SemRel 2024: A Collection of Semantic Textual Relatedness Datasets for 13 Languages

Nedjma Ousidhoum*, Shamsuddeen Hassan Muhammad*, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, Saif M. Mohammad

Semantic Textual Relatedness (STR)

There was a lemon tree next to the house	*I have a green hat*
I am feeling sick	*Get well soon*

- Central to understanding meaning in text.
- STR involves Semantic Textual Similarity (**STS**) and all **commonalities** between two units of text: **topic, view, time period**, **…**
- Previous work focused on semantic textual similarity (STS) except the STR dataset by Abdalla et al. (2023) that focuses on English.
- Applications include evaluating sentence representation, QA, summarisation.

Dataset Creation: Key Steps

- **Create instances**
  - Identify data sources.
    - Wikipedia, social media, reviews, …
  - Extract and **pair sentences.**
- **Comparative annotation** using Best-Worst Scaling (BWS).
- **Quality control** sanity check and postprocessing.

Lang	afr	amh	arb	arq	ary	eng	esp	hau	hin	ind	kin	mar	tel
Size	700	1,258	627	1,941	1,421	8,350	2,302	2,551	1,256	504	1,102	1,791	1,597
SHR	0.85	0.90	0.86	0.64	0.77	0.82	0.70	0.74	0.93	0.68	0.74	0.94	0.87

Data Selection

We pair average-length sentences from various sources using heuristics such as lexical overlap, contiguity, (MT) paraphrases, STS, random selection, manual checks.

Data Annotation using BWS

That's difficult. They're both great. That's hard they as they are great!	👍
That's difficult. I think it's easy.	
There is a tree next to the house. I love reading next to the tree.	
I was travelling. She bought a new phone.	👎

Choose the **most related** and the **least related** pair from **a tuple of 4 sentence pairs**.

Data Aggregation

We generate real-valued scores based on **the number of times a pair was chosen as best** and **the number of times it was chosen worst.**

Datasets

Team effort to build datasets for 13 languages from 5 language families.
The datasets were used in a SemEval shared task (with >160 participants).

Experiments

- Determine relatedness scores.
- Assess how well do system-predicted rankings of test instances align with human judgments using Spearman rank correlation.
- Baselines Lexical Overlap, Supervised, Unsupervised, and Crosslingual settings.

Insights

Results show limitations of current multilingual and language-specific language models.

Performance highly language dependent.