SLEDDED: A Proposed Dataset of Event Descriptions for Evaluating Phrase Representations

Laura Rimell
Computer Laboratory
University of Cambridge
laura.rimell@cl.cam.ac.uk

Eva Maria Vecchi
Computer Laboratory
University of Cambridge
eva.vecchi@cl.cam.ac.uk

Abstract

Measuring the semantic relatedness of phrase pairs is important for evaluating compositional distributional semantic representations. Many existing phrase relatedness datasets are limited to either lexical or syntactic alternations between phrase pairs, which limits the power of the evaluation. We propose SLEDDED (Syntactically and LExically Divergent Dataset of Event Descriptions), a dataset of event descriptions in which related phrase pairs are designed to exhibit minimal lexical and syntactic overlap; for example, *a decisive victory — won the match clearly*. We also propose a subset of the data aimed at distinguishing event descriptions from related but dissimilar phrases; for example, *vowing to fight to the death — a new training regime for soldiers*, which serves as a proxy for the tasks of narrative generation, event sequencing, and summarization. We describe a method for extracting candidate pairs from a corpus based on occurrences of event nouns (e.g. *war*) and a two-step annotation process consisting of expert annotation followed by crowdsourcing. We present examples from a pilot of the expert annotation step.

1 Introduction

Measuring the semantic relatedness of phrase pairs is an important means of evaluation for vector space representations, particularly in Compositional Distributional Semantics (CDS). However, existing phrase relatedness datasets are not often designed to test lexical and syntactic divergence simultaneously. On the one hand are datasets which hold syntactic structure fixed while varying lexical items, e.g. the adjective-noun dataset of Mitchell and Lapata (2010) (1) and the subject-verb-object dataset of Kartsaklis and Sadrzadeh (2014) (2).

(1) a. *new information*
b. *further evidence*

(2) a. *programme offer support*
b. *service provide help*

Such datasets are useful for examination of targeted syntactic structures, especially in type-based CDS models, but fail to challenge CDS models to compose longer phrases with realistic sentence structure.

On the other hand, the datasets with the most complex and varied syntactic structures tend to exhibit a great deal of lexical overlap across the highly-related pairs, e.g. MSRPar (Dolan et al., 2004) (3) and SICK (Marelli et al., 2014b) (4).

(3) a. *The unions also staged a five-day strike in March that forced all but one of Yale's dining halls to close.*
b. *The unions also staged a five-day strike in March; strikes have preceded eight of the last 10 contracts.*

(4) a. *A hiker is on top of the mountain and is doing a joyful dance.*
b. *A hiker is on top of the mountain and is dancing.*

This phenomenon is not intentional, but a function of the data collection methodology. However, the high degree of lexical overlap makes it difficult to evaluate CDS models, since lexical overlap baselines are challenging to beat (Rus et al., 2014); and non-compositional or semi-compositional methods can perform better than fully compositional ones (Marelli et al., 2014a).

While sentence pairs with high lexical overlap may be common in some tasks – extractive sum-

Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 140–144,
Berlin, Germany, August 12, 2016. ©2016 Association for Computational Linguistics

marization of multiple similar news stories, for example – we believe that datasets with this characteristic are not able to make clear distinctions between CDS models. We therefore propose a new dataset exhibiting both lexical and syntactic variation across related phrases.

2 Proposal

We propose SLEDDED (Syntactically and LExically Divergent Dataset of Event Descriptions), a phrase relatedness dataset in which semantically related phrase pairs are carefully curated to exhibit both syntactic and lexical divergence. Specifically, we propose to base the related pairs on *event descriptions*, where one description is centered around a *non-deverbal event noun* and its counterpart centered around a *verb*. Example noun-verb pairs are shown in Figure 1.

victory – win
ceremony – celebrate
meal – eat
war – fight

Figure 1: Example pairs of non-deverbal event nouns and counterpart verbs (idealized, not from corpus data).

Non-deverbal event nouns describe events, but in contrast to deverbal nouns such as *celebration* or *fighting*, are not morphologically derived from verbs. The use of non-deverbal event nouns ensures that related nouns and verbs cannot be trivially equated by stemming. In the proposed dataset, we aim for minimal shared lemmas in every phrase pair. Example phrase pairs are shown in Figure 2.

a decisive victory – won the match clearly
graduation ceremony – celebrated her degree
a delicious meal – ate pasta bolognese
war between neighbors – fought over borders

Figure 2: Example pairs of short phrases (idealized, not from corpus data).

Although related phrases similar to those described here can be found within many large paraphrase datasets, they are not readily separable from other kinds of related pairs. We believe that more focused datasets like SLEDDED can provide a good complement to larger, less controlled paraphrase datasets.

SLEDDED is aimed primarily at providing a new challenge for CDS. We expect vector addition to be a challenging baseline, as it has been for many other tasks, since simple addition captures word relatedness without regard to syntax. Composition with Recursive Neural Networks (RNNs) may also do well. We consider the dataset to be a particular challenge for type-based (e.g. tensors) and syntax-based (e.g. tree kernels) composition methods. We also propose a subset of confounders that require a distinction between relatedness and similarity for events, that can serve as a proxy for tasks such as narrative generation or event sequencing, and may be challenging for all models; see Section 3.4.

3 Methods

In this section we describe our proposed method for building SLEDDED, and present examples from a pilot involving corpus data extraction and expert annotation.

We choose to extract target phrases from a corpus rather than elicit phrases by crowdsourcing, since we expect the notion of event nouns to be confusing for non-experts, and also expect a wider range of realistic examples from corpus data. We considered several existing methods for automatic extraction of paraphrases that are lexically or syntactically divergent; however, none are exactly suited for our proposed dataset. Bunescu and Mooney (2007) use named entity pairs as anchors for diverse expressions of semantic relations, e.g. *Pfizer buys Rinat, Pfizer paid several hundred million dollars for Rinat, Pfizer Expands With Acquisition of Rinat*. We do not wish to use named entity anchors and this format limits the dataset to binary relations. Xu et al. (2014) use multi-instance learning to jointly model word and sentence relatedness in Twitter data, but require a large corpus of crowdsourced sentence similarity judgements. We do not want to invest in large numbers of sentence-level judgements when it is not certain how many word alignments involving event nouns could be subsequently learned.

Instead, we choose to capitalize on the fact that event nouns can co-refer with verbal descriptions of events, either anaphorically (backwards referring) or cataphorically (forwards referring). An example would be *The two countries* **fought** *savagely over their borders. The* **war** *lasted for years.* Identifying such pairs falls within the task of event coreference resolution (Bagga and Baldwin, 1999; Chen and Ji, 2009; Bejan and Harabagiu, 2014), but focuses on cases where one event mention is a noun. Moreover, we do not care about optimal clusterings of event mentions, but rather a set of candidates for related nouns and verbs, which can be manually filtered to create the dataset. For our pilot, we used a simple supervised method to identify event nouns, following Bel et al. (2010), and investigated the adjacent sentences for co-referring verbs.

3.1 Event Nouns

Our goal was a wide variety of event nouns covering various topics. We began with a small seed set of 73 positives (event nouns) and 94 negatives (non-event nouns), manually curated by Bel et al. (2010). We expanded the seed set using FrameNet (Fillmore and Baker, 2010), labeling nouns belonging to the *activity* or *process* classes as positive, and nouns belonging to the *entity* or *locale* classes as negative. This combination resulted in 1028 seed nouns, half positive and half negative (after downsampling the negatives).

We then bootstrapped additional nouns using the NYT portion of the Gigaword Corpus (Graff et al., 2005) by training an SVM on our seed set, using 126 syntactic features. This approach is similar to that of Bel et al. (2010), who trained a decision tree classifier with a dozen features. We made use of linguistic features previously found useful for identifying non-deverbal event nouns (Resnik and Bel, 2009; Bel et al., 2010), including the ability to occur as the subject of aspectual verbs (*the ceremony* **lasted** *for an hour*, *the meal* **began** *at 7:00*) and the object of temporal prepositions (**during** *the war*). The SVM achieved 78% accuracy using cross-validation on the seed set.

We used the SVM to classify 500 frequent nouns from NYT Gigaword that were not in our seed set. Of these, 286 were predicted as negative and 214 positive; we manually edited the positives down to 185. The resulting 699 positives were used for corpus extraction, and the 800 negatives will be used for confounders.

3.2 Corpus Extraction

After preprocessing NYT Gigaword, sentences containing positive event nouns were extracted. Expert annotators will see the extracted target sentences in random order, and each target sentence will be accompanied by its immediately preceding and following sentences, which will be inspected for co-referring verbs.

3.3 Two-Stage Annotation

Positive examples are still sparse among our candidate pairs. This leads us to propose a two-stage annotation process where the initial candidates are filtered by experts, after which the relatedness ratings are obtained by crowdsourcing. The goal of the first phase is for experts to choose phrase pairs that exhibit lexical and syntactic divergence, and appear to have moderate to high relatedness. The experts also shorten full sentences to phrases of at most 10 words.

Expert annotation can be a bottleneck for dataset creation. However, in cases where the source data is unbalanced, expert annotation can actually increase the potential size of the dataset, since funds are not wasted on crowdsourcing to rule out a large number of negatives. As mentioned above, the initial expert filtering also ensures high quality examples despite the potentially difficult concept of non-deverbal event nouns.

The authors have performed a short pilot of the expert annotation stage. In a couple of hours we produced approximately fifty positive examples, suggesting that in less than a month of part-time expert annotation we could produce a dataset of a few thousand pairs (including confounders; see Section 3.4) to proceed to crowdsourcing. The annotation guidelines developed for this pilot are shown in Figure 3. On a sample of the data we obtained inter-annotator agreement of 0.89, reported as $2P(A) - 1$ for unbalanced data (Eugenio and Glass, 2004). Table 1 provides a sample of phrase pairs that the annotators considered moderately or highly related.

3.4 Confounders

We propose two sets of confounders. The first set consists of standard low-relatedness pairs, created by shuffling related pairs, by pairing event nouns with unrelated adjacent sentences (the unrelated pairs from the expert annotation stage), and by pairing phrases centered around non-event nouns with adjacent sentences. Non-event noun phrases can be extracted from the corpus using our negatives list from (Bel et al., 2010), FrameNet, and bootstrapping. The data passed along for crowdsourcing will consist of the positives from expert annotation along with an equal number of confounders.

- Target sentence S_{targ} contains an event noun or noun phrase.
- Mark as a positive pair if S_{prev} or S_{next} contains a verb or verb phrase which is related in meaning to the noun or noun phrase in S_{targ}.
- Short phrases around the noun or verb can be considered in the relatedness decision.
  - Noun phrase can include an adjectival or nominal modifier, or short PP, which identifies the relevant sense (e.g. *welfare program* vs. *TV program*).
  - Event noun must be the head of the noun phrase, e.g. *earned income*, not *income tax*.
  - Verb phrase can include an object, other argument, or short PP, which identifies the relevant sense (e.g. *provide aid*).
- The noun (phrase) and verb (phrase) must be topically similar, but do not need to be paraphrases (e.g. *disease/diagnosis, disease/donate organ, trial/convict* are positives).
- Do not include antonymous related items, e.g. *loss/win*.
- Do not include cases where the noun and verb share a root, e.g. *fight/fight, presumption/presume*.
- Shorten the two sentences to phrases of maximum 10 words.

Figure 3: Annotation guidelines used for pilot expert annotation.

the comfort of a KLM **flight** from Belfast	they **returned** to their home in Northern Ireland
the peso **crisis** erupted	Mexican stocks **slipped**
he heads an **outreach program**	he **works with refugees**
starting a **workout program**	**walk** at a medium pace for an hour
we have won this **war**	vowing to **fight** to the death
passengers in New York have no **choice**	passengers can **decide** whether to avoid Kennedy
the **political battle** underlined the role that settlements play	Cabinet members **argued** that construction projects might be in jeopardy
enjoy a sound **meal**	**nibble** on snacks
Clinton gave a **speech**	the White House **announced** its members
the son died of heart **disease**	**donate** an organ for a family
a first-round playoff **loss**	**win** one last Super Bowl

Table 1: Sample candidates for highly and moderately related phrase pairs as judged by the authors, from pilot annotation. The counterpart noun and verb, with modifiers when relevant, are in bold.

The second proposed set of confounders is aimed at evaluating whether CDS models can distinguish between *relatedness* and *similarity* with regard to event descriptions. Here, we choose phrases centered around a common argument of a verb, but where the phrase does not describe the same event. For example, *the two countries* **fought** *savagely* might be paired with *many* **soldiers** *required training*, rather than *the* **war** *lasted for years*; or *we* **ate** *at a new restaurant* might be paired with *the art of making* **pizza**, rather than *the* **meal** *was delicious*. We conceive this as an alternative subset of the data, where the task is to assign a lower score to the phrases containing a non-event noun, a much harder task than simple relatedness. This task is a proxy for downstream applications such as event sequencing, narrative generation, and summarization, where it is necessary to identify when multiple phrases describe the same event. We emphasize that this confounder set is speculative; we expect that its development will be complex and will introduce interesting problems which will undoubtedly result in modifications to the approach as we work with the data.

4 Conclusion

SLEDDED is a targeted dataset of event descriptions which focuses on semantic relatedness under lexical and syntactic divergence. Although SLEDDED is aimed primarily at CDS, it would also be suitable for evaluating representations used for tasks such as Recognizing Textual Entailment (RTE) or Machine Translation (MT). We believe phrase relatedness tasks have continued potential for evaluating the next generation of vector space representations, if they are carefully designed to isolate the behavior of different representations under specific linguistic conditions.

Acknowledgments

This work is supported by ERC Starting Grant DisCoTex (306920). We thank David Weir and the participants in the IWCS 2015 Advances in Distributional Semantics Workshop for their suggestions on a previous version of this proposal, and the anonymous reviewers for their helpful comments.

References

Amit Bagga and Breck Baldwin. 1999. Cross-document event coreference: Annotations, experiments, and observations. In *Proceedings of the ACL Workshop on Coreference and its Applications*, page 18, College Park, MD.

Cosmin Adrian Bejan and Sanda Harabagiu. 2014. Unsupervised event coreference resolution. *Computational Linguistics*, 40:311–347.

Núria Bel, Maria Coll, and Gabriela Resnik. 2010. Automatic detection of non-deverbal event nouns for quick lexicon production. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 46–52, Beijing, China.

Razvan C. Bunescu and Raymond J. Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 576–583.

Zheng Chen and Heng Ji. 2009. Graph-based event coreference resolution. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 54–57, Singapore.

William Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30:95–101.

C. J. Fillmore and C. Baker. 2010. A frames approach to semantic analysis. In *The Oxford Handbook of Linguistic Analysis*, pages 791–816. Oxford University Press, Oxford.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2005. English gigaword second edition. LDC2005T12.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2014. A study of entanglement in a categorical framework of natural language. In *Proceedings of the 11th Workshop on Quantum Physics and Logic (QPL)*, Kyoto, Japan.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014a. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of SemEval*, pages 1–8.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014b. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC*.

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

Gabriela Resnik and Núria Bel. 2009. Automatic detection of non-deverbal event nouns in spanish. In *Proceedings of GL2009: 5th International Conference on Generative Approaches to the Lexicon*, Pisa.

Vasile Rus, Rajendra Banjade, and Mihai Lintean. 2014. On paraphrase identification corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from twitter. *TACL*, 2:435–448.