Patents Phrase to Phrase Semantic Matching Dataset

SIGIR PatentSemTech 2022 Workshop

Grigor Aslanyan

Ian Wetherbee

Introduction

Google

- Semantic Textual Similarity (STS) measures how similar two pieces of text are.
- There are multiple benchmark datasets for STS (e.g. STS-B, SICK, MRPS, PIT) but they are all general purpose.
- We are publicly releasing a new phrase-phrase dataset:
  - Human rated
  - Contextual
  - Focused on technical terms from patents
  - Includes similarity scores
  - Includes granular ratings (e.g. synonym, antonym, hypernym)
- The dataset was successfully used in the U.S. Patent Phrase to Phrase Matching Kaggle Competition:
  - Ran from March 21 - June 20, 2022
  - Had about 2,000 participants from all over the world

Dataset Focus

Google

- **Phrase disambiguation**: certain phrases can have multiple different meanings (e.g. “mouse”).
  - We include Cooperative Patent Classification (CPC) classes as context.
- **Adversarial keyword match**: there are phrases that have matching keywords but are otherwise unrelated (e.g. “container section” → “kitchen container”, “offset table” → “table fan”).
  - We include many such examples in our data.
- **Hard negatives**: We created our dataset with the aim to improve upon current state of the art language models.
  - We have used the BERT model to generate some of the target phrases.

Data Sample

Google

anchor	target	context	rating	score
acid absorption	absorption of acid	B08	exact	1.00
acid absorption	acid immersion	B08	synonym	0.75
acid absorption	chemically soaked	B08	domain	0.25
acid absorption	acid reflux	B08	not rel.	0.00
gasoline blend	petrol blend	C10	synonym	0.75
gasoline blend	fuel blend	C10	hypernym	0.50
gasoline blend	fruit blend	C10	not rel.	0.00
faucet assembly	water tap	A22	hyponym	0.50
faucet assembly	water supply	A22	holonym	0.25
faucet assembly	school assembly	A22	not rel.	0.00

Dataset Size

- **48,548** phrase pairs.
- **973** unique anchors.
- **106** different context CPC classes.
- Split (all the pairs with the same anchor go into the same split):
  - Training - 75%
  - Validation - 5%
  - Test - 25%

Dataset Generation

Google

- For each patent extract important (salient) terms, typically either:
  - Noun phrases (e.g. “fastener”, “lifting assembly”), or
  - Functional phrases (e.g. “food processing”, “ink printing”).
- Keep only phrases that appear in at least 100 patents.
- Randomly sample about 1,000 from the remaining phrases - these become the anchor phrases.
- For each anchor find all the matching patents and their CPCs, randomly sample up to 4 CPC classes for context.
- For each anchor generate targets:
  - Partial match - randomly select phrases (from the entire corpus) with a partial keyword match with the anchor (e.g. “abatement” → “noise abatement”, “material formation” → “formation material”)
  - Masked Language Model (MLM) - find sentences from patent text containing the anchor phrase, mask it out, and let BERT generate candidate phrases for the mask (we use the Patent-BERT model).

Ratings

Google

- Very high (exact)
- High (close synonym)
- Medium
  - Hyponym (broad-narrow)
  - Hypernym (narrow-broad)
  - Structural match
- Low
  - Antonym
  - Meronym (part of)
  - Holonym (whole of)
  - Domain related
- Not related

Ratings

Google

- Each pair is rated independently by two raters. Afterwards they met and discussed disagreements and came up with final ratings.
- Each rater also generated new target phrases with different ratings.

Baselines

Google

- Baselines computed with pretrained models:
  - Context is ignored
  - Dual tower architecture - embed anchor and target, compute cosine similarity
  - Use mean pooling of individual keyword embeddings
- Results on the test set:

Model	Dim.	Pearson cor.	Spearman cor.
GloVe	300	0.429	0.444
FastText	300	0.402	0.467
Word2Vec	250	0.437	0.483
BERT	1024	0.418	0.409
Patent-BERT	1024	0.528	0.535
Sentence-BERT	768	0.598	0.577

Kaggle Results

Google

Prize Winners

#	△	Team	Members	Score	Entries	Last	Code
1	▲ 2	gezi		0.8782	100	16d	< >
2	▲ 2	Z & T		0.8775	82	16d	< >
3	▼ 1	S.X.R.D(gsdata)		0.8772	119	16d	< >
4	▼ 3	Ria~		0.8771	91	16d	
5	—	prompt is all you need		0.8766	87	16d	
6	—	hyd		0.8765	219	17d	
7	—	vialactea		0.8761	80	16d	
8	—	Team N		0.8757	104	16d	< >
9	▲ 1	Q. S.		0.8753	55	16d	
10	▼ 1	We did it!		0.8750	406	16d	

Google

Thank you!

Questions?