DISTILLING EXPLAINABLE SEMANTIC TEXTUAL SIMILARITY FUNCTIONS FROM PRETRAINED TRANSFORMERS

Henri Iser

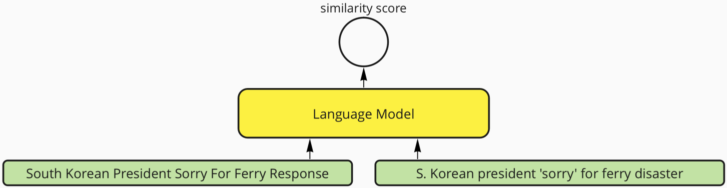
Supervisor: Eduardo Brito

December 6, 2021

University of Bonn

Sentence Similarity

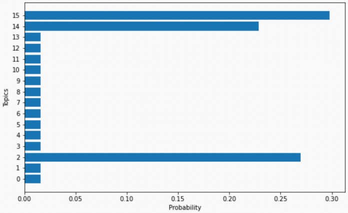
- *SOTA* model lack explainability
- Too slow in some setups


similarity score
Language Model
South Korean President Sorry For Ferry Response
S. Korean president 'sorry' for ferry disaster

Explainable Semantic Features

- Topic modeling
  - Latent-dirichlet-Allocation (LDA)[1]
  - Anchored Correlation Explanation[2]

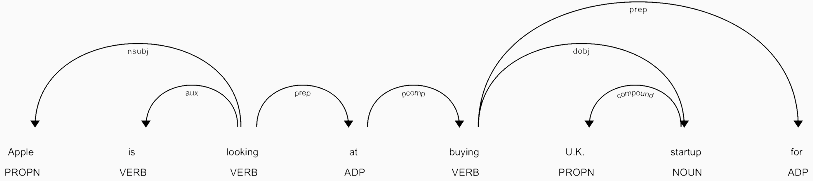
Men are running a race. ——— LDA ——→



[1] Blei, Ng, and Jordan, 2003 [2] Gallagher et al., 2017

Explainable Semantic Features

- Topic modeling
    - Latent-dirichlet-Allocation (LDA)[1]
    - Anchored Correlation Explanation[2]
- Part-Of-Speech (POS) tags



[1] Blei, Ng, and Jordan, 2003 [2] Gallagher et al., 2017

Explainable Semantic Features

- Topic modeling
  - Latent-dirichlet-Allocation (LDA)[1]
  - Anchored Correlation Explanation[2]
- Part-Of-Speech (POS) tags
- Regular Expressions
- other ...

[1] Blei, Ng, and Jordan, 2003 [2] Gallagher et al., 2017

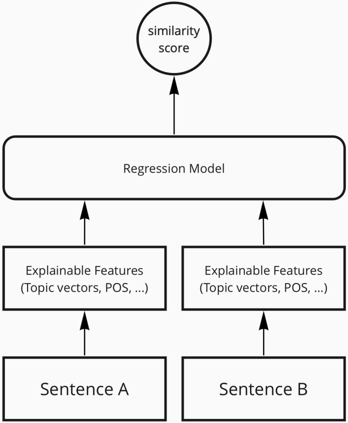
Explainable Semantic Features

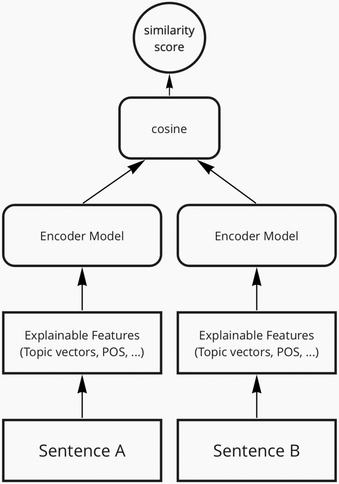
- Topic modeling
  - Latent-dirichlet-Allocation (LDA)[1]
  - Anchored Correlation Explanation[2]
- Part-Of-Speech (POS) tags
- Regular Expressions
- other …

How to combine these features?

[1] Blei, Ng, and Jordan, 2003 [2] Gallagher et al., 2017

Framework


similarity score
Regression Model
Explainable Features (Topic vectors, POS, ...)
Explainable Features (Topic vectors, POS, ...)
Sentence A
Sentence B


similarity score
cosine
Encoder Model
Encoder Model
Explainable Features (Topic vectors, POS, ...)
Explainable Features (Topic vectors, POS, ...)
Sentence A
Sentence B

Data

Datasets:

- STS benchmark[1]
- Quora question pairs[2]
- BWS Argument Similarity Corpus[3]
- Microsoft Research Paraphrase Corpus[4]

[1] http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark

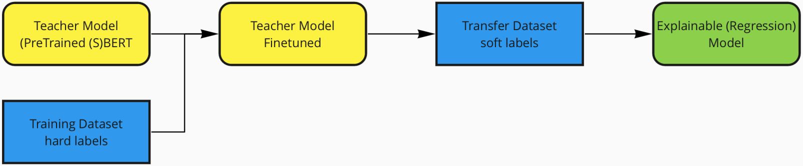
[2] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pair

[3] https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2496.2

[4] https://github.com/wasiahmad/paraphrase_identification

Data Augmentation

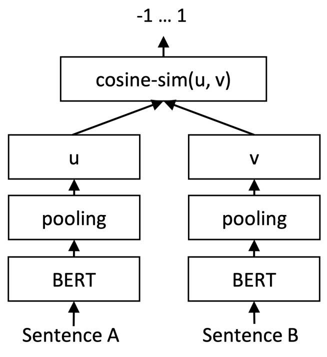
- i.e. train set of STS benchmark contains only #5552 scored sentence pairs.
- We need more data to train our model
- Use Pre-Trained (Sentence)-BERT Model to create soft labels



Data Augmentation

Sentence-BERT as Teacher Model

- Outperforms *SOTA* sentence embedding methods
- High efficiency
- Maintains BERT's accuracy

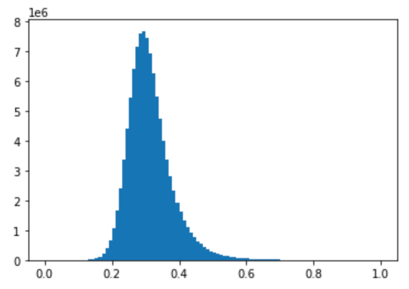


Reimers and Gurevych, 2019

Data Augmentation

Sentence-BERT as Teacher Model

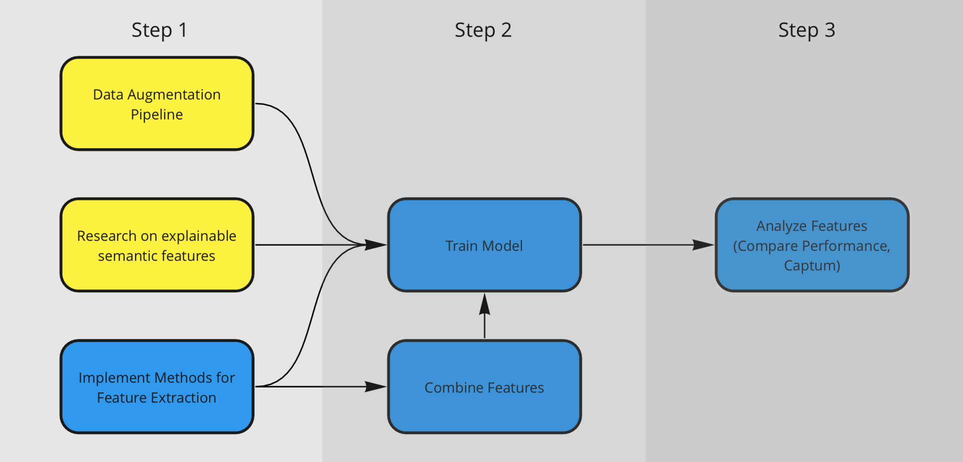
- Outperforms *SOTA* sentence embedding methods
- High efficiency
- Maintains BERT's accuracy



Goals of this Work

- Evaluate similarity scoring task using different semantic features
- Compare explainable model against *SOTA*
  - Performance
  - Runtime
- Analyse effect of each feature on performance

Roadmap


Step 1
Step 2
Step 3
Data Augmentation Pipeline
Research on explainable semantic features
Implement Methods for Feature Extraction
Train Model
Combine Features
Analyze Features (Compare Performance, Captum)

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (Mar. 2003). “Latent Dirichlet Allocation”. In: *J. Mach. Learn. Res.* 3.null, pp. 993–1022. ISSN: 1532-4435.

Gallagher, Ryan et al. (2017). “Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge”. In: *Transactions of the Association for Computational Linguistics* 5.0, pp. 529–542. ISSN: 2307-387X. URL: `https://transacl.org/ojs/index.php/tacl/article/view/1244`.

Reimers, Nils and Iryna Gurevych (2019). “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *CoRR* abs/1908.10084. arXiv: `1908.10084`. URL: `http://arxiv.org/abs/1908.10084`.

Questions?