Data Statement for STR-2021:
A Dataset for Semantic Textual Relatedness

Mohamed Abdalla
Department of Computer Science
University of Toronto
msa@cs.toronto.edu

Krishnapriya Vishnubhotla
Department of Computer Science
University of Toronto
vkpriya@cs.toronto.edu

Saif M. Mohammad
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

October 2021

Project Homepage: https://arxiv.org/abs/1803.09010
Link to ArXiv: https://arxiv.org/abs/1803.09010
How to cite:

Title
Authors:
Venue:
Year:

Motivation

Q1. For what purpose was the dataset created?

The degree of semantic relatedness (or, closeness in meaning) between two units of language has long been considered fundamental to understanding meaning. It is also central to textual coherence and useful for computational applications such as question answering and summarization. However, much of the past NLP work has focused on semantic similarity (a much smaller subset of semantic relatedness), in no small part because of a paucity of relatedness datasets. Here for the first time, we created a dataset of semantic relatedness for sentence pairs. This dataset has 5,500 English sentence pairs manually annotated for semantic relatedness using comparative annotations (a setup that we show produces higher-annotation reliability). We use the dataset to explore a number of research questions on what makes two sentences more semantically related. We also evaluate a suite of sentence representation methods on their ability to place pairs that are more related closer to each other in vector space.

Q2. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This dataset was created by Mohamed Abdalla (PhD Candidate, University of Toronto), Krishnapriya Vishnubhotla (PhD Candidate, University of Toronto), Saif Mohammad (Senior Research Scientist, National Research Council of Canada) on behalf of ourselves.

Q3. Who funded the creation of the dataset?

Dataset annotation was funded by National Research Council Canada.

Q4. Any other comments?

None.

Composition

Q5. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

The dataset is composed of English sentence pairs and a human annotated relatedness score.

Q6. How many instances are there in total (of each type, if appropriate)?

There are 5500 sentence pairs with associated human relatedness scores in the dataset.

Q7. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The sentences are sampled from one of seven data sources. We chose to construct our dataset by sampling sentences from many sources to capture a wide variety of text in terms of sentence structure, formality, and grammaticality. Pairs of sentences were created from the source sentences in a number of ways as described ahead. The sources are:

1. **Formality** (Rao and Tetreault, 2018): Pairs of sentences having the same meaning but differing in formality (one formal, one informal).
2. **Goodreads** (Wan and McAuley, 2018): Book reviews from the Goodreads website.[1]
3. **ParaNMT** (Wieting and Gimpel, 2018): Paraphrases from a machine translation system.
4. **SNLI** (Bowman et al., 2015): Pairs of premises and hypotheses, created from image captions, for natural language inference.
5. **STS** (Cer et al., 2017): Pairs of sentences with semantic similarity scores. (Integer label responses, 0 to 5, from multiple annotators were averaged to obtain the similarity scores.)
6. **Stance** (Mohammad et al., 2016): Tweets labelled for both sentiment (*positive*, *negative*, *neutral*) and stance (*for*, *against*, *neither*) towards targets (e.g., *Donald Trump*, *Feminism*).
7. **Wikipedia** Text Simplification Dataset (Horn et al., 2014): Pairs of Wikipedia sentences and their simplified forms.

From each source, we sampled sentences that were between 5 and 25 words long. For the paraphrase datasets (Formality, ParaNMT, and Wikipedia), we obtained sentence pairs in two ways: by directly taking the paraphrase pairs (indicated by the suffix *pp*, and by randomly pairing sentences from two different paraphrase pairs (suffixed by *r*). The paraphrase pairs were selected at random from the source dataset, whereas the lexical overlap strategy was applied in the creation of the random pairs.

The number of sentences pairs sampled from each dataset is presented in Table 1. We discuss how pairs were selected in more detail in the “Collection Proccess” section below.

Q8. What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features?

Each instance contains a pair of two English sentences.

Q9. Is there a label or target associated with each instance?

[1] We only accessed sentences from the ‘Fantasy and Paranormal’ genre, since it contained the most reviews per book (and thus potentially rich in related sentence pairs).

Types of Pairs	Key Attributes	# pairs
1. Formality	paraphrases, style	1000
Formality_pp	paraphrases, differ in style	300
Formality_r	random pairs	700
2. Goodreads	reviews, informal	1000
3. ParaNMT	automatic paraphrases	750
ParaNMT_pp	automatic paraphrases	450
ParaNMT_r	random pairs	300
4. SNLI	captions of images	750
5. STS	have similarity scores	250
6. Stance	tweet pairs with same hashtag, less grammatical	750
7. Wikipedia	formal	1000
Wiki_pp	paraphrases, formal	500
Wiki_r	random pairs, formal	500
ALL		5500

Table 1: Summary of sentence pair types in STR-2021.

With each sentence pair, we provide a human annotated semantic relatedness score. The score presents how related two sentences are on a scale of 0–1 relative to other sentence pairs. Note: The scores are not absolute. The annotations are relative to other sentence pairs in the dataset.

Q10. Is any information missing from individual instances?

No.

Q11. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

No.

Q12. Are there recommended data splits (e.g., training, development/validation, testing)?

We recommend using 5-fold cross validation. We have provided fold-labels indicating the folds we used. If you must use a single train–test split, then we recommend using one or more of the five folds for creating such splits and reporting how the splits were created. This will aid comparison of results.

Q13. Are there any errors, sources of noise, or redundancies in the dataset?

None that we are aware of. Please contact us if you discover any.

Q14. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained.

Q15. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)?

No.

Q16. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No.

Q17. Does the dataset relate to people?

The individual sentences were written by people. In this way, the dataset relates to people.

Q18. Does the dataset identify any subpopulations (e.g., by age, gender)?

No.

Q19. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

No.

Q20. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

No.

Q21. Any other comments?

None.

Collection Process

Q22. How was the data associated with each instance acquired?

The sentences were sampled from the sources listed earlier.

Stance Data

We created 750 sentence pairs from Mohammad et al. (2016)'s dataset of tweets labeled for stance. The original dataset is composed of individual tweets labelled for both stance ('For', 'Against', 'Neither Inference Likely') and sentiment ('Positive', 'Negative', 'Neutral'). The dataset was built from tweets focused on six targets: 'Atheism', 'Climate Change', 'Donald Trump', 'Feminism', 'Hillary Clinton', 'Abortion'.

When curating our sentence pairs, we limited the possible targets to 'Hillary Clinton', 'Donald Trump', and 'Abortion'. Sentence pairs were chosen such that both sentences shared the same target. 500 sentence pairs shared their stance towards their target (250 for and 250 against). 250 sentences pairs differed on their stance. We did not use any lexical overlap heuristic to specify which tweets should be paired with each other because we were interested in studying if overlap in topic was a strong enough signal to impact relatedness. That is, by choosing pairs with the same target, we were already pre-selecting for various degrees of relatedness.

SNLI Data

We created 750 sentence pairs from the Stanford Natural Language Inference (SNLI) Dataset (Bowman et al., 2015). SNLI is composed of image description captions; for each caption, multiple premise sentences are generated, along with multiple possible hypotheses sentences that could possibly belong to each premise. To build our sentence pairs we sought to pair different premise sentences together. We did not wish to pair between premise and hypothesis sentences as the sentence structure was significantly different (and simpler for the hypothesis sentences), as noted by the creators of the dataset. Even still, the majority of premise sentences are very short (with a mean token count of 14), often following very simple (and similar) grammatical structure.

To generate the sentence pairs, first we removed all sentences with less than 5 or more than 25 tokens. Then, for each token in all remaining sentences, we replaced each token with it's most frequent synonym, using Roget's thesaurus (Roget, 1911), to define synonymous relationships. Words which did not have synonyms were left unchanged. The intention behind replacing each word with its most frequent synonym was to ensure that synonymous phrasings would count as overlaps when we measure it. We then randomly selected 750 sentences to serve as the first sentence of our final pairings. To find the second sentence to each pairing we looped through all premise sentences and returned the first sentence which satisfied two conditions: 1) The unigram overlap was greater than or equal to 25% and less than 75% of the first sentence, and 2) the difference in length between both sentences did not exceed 25%.

Wikipedia Data

We sampled 1000 sentence pairs from a dataset that pairs sentences from English Wikipedia with sentences from Simple English Wikipedia. Created to enable the task of sentence simplification the paired sentences, paired using rules-based classification, are often very closely related. We used this dataset in two ways: 1. Extracting sentence pairs which serve as paraphrases or near paraphrases (we refer to these as Wiki_pp), and 2. pairing sentences to other random sentences in the dataset (we refer to these as Wiki_r).

Wiki_PP: First, we removed any pairings for which either sentence was less than 5 words or more than 25 words. Then we narrowed the list of pairings further by removing any pairings that did not share more than 25% and less than 75% of unique unigrams. From the remaining sentence pairs, we randomly selected 500 paired sentences.

Wiki_R: Here, we only make use of the full sentences from the original Wikipedia, discarding sentences from Simple Wikipedia. We remove all sentences that have less than 5 or more than 25 tokens. To create the sentence pairs, we loop in a random order through all possible pairing of sentences. We pair two sentences if they share at least 25% of their tokens and less than 75% of their tokens AND the difference in length between both sentences did not exceed 25%. We stop once we have generated 500 sentence pairs.

Goodreads Data

We created 1000 sentence pairs from the UCSD Goodreads Dataset (Wan and McAuley, 2018; Wan et al., 2019), which has book reviews from the Goodreads website. We limit the sampling to the 'Fantasy and Paranormal' genre, since it contained a relatively higher number of reviews per book, allowing for a higher possibility of sampling more related sentence pairs. Each review was first split into sentences using the default NLTK sentence tokenizer; we keep only those sentences with the number of tokens between 5 and 25. We then randomly examine pairs of sentences, and quantify the lexical overlap between then with an IDF-weighted Dice overlap score. The pairs are then assigned to buckets based on this overlap score; the range of each bucket is obtained by first finding 50 equally-spaced percentiles of the entire score distribution. We then sample exponentially increasing number of sentences from low to high weighted Dice overlap bins such that a total of 1000 sentence pairs are included.

ParaNMT Data

ParaNMT (Wieting and Gimpel, 2018) is a dataset of 51 million sentential paraphrases that were automatically generated using a neural machine translation system. We generated two sets of pairs from these sentences corresponding to paraphrases and random pairs:

ParaNMT_PP: We assign paraphrases to buckets based on the Dice score between the two sentences. We divided the range of scores into 100 equally-sized percentiles. We then sample pairs uniformly from each bucket, for a total of 450 sentence pairs.

ParaNMT_R: For the random, non-paraphrase sentence pairings, we use the Dice score to extract 300 pairs, analogous to the creation of the **Wiki_R** pairs.

Formality Data

Our third paraphrase corpus is the Formality dataset from Rao and Tetreault (2018) (They refer to it as GYAFC). This consists of human-written formal and informal paraphrases for sentences sourced from the Yahoo! Answers platform. Our sampling procedure for this dataset follows that of the ParaNMT dataset.

Formality_PP: We assign sentences to one of 50 buckets based on their lexical overlap score as before. We then uniformly sample from each bucket to extract 300 sentence pairs.

Formality_R: We sample random pairings of sentences that define the token overlap and length difference conditions as defined for Wiki_R and ParaNMT_R. We extract 700 such sentence pairs.

Relatedness Annotations

From the list of 5,500 sentence pairs, we generated 11,000 unique 4-tuples (each 4-tuple consists of 4 distinct sentence pairs) such that each sentence pair occurs in around eight 4-tuples. The tuples were generated using Best–Worst Scaling as described in the associated publication.

Q23. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

The sentences were sampled from existing public datasets available for download. The annotations were done using Amazon Mechanical Turk.

Q24. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The sampling strategy is described in Q22.

Q25. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

We used Amazon Mechanical Turk (a crowdsourcing platform) for obtaining relatedness annotations. This project was approved by the first author's Institutional Research Ethics Board (Protocol#:40736). The annotators were based in the United States of America and were paid the federal minimum wage of $7.25.

Q26. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?

The sampled sentences come from texts found online from 2000–2020. The human annotations were collected in 2021.

Q27. Were any ethical review processes conducted (e.g., by an institutional review board)?

This project was approved by the University of Toronto Research Ethics Board (Protocol#:40736).

Q28. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Q29. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A.

Q30. Were the individuals in question notified about the data collection?

N/A.

Q31. Did the individuals in question consent to the collection and use of their data?

N/A.

Q32. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

N/A.

Q33. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

N/A.

Q34. Any other comments?

None.

Preprocessing/cleaning/labeling

Q35. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

There was no special cleaning or preprocessing done to the data.

Q36. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

N/A.

Q37. Is the software used to preprocess/clean/label the instances available?

N/A.

Q38. Any other comments?

None.

Uses

Q39. Has the dataset been used for any tasks already?

This dataset was first used in the publication detailing its release. In the publication, we explored a number of research questions on what makes two sentences more semantically related. We also evaluated a suite of sentence representation methods on their ability to place pairs that are more related closer to each other in vector space.

Q40. Is there a repository that links to any or all papers or systems that use the dataset?

No.

Q41. What (other) tasks could the dataset be used for?

In our eyes, this dataset is best suited to exploring human notions of meaning or semantic relatedness. It can also be used for evaluating approaches that attempt to emulate this notion. Note, however, that because the dataset is of a limited size and drawn from a certain population at a certain time, evidence from using this dataset should be used as one among many to draw broader conclusions about language.

Q42. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Any dataset of semantic relatedness entails several ethical considerations. We list some notable ones below:

- *Coverage:* We sampled English sentences from a diverse array of sources from the internet, with a focus on social media. Yet, it is likely that several types of sentences (and several demographic groups) are not well-represented in STR-2021. The dataset likely includes more sentences by people from the United States and Europe and with a socio-economic and educational background that allows for social media access.
- *Not Immutable:* The relatedness scores do not indicate an inherent unchangeable attribute. The relatedness can change with time, but the dataset entries are largely fixed. They pertain to the time they are created.
- *Socio-Cultural Biases:* The annotations of relatedness capture various human biases. These biases may be systematically different for different socio-cultural groups. Our data was annotated by US annotators, but even with the US there are different socio-cultural groups.
- *Inappropriate Biases:* Our biases impact how we view the world, and some of the biases of an individual may be inappropriate. For example, one may have race or gender-related biases that may percolate subtly into one's notions of how related two units of text are. Our dataset curation was careful to avoid sentences from problematic sources, and we have not seen any inappropriate relatedness judgments, but it is possible that some subtle inappropriate biases still remain. Thus, as with any approach for sentence representation or semantic relatedness, we caution users to explicitly check for such biases in their system regardless of whether they use STR-2021.
- *Perceptions (not "right" or "correct" labels):* Our goal here was to identify common perceptions of semantic relatedness. These are not meant to be "correct" or "right" answers, but rather what the majority of the annotators believe based on their intuitions of language.
- *Relative (not Absolute):* The absolute values of the relatedness scores themselves have no meaning. The scores help order the sentence pairs relative to each other. For example, a pair with a higher relatedness score should be considered more related than a pair with with a lower score. Further, no claim is made that the mid-point (relatedness score of 0.5) separates related words from unrelated words. One may determine categories such as *related* or *unrelated* by finding thresholds of relatedness scores optimal for their use/task.

The authors welcome feedback on further considerations, especially based on the use of the dataset.

Q43. Are there tasks for which the dataset should not be used?

Users of this dataset should familiarize themselves with the limitations defined above and ensure that these limitations are not problematic for their particular use case. We do not recommend commercial use or deployment of systems trained on this dataset. Any such use must first be thoroughly examined to ensure that any of the biases in the dataset do not negatively affect people. Contact the authors to obtain permission if you wish to use this dataset for commercial purposes.

Q44. Any other comments?

None.

Other

This is an early release version of the datastatement. Further details about distribution, maintenance, and other aspects of the data will be added in the near future.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a lexical simplifier using wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 458–463.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140.

Peter Mark Roget. 1911. *Roget's Thesaurus of English Words and Phrases...* TY Crowell Company.

Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.