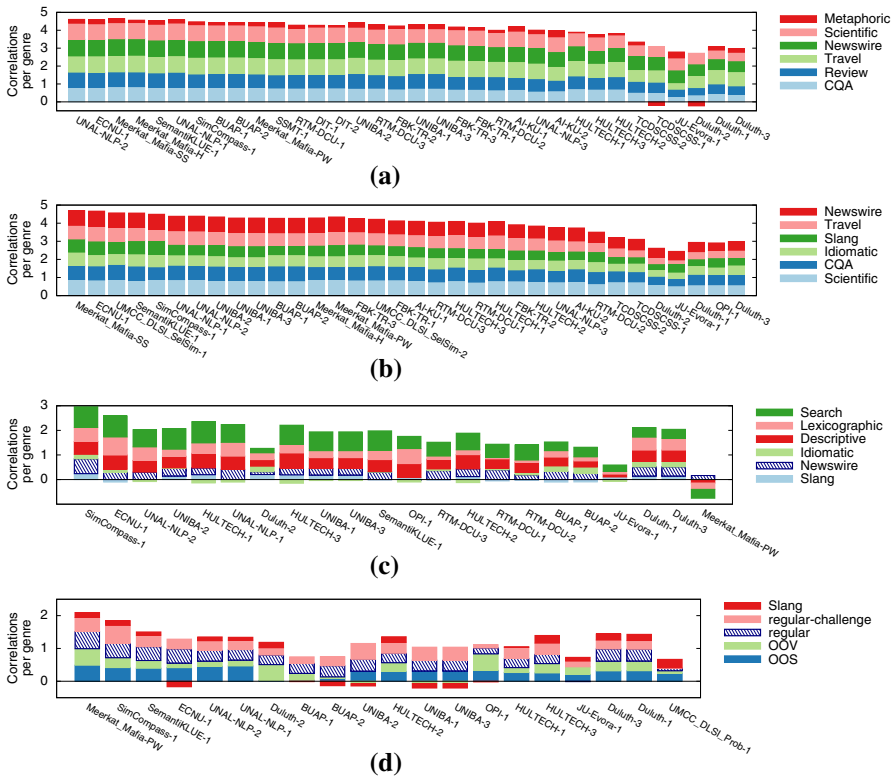

Fig. 3 A stacked histogram for each system, showing its Pearson correlations for genre-specific portions of the gold-standard data, which may also be negative. **a** Paragraph-to-sentence, **b** sentence-to-phrase, **c** phrase-to-word, **d** word-to-sense

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.

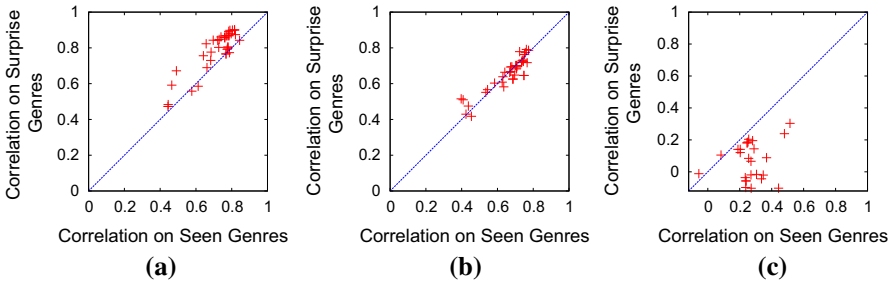

Fig. 4 All systems' Pearson correlations on the test data subsets for (1) genres observed in the training data versus (2) surprise genres not seen in the training data. **a** Paragraph-to-sentence, **b** sentence-to-phrase, **c** phrase-to-word

6.4 Spearman rank analysis

Although the goal of CLSS-2014 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 suggest 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. Nevertheless, Pearson correlation is still an important evaluation metric since it requires systems to judge the similarity of a pair independently from all other pairs, which is essential when a single, arbitrary pair's similarity value is needed as a feature in further applications.

6.5 Analysis of held-out genres

The test datasets for the three text-based levels featured at least one text genre not seen in the training data (cf. Table 4). These held-out genres provide a way to assess the generalizability of a system to novel text styles. The paragraph-to-sentence and sentence-to-phrase test sets contained data from the SCIENTIFIC genre, which was gathered from Wikipedia articles marked with scientific categories and frequently featured jargon or domain-specific terminology. The sentence-to-phrase and phrase-to-word test sets contained texts from the SLANG genre, whose texts were designed to include many colloquial expressions or slang usages of common words.

Performance on the surprise genres differed, with systems having higher Pearson correlation with the gold standard on SCIENTIFIC text than with the observed genres, while lower correlation on SLANG text. Figure 4 shows the relative differences in each system's performance on the unseen genres versus performance on genres observed in the training data. An analysis of the texts for the SCIENTIFIC genre

revealed that its performance was improved due to the presence of jargon and domain-specific terms (Fig. 4a); because these terms are difficult to summarize, similar pairs tend to contain identical jargon terms in both texts. As a result, string similarity measures provide a more accurate estimation of semantic similarity than with other genres. In contrast to SCIENTIFIC texts, SLANG pairs often have little string resemblance to one another. As a result, a meaningful semantic comparison cannot be performed using string similarity and requires comparing alternate representations. A major challenge therefore is to have lexical resources that correctly recognize and represent the slang usage of the word or phrase. We found that many systems did not include special resources for slang text and therefore were unable to recognize and compare these texts meaningfully, resulting in lower scores (Fig. 4c).

7 Future work

The success of this pilot task in CLSS provides several avenues for future work. First, the current evaluation is only based on comparing similarity scores, which omits information on why two items are similar. In a future extension, we plan to develop a complementary subtask based on semantic alignment where systems must identify the portions of a pair's items that are semantically similar and the cause of their degree of similarity (e.g., synonymy, slang paraphrase).

Second, the methods developed for the CLSS task are intended to have practical utility for other NLP tasks, such as summarization. Therefore, in future versions of the task, we plan to include an application-based evaluation where a CLSS system's similarity scores on the pairs in the test set are used in a downstream application (e.g., used to rank the quality of summaries) and the CLSS system is evaluated based on how well the application performed.[10]

Third, CLSS-2014 included a variety of corpora, which revealed notable differences in the capabilities of systems. In particular, systems performed worst on (1) informal texts, such as those from CQA and those containing slang and idioms and (2) on METAPHORIC comparison that require deeper semantic interpretation. In future work, we plan to develop CLSS datasets targeting these two particular aspects. The first dataset will use informal texts such as microtext and email where the medium lends itself to more lexically-compressed writing style. The second will focus on comparison between news stories and their analytical summaries, which may be thematic interpretations of the story content.

Fourth, a key objective will be to develop annotation methodologies that do not require expert intervention. The current annotation process proved time-intensive which prevented the creation of larger datasets. Furthermore, our pre-task investigations and later replication study (Sect. 4.7) showed that crowdsourcing using rating-scale questions did not produce annotations of sufficiently high quality. Therefore, we plan to investigate further adapting the annotation task to the crowdsourced setting, such as requiring workers to explicitly comment on why two

[10] A similar setup was used in SemEval-2013 Task 11 (Navigli and Vannella 2013) that evaluated Word Sense Induction and Disambiguation within an end-user application of search results clustering.

items are similar; furthermore, we plan to pursue annotation using the video-game annotation methods (Vannella et al. 2014; Jurgens and Navigli 2014), which have proven highly successful for other linguistic annotations.

8 Conclusion

This paper introduces a new semantic similarity task, Cross-Level Semantic Similarity, for measuring the semantic similarity of linguistic items of different sizes. Using a multi-phase annotation procedure, we have produced a high-quality dataset of 4000 items drawn from 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. The task was organized as a part of SemEval-2014 and 19 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 many systems perform well only when the text itself is similar, rather than its underlying meaning. Nevertheless, the results of CLSS-2014 are highly encouraging and point to clear future objectives for developing CLSS systems that operate more on semantic representations rather than text. All task data and resources are available at http://alt.qcri.org/semeval2014/task3/.

Acknowledgments The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of NAACL, Boulder, CO* (pp. 19–27).

Agirre, E., Cer, D., Diab, M., & Gonzalez-Agirre, A. (2012). SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the 6th international workshop on semantic evaluation (SemEval-2012), Montréal, Canada* (pp. 385–393).

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., & Guo, W. (2013). *SEM 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity. In *Proceedings of the second joint conference on lexical and computational semantics (*SEM), Atlanta, GA* (pp. 32–43).

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., et al. (2014). SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), Dublin, Ireland* (pp. 81–91).

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Bär, D., Biemann, C., Gurevych, I., & Zesch, T. (2012). UKP: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of SemEval-2012, Montréal, Canada* (pp. 435–440).

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

Diab, M. (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.

Dolan, B., Quirk, C., & Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on computational linguistics, Geneva, Switzerland* (pp. 350–356).

Erk, K., & McCarthy, D. (2009). Graded word sense assignment. In *Proceedings of the 2009 conference on empirical methods in natural language processing (EMNLP), Singapore* (pp. 440–449).

Erk, K., McCarthy, D., & Gaylord, N. (2013). Measuring word meaning in context. *Computational Linguistics*, *39*(3), 511–554.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic database*. Cambridge, MA: MIT Press.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., et al. (2001). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*(1), 116–131.

Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of NAACL, Atlanta, GA* (pp. 758–764).

Hill, F., Reichart, R., & Korhonen, A. (2014). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. arXiv:1408.3456.

Ide, N., & Suderman, K. (2004). The American National Corpus first release. In *Proceedings of the 4th language resources and evaluation conference (LREC), Lisbon, Portugal* (pp. 1681–1684).

Jimenez, S., Gonzalez, F., & Gelbukh, A. (2010). Text comparison using soft cardinality. In *Proceedings of the 17th international conference on string processing and information retrieval* (pp. 297–302). Berlin: Springer.

Jurgens, D., & Klapaftis, I. (2013). SemEval-2013 task 13: Word sense induction for graded and non-graded senses. In *Second joint conference on lexical and computational semantics (*SEM). Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013), Atlanta, GA, USA* (Vol. 2, pp. 290–299).

Jurgens, D., & Navigli, R. (2014). It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *Transactions of the Association for Computational Linguistics (TACL)*, *2*, 449–464.

Jurgens, D., & Pilehvar, M. T. (2015). Reserating the awesometastic: An automatic extension of the WordNet taxonomy for novel terms. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies, Denver, CO* (pp. 1459–1465).

Jurgens, D., Pilehvar, M. T., & Navigli, R. (2014). SemEval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th international workshop on semantic evaluation, Dublin, Ireland* (pp. 17–26).

Jurgens, D., Mohammad, S., Turney, P., & Holyoak, K. (2012). SemEval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the 6th international workshop on semantic evaluation (SemEval-2012), Montréal, Canada* (pp. 356–364).

Kilgarriff, A. (2001). English lexical sample task description. In *The proceedings of the second international workshop on evaluating word sense disambiguation systems (SENSEVAL-2), Toulouse, France* (pp. 17–20).

Kim, S. N., Medelyan, O., Kan, M. Y., & Baldwin, T. (2010). SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th international workshop on semantic evaluation (SemEval-2010), Los Angeles, CA* (pp. 21–26).

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit X, Phuket, Thailand* (pp. 79–86).

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, *18*(8), 1138–1150.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the fifteenth international conference on machine learning, San Francisco, CA* (pp. 296–304).

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., & Zamparelli, R. (2014). SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through

semantic relatedness and textual entailment. In *Proceedings of SemEval-2014, Dublin, Ireland* (pp. 1–8).

McAuley, J.J., Leskovec, J. (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, Rio de Janeiro, Brazil* (pp. 897–908).

McCarthy, D., & Navigli, R. (2009). The English lexical substitution task. *Language Resources and Evaluation*, *43*(2), 139–159.

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the conference of the North American chapter of the association for computational linguistics (NAACL), Atlanta, GA* (pp. 746–751).

Navigli, R. (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), Sydney, Australia* (pp. 105–112).

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, *41*(2), 1–69.

Navigli, R., & Vannella, D. (2013). SemEval-2013 task 11: Evaluating word sense induction and disambiguation within an end-user application. In *Proceedings of the 7th international workshop on semantic evaluation (SemEval 2013), in conjunction with the second joint conference on lexical and computational semantics (*SEM 2013), Atlanta, USA* (pp. 193–201).

Pavlick, E., Post, M., Irvine, A., Kachaev, D., & Callison-Burch, C. (2014). The language demographics of amazon mechanical turk. *Transactions of the Association for Computational Linguistics*, *2*, 79–92.

Pilehvar, M. T., & Navigli, R. (2014a). A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation. *Computational Linguistics*, *40*(4), 837–881.

Pilehvar, M. T., & Navigli, R. (2014b). A robust approach to aligning heterogeneous lexical resources. In *Proceedings of the 52nd annual meeting of the association for computational linguistics, Baltimore, USA* (pp. 468–478).

Pilehvar, M. T., & Navigli, R. (2015). From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, *228*, 95–128.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633.

Šarić, F., Glavaš, G., Karan, M., Šnajder, J., & Dalbelo Bašić, B. (2012). Takelab: Systems for measuring semantic text similarity. In *Proceedings of SemEval-2012, Montréal, Canada* (pp. 441–448).

Snow, R., Prakash, S., Jurafsky, D., & Ng, A. Y. (2007). Learning to merge word senses. In *The 2012 conference on empirical methods on natural language processing and computational natural language learning, Prague, Czech Republic* (pp. 1005–1014).

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

Specia, L., Jauhar, S. K., & Mihalcea, R. (2012). SemEval-2012 task 1: English lexical simplification. In *Proceedings of the sixth international workshop on semantic evaluation (SemEval-2012), Montréal, Canada* (pp. 347–355).

Sultan, M. A., Bethard, S., & Sumner, T. (2014). Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*, *2*, 219–230.

Sultan, M. A., Bethard, S., & Sumner, T. (2015). DLS@CU: Sentence similarity from word alignment and semantic vector composition. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), Denver, CO* (pp. 148–153).

Vannella, D., Jurgens, D., Scarfini, D., Toscani, D., & Navigli, R. (2014). Validating and extending semantic knowledge bases using video games with a purpose. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (ACL 2014), Baltimore, MD* (pp. 1294–1304).

Wise, M. J. (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, Philadelphia, PA, USA* (pp. 130–134).