CoSimLex : A Resource for Evaluating Graded Word Similarity in Context

Armendariz, Carlos S.

EUROPEAN LANGUAGE RESOURCES ASSOC-ELRA
2020

Armendariz , C S , Purver , M , Ulcar , M , Pollak , S , Ljubesic , N , Robnik-Sikonja , M , Granroth-Wilding , M & Vaik , K 2020 , CoSimLex : A Resource for Evaluating Graded Word Similarity in Context . in N Calzolari , F Bechet , P Blache , K Choukri , C Cieri , T Declerck , S Goggi , H Isahara , B Maegaard , J Mariani , H Mazo , A Moreno , J Odijk & S Piperidis (eds) , PROCEEDINGS OF THE 12TH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2020) . EUROPEAN LANGUAGE RESOURCES ASSOC-ELRA , pp. 5878-5886 , 12th International Conference on Language Resources and Evaluation (LREC) , Marseille , France , 11/05/2020 . <
https://aclanthology.org/2020.lrec-1.720 >

http://hdl.handle.net/10138/344267

CoSimLex: A Resource for Evaluating Graded Word Similarity in Context

Carlos S. Armendariz, Matthew Purver[y],
Matej Ulčar[z], Senja Pollak[y], Nikola Ljubešić[y], Marko Robnik-Šikonja[z],
Mark Granroth-Wilding[}], Kristiina Vaik[x]
Cognitive Science Research Group, Queen Mary University of London, London, UK
{c.santosarmendariz, m.purver}@qmul.ac.uk
[y]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
{senja.pollak, nikola.ljubesic}@ijs.si
[z]University of Ljubljana, Faculty of Computer and Information Science, Slovenia
{matej.ulcar, marko.robnik}@fri.uni-lj.si
[}] Department of Computer Science, University of Helsinki, Finland
mark.granroth-wilding@helsinki.fi
[x]Department of Data Analysis, Texta, Estonia
kristiina.vaik@ut.ee

Abstract

State of the art natural language processing tools are built on context-dependent word embeddings, but no direct method for evaluating these representations currently exists. Standard tasks and datasets for intrinsic evaluation of embeddings are based on judgements of similarity, but ignore context; standard tasks for word sense disambiguation take account of context but do not provide continuous measures of meaning similarity. This paper describes an effort to build a new dataset, CoSimLex, intended to fill this gap. Building on the standard pairwise similarity task of SimLex-999, it provides context-dependent similarity measures; covers not only discrete differences in word sense but more subtle, graded changes in meaning; and covers not only a well-resourced language (English) but a number of less-resourced languages. We define the task and evaluation metrics, outline the dataset collection methodology, and describe the status of the dataset so far.

Keywords: corpus, annotation, semantics, similarity, context, salience, context-dependence

1. Introduction

Recent work in language modelling and word embeddings has led to a sharp increase in use of context-dependent models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). These models, by providing representations of words which depend on the surrounding context, allow us to take account of the effects not only of discrete differences in word sense but of the more graded effects of context. However, evaluation of these models has generally been in terms of either their performance as language models, or their effect on downstream tasks such as sentiment classification (Peters et al., 2018): there are few resources available which allow evaluation in terms of the properties of the embeddings themselves, or in terms of their ability to model human perceptions of meaning. There are established methods to evaluate word embedding models intrinsically via their ability to reflect human similarity judgements (see e.g. WordSim-353 (Finkelstein et al., 2002) and SimLex-999 (Hill et al., 2015)) or model analogies (Mikolov et al., 2013); however, these have generally ignored context and treated words in isolation. The few that do provide context (e.g. SCWS (Huang et al., 2012) and WiC (Pilehvar and Camacho-Collados, 2019)) focus on word sense and discrete effects, thus missing some of the effects that context has on words in general, and some of the benefits of context-dependent models. To evaluate current models, we need a way to evaluate their ability to reflect similarity judgements *in context*: how well do they model the effects that context has on word meaning?

In this paper we present our ongoing efforts to define and build a new dataset that tries to fill that gap: CoSimLex (Armendariz et al., 2020). CoSimLex builds on the familiar pairwise, graded similarity task of SimLex-999, but extends it to pairs of words as they occur in context, and specifically provides two different shared contexts for each pair of words. This will provide a dataset suitable for intrinsic evaluation of state-of-the-art contextual word embedding models, by testing their ability to reflect human judgements of word meaning similarity in context, and crucially, the way in which this varies as context is changed. It goes beyond other existing context-based datasets by taking the *gradedness* of human judgements into account, thus applying not only to polysemous words, or words with distinct senses, but to the phenomenon of context-dependency of word meaning in general. The dataset is also multilingual, and includes three less-resourced European languages: Croatian, Finnish and Slovene. It is to be used as the gold standard for evaluation of a task at SemEval2020: Task 3, Graded Word Similarity in Context[1].

2. Background

From the outset, our main motivation for the development of this dataset came from an interest in the cognitive and psychological mechanisms by which context affects our perception of the meaning of words. There have been many different ways in the literature to look at this phenomenon,

[1]https://competitions.codalab.org/competitions/20905

which lie in the intersection of several different fields of research, and a detailed discussion of the different approaches to this problem is out of the scope of this paper; here, we present two of the most prominent ideas that helped define what we were trying to capture, and made an impact in the design of the dataset and its annotation process. We then look at previous datasets that deal with similarity in context.

2.1. Contextual Modulation

Within the field of lexical semantics, Cruse (1986) proposed an interesting compromise between those linguists that saw words as associated with a number of discrete senses and those that thought that the perceived discreteness of lexical senses is just an illusion. He distinguishes two different manners in which sentential context modifies the meaning of a word. First, the context can select for different discrete senses; if that is the case, the word is described as *ambiguous*, and the process is referred as *contextual selection of senses*. This effect is well known, and is the basis of many word-sense disambiguation tasks.

1. We finally reached the bank.
2. At this point, the bank was covered with brambles.

In example (1), the word *bank* can have the *financial* or *river bank* sense; and here, the context doesn't really help us select the correct sense. This creates some tension on the part of the reader: we need to select a sense in order for the sentence to properly work, and without this we may feel that the sentence has not been fully understood. This is an example of *ambiguity*. In example (2), in contrast, the context makes one of the senses *more normal* than the other. Cruse (1986) sees the evaluation of *contextual normality* as the main mechanism for sense selection.

The second way in which context can modify the meaning of a word works within the scope of a single sense, modifying it in an unlimited number of ways by *highlighting* certain semantic traits and *backgrounding* others. This process is called *contextual modulation of meaning* and the word is said to be *general* with respect to the traits that are being modulated. This effect is by nature not discrete but continuous and fluid, and since every word is *general* to some extent: it can be argued that a word has a different meaning in every context in which it appears.

3. Sue is visiting her pregnant cousin.
4. Peter doesn't like his cousin.
5. Arthur poured the butter into a dish.

In example (3), the context tells us that the cousin is female. The meaning of *cousin* is being *modulated* by the context to promote the "female" trait. *Cousin* is a *general* word that includes male and female, but also tall, short, happy and sad cousins. However, as we can see in example (4), the absence of information about these traits doesn't produce the type of tension we saw in (1) above; there is a distinction between meaning modulation and sense selection.

The last example (5) is another case of *contextual modulation* in which *poured* highlights the "liquid" trait for *butter*. It is interesting to notice that in this case not only "liquid" is highlighted, related traits like "warm" can be highlighted as a consequence.

It seems clear that the contextual selection of senses would modify human judgements of similarity. For example, the word *bank*, when used in a context which selects its financial institution sense, should be scored as more similar to other kinds of financial institution (e.g. *building society*) than when in a context which selects the geographic sense of the word. However, we should also expect that a word like *butter*, when contextually modulated to highlight its "liquid", "hot" and "frying" traits, should score more similar to *vegetable oil* than when contextually modulated to highlight its "animal sourced", "dairy", and "creamy" traits. This kind of hypothesis would be testable given a new context-dependent similarity dataset.

Both *sense selection* and *meaning modulation* happen very commonly together, with the same context forcing a sense and then modulating its expression. Many different explanations have been proposed for the emergence of these discrete senses, and some may have their origins in very commonly modulated meaning but, according to Cruse, once a discrete sense is established it becomes something different and follows different rules:

6. John prefers bitches to dogs.
7. John prefers bitches to canines.
8. Mary likes mares better than horses.

Here example (6) works because one of the discrete senses associated to the word *dog* refers only to male dogs. This cannot be explained by *contextual modulation*: if that was the case, example (7), which replaces *dog* with *canine*, should also work, as *canine* could be modulated in the same way that *dog* was; and similarly example (8). However, both seem unnatural at best. The fact that neither *canine* nor *horse* can be modulated in this same way indicates that meaning modulation and sense selection are two, strongly interconnected, but distinctive mechanisms of contextual variability.

A final interesting point about Cruse's view is that he doesn't find the contrast between polysemy and homonymy particularly helpful, and dislikes the use of these terms because they promote the idea that the primary semantic unit is some common lexeme and each of the different senses are just variants of it. He instead believes the primary semantic unit should be the *lexical units*, a union of a single sense and a lexical form, and finds it more useful to look at the contrast between discrete and continuous semantic variability.

2.2. Salience Manipulation

Until now we have looked at contextual variability as an exclusively linguistic phenomenon, a point of view rooted in lexical semantics. We looked at how the context of the sentence affects the meaning of the word. In contrast, cognitive linguistics, and the more specific cognitive semantics, look at language and meaning as a more general expression of human cognition (Evans and Green, 2018).

This approach champions concepts, more specifically *conceptual structures*, as the true recipient of meaning, replacing words or lexical units. These linguistic units no longer refer to objects in an external world but to concepts in the mind of the speaker. Words get their meaning only by association with *conceptual structures* in our minds. The process by which we construct meaning is called conceptualisation, an embodied phenomenon based in social interaction and sensory experience.

Cognitive linguists gravitate to themes that focus on the flexibility and the ability of the interaction between language and conceptual structures to model continuous phenomena, like prototyping effects, categories, metaphor theory and new ways to look at polysemy. Within the cognitive tradition, the idea of *conceptual spaces*, characterised by conceptual dimensions, has been especially influential (Gärdenfors, 2000; Gärdenfors, 2014). These dimensions can range from concrete ones like weight, temperature and brightness, to very abstract ones like awkwardness or goodness. Once a domain, or selection of dimensions is established, a concept is defined as a region (usually a convex one) of the conceptual space. An example would be to define the colour *brown* as a region of a space made of the dimensions *Red*, *Green* and *Blue*. This geometric approach lends itself perfectly to model phenomena like prototyping (central point of the region), similarity (distance), metaphor (projection between different dimensions) and, more importantly for our concerns here, fluid changes in meaning due to the effects of context.

Warglien and Gärdenfors (2015) use conceptual spaces to look at *meaning negotiation* in conversation. They investigate the mechanisms, consciously or unconsciously, employed by the people involved in conversation to negotiate meaning of vague predicates, in order to satisfy the coordination needed for communication. These tools help them to decide areas in which they don't agree as well. All these processes work by manipulating the conceptual dimensions in which meaning is represented. We will refer to them as *salience manipulation* because their main role is to dynamically rise or lower the perceived importance of certain conceptual dimensions.

The main mechanism by which speakers can modify salience of conceptual dimensions are the automatic *priming* effects described by, for example, Pickering and Garrod (2004): mentioning specific words early in the conversation can make the dimensions associated with such words more relevant. Speakers can also explicitly try to remove dimensions from the domain in order to promote agreement or bring in new dimensions by using *metaphoric projections*. Because metaphors can be understood as mappings that transfer structure from one domain to another, they can introduce new dimensions and meaning to the conversation.

> The lion Ulysses emphasizes Ulysses' courage but hides his condition of a castaway in Ogiya. Thus metaphors act by orienting communication and selecting dimensions that may be more or less favorable to the speaker. By suggesting that a storm hit the financial markets, a bank manager can move the conversation away from dimensions pertaining to his own responsibilities and instead focus on dimensions over which he has no control. (Warglien and Gärdenfors, 2015)

From this perspective, then, the change in meaning is no longer a change in the meaning of a specific word, but a change in the mind of the hearer (or reader), a change in their mental state triggered by their interaction with the context. We saw an example of the meaning of the word "butter" being *contextually modulated* before, lets see some examples of *salience manipulation* having an effect on the same word:

9. My muffins were a failure, I should have used butter or margarine instead of olive oil.

10. Vegan chefs replace animal fats, like butter, with plant based ones like olive oil or margarine.

11. Vegan influencers believe the consumption of animal products is cruel and unnecessary.

In example (9), in the context of a baking recipe, important dimensions are related to the physical properties of butter, margarine and olive oil. When focusing on these type of dimensions butter and margarine seem more similar because they are both solid while olive oil is liquid. In contrast, in the following example (10) we bring up ideas about veganism and the dimension of animal versus based plant products becomes very salient. This could bring margarine and olive oil closer together and distance both of them from butter, which is an animal product.

There are important differences between this *salience manipulation* effect and the similarly "graded" *contextual modulation* effect. In the previous example (5) *poured* modulated the meaning of the word butter by promoting its "liquid" trait. This effect is limited to the word butter. On the contrary, if the context triggers changes in the salience of conceptual dimensions, any word the annotator evaluates after the change takes place will be affected by it. Once the idea of animal vs plant based is introduced, the change takes place in the mind of the annotator and the perception of the meaning of not only butter, but margarine and olive oil is impacted as well. Our hypothesis is that, by using *salience manipulation*, a context like example (11) can have a impact in the scoring of the similarity of butter, margarine and olive oil without these words even being present in the context. Something that would be impossible if we were looking only at the *contextual modulation* and *sense selection* effects.

The expectation that priming is the main mechanism for modifying salience has its own implications: Branigan et al. (2000) found that priming effects are much stronger in the context of as natural dialog as possible, when speakers had no time constraints and could respond at their own pace. These results were taken into account when designing our dataset and annotation methodology: it is crucial for us to create an annotation process in which the annotator interacts with the context, and does so in as natural a way as possible, before they rate the similarity. Because priming is an automatic process, them knowing that they should be annotating similarity in context becomes a lot less important.

Word1: bank   Word2: money
Context1 Located downtown along the east *bank* of the Des Moines River ...
Context2 This is the basis of all *money* laundering, a track record of depositing clean money before slipping through dirty money ...

Figure 1: Example from the SCWS dataset, the focus is in the different senses of the word *bank* and there is one independent context per word.

2.3. Existing Datasets

There are a few examples of datasets which take context into account. However, so far these have been motivated by discrete sense disambiguation, and therefore take a view of word meaning as discrete (taking one of a finite set of senses) rather than continuous; they are therefore not suited for the more graded effects we are interested to look into.

The Stanford Contextual Word Similarity (SCWS) dataset (Huang et al., 2012) does contain graded similarity judgements of pairs of words in the context of organically occurring sentences (from Wikipedia). However it was designed to evaluate a discrete multi-prototype model, so the focus was on the contexts selecting for one of the word senses. This resulted in them presenting each of the two words of the pair in their own distinct context. From our point of view this approach has some drawbacks: First, even in the cases where they annotated the same pair twice, we find ourselves with four different contexts, each affecting the meaning of each of the instances of the words independently, and it is not possible to produce a systematic comparison of contextual effects on pairwise similarity. Second, beyond the independent lexical semantics of each word being affected by their independent *local context*, the annotator is being presented with two completely independently occurring contexts at the same time. Even if the two contexts did organically occur on their own, this combination of the two did not, and we have seen before how crucial we think keeping the interaction with the context as natural as possible is. There is no easy way to know how this newly assembled *global context* affects the cognitive state of the annotators and their perception of similarity. The same goes for the contextually-aware models trying to predict their results. Joining the contexts before feeding them to the model could create conflicting, difficult to predict effects, but feeding each context independently is fundamentally different to what humans annotators were presented with.

In addition to these limitations of the independent contexts approach, the scores found in SCWS show a worryingly low inter-rater agreement (IRA), measured as the Spearman correlation between different annotators. As pointed out by (Pilehvar and Camacho-Collados, 2019), the mean IRA between each annotator and the average of the rest, which is considered a human-level upper bound for model's performance, is 0.52; while the performance of a simple context-independent model like word2vec (Mikolov et al., 2013) is 0.65. Examining the scores more in detail, we find that many scores show a very large standard deviation, with annotators rating the same pair very differently. One possible reason for this may lie in the annotation design: the task itself does not directly enforce engagement with the context, and the words were presented to annotators highlighted in boldface, making it easy to pick them out from the context without reading it; thus potentially leading to a lack of engagement of the annotators with the context.

A lot of these limitations were addressed by the more recent Words-in-Context (WiC) dataset (Pilehvar and Camacho-Collados, 2019). With a more direct and straightforward take on word sense disambiguation, each entry of the dataset is made of two lexicographer examples of the same word. The entry is completed with a positive value (T) if the word sense in the two examples/context is the same, or with a negative value (F) if the contexts point to different word senses. One advantage of this design is that it forces engagement with the context; another is that it creates a task in which context-independent models like word2vec "would perform no better than a random baseline". Human annotators are shown to produce healthy inter-rater agreement scores for this dataset. However the dataset is again focused in looking at discrete word senses and cannot therefore capture continuous effects of context in the judgements of similarity between different words.

These datasets are also available only in English, and do not allow models to be evaluated across different languages.

3. Dataset and Task Design

CoSimLex will be based on pairs of words from SimLex-999 (Hill et al., 2015); the reliability and common use of this dataset makes it a good starting point and allows comparison of judgements and model outputs to the context-independent case. For Croatian and Finnish we use existing translations of Simlex-999 (Mrkšić et al., 2017; Venekoski and Vankka, 2017; Kittask, 2019). In the case of Slovene, we have produced our own new translation (Pollak et al., 2020), following the methodology used by Mrkšić et al. (2017) for Croatian.

The English dataset consists of 333 pairs; the Croatian, Finnish and Slovene datasets of 111 pairs each. Each pair is rated within two different contexts, giving a total of 1554 scores of contextual similarity. This poses a difficult task: to find suitable, organically occurring contexts for each pair; this task is more pronounced for languages with less resources, and as a result the selection of pairs is different for each language.

Each line of CoSimLex will be made of a pair of words selected from Simlex-999; two different contexts extracted from Wikipedia in which these two words appear; two scores of similarity, each one related to one of the contexts; and two scores of standard deviation. Please see Figure 2 for an example from our English pilot.

Word1: population Word2: people	SimLex: 7.68 0.80
Context1	Context1: 6.49 1.40
Disease also kills off a lot of the gazelle population. There are many people and domesticated animals that come onto their land. If they pick up a disease from one of these domesticated species they may not be able to fight it off and die. Also, a big reason for the decline of this gazelle population is habitat destruction.	
Context2	Context2: 7.73 1.77
But the discontent of the underprivileged, landless and the unemployed sections remained even after the reforms. The crumbling industries give rise to extreme unemployment, in addition to the rapidly growing population. These people mostly belong to the SC/ST or the OBC. In most cases, they join the extremist organizations, mentioned earlier, as an alternative to earn their livelihoods.	

Figure 2: Example from the English pilot, showing a word pair with two contexts, each with mean and standard deviation of human similarity judgements. The original SimLex values for the same word pair without context are shown for comparison.

Evaluation Tasks and Metrics The first practical use of CoSimLex will be as a gold standard for the public SemEval 2020 task 3: *Graded Word Similarity in Context*. The goal of this task is to evaluate how well modern context-dependent embeddings can predict the effect of context in human perception of similarity. In order to do so we define two subtasks and two metrics:

Subtask 1 - Predicting Changes: In subtask 1, participants must predict the *change* in similarity ratings between the two contexts. In order to evaluate it we calculate the difference between the scores produced by the model when the pair is rated within each one of the two contexts. We do the same with the average of the scores produced by the human annotators. Finally we calculate the uncentered Pearson correlation. A key property of this method is that any context-independent model will predict no change and get strongly penalised in this task.

Subtask 2 - Predicting Ratings: In subtask 2, participants must predict the absolute similarity rating for each pair in each context. This will be evaluated using Spearman correlation with gold-standard judgements, following the standard evaluation methodology for similarity datasets (Hill et al., 2015; Huang et al., 2012). Good context-independent models could theoretically give competitive results in this task, however we still expect context-dependent models to have a considerable advantage.

4. Annotation Methodology

As starting point for our annotation methodology, we adapted the annotation instructions used for SimLex-999. This way we benefit from its tested method of explaining how to focus on *similarity* rather than *relatedness* or association (Hill et al., 2015). As explained in their original paper, *cup* and *mug* are very similar, while *coffee* and *cup* are strongly related but not similar at all. For English we adopted a modified version of their crowd-sourcing process: we use *Amazon Mechanical Turk*[1] with the same scoring scale (0 to 6), the same post-processing and cleaning of the data (a necessary step when working with this kind of crowd-sourcing platform), and achieve similarly good inter-annotator agreement. For the less-resourced languages, crowdsourcing is not a viable option due to lack of available speakers, and we recruit annotators directly. This means fewer annotators (for Croatian, Finnish and Slovene, 12 annotators vs 27 in English), however the average quality of annotation is a lot higher and the data requires less post-processing - see Section 5. for details.

4.1. Finding Suitable Contexts

For each word pair we need to find two suitable contexts. These contexts are extracted from each language's Wikipedia. They are made of three consecutive sentences and they need to contain the pair of words, appearing only once each. English is by far the easiest language to work with, not only because of the amount and quality of the text contained in the English version of Wikipedia but because the other four languages are highly inflected (Croatian, Finnish and Slovene). To overcome this, we work with data from (Ginter et al., 2017)[2] which contains tokenised and lemmatised versions of Wikipedia for 45 languages.

We first find all the possible candidate contexts for each word pair, and then select those candidates that are most likely to produce different ratings of similarity. The differences are expected to be small, especially in words that don't present several senses and are not highly polysemous, so we need a process that has the most chances of finding contexts that make a difference. We use a dual process in which we use ELMo and BERT to rate the similarity between the target pair within each of the candidate contexts. Then we select the 2 contexts in which ELMo scored the pair as the most similar, and the 2 contexts in which it scored them as most different. We do the same using BERT scores. This gives us 4 contexts in which our target words are scored as very similar by the models and 4 contexts in which they are scored as very different.

The final selection of two contexts is made by expert human annotators, one per language. We construct online surveys with these 8 contexts and ask them to select the two in which they think the word pair is the most and the least similar, trying to maximise the potential contrast in similarity. In addition, we ask them how much potential for a difference they see in the contexts selected. This gives us not only the contexts we need, but a predicted performance and direction of change for use in later analysis.

In the case of less resourced languages, the smaller size and lower quality of the Wikipedia text resources require some extra steps to ensure the quality of the final annota-