14th September 2023

This Is Not Correct! Negation-Aware Evaluation Of Language Generation Systems

Miriam Anschütz, Diego Miguel Lozano, Georg Groh

Technical University of Munich
School of Computation, Information and Technology
Research Group Social Computing

1

Motivation

Modern embeddings are great.

Modern embeddings are great.

But almost always **fail to capture negation**.

Example

Specialized Models

I like rainy days because they make me feel relaxed.

I **don't** like rainy days because they **don't** make me feel relaxed.

BERTScore	microsoft/deberta-xlarge-mnli	0.879
Sentence Transformers	all-mpnet-base-v2	0.879
Universal Sentence Encoders	en_use_lg	0.900

Example
LLMs

I like rainy days because they make me feel relaxed.

I **don't** like rainy days because they **don't** make me feel relaxed.

text-embedding-ada-002	Open AI	0.948
embed-multilingual-v2.0	Cohere	0.989
textembedding-gecko@001	en_use_lg	0.935

We set as our **goal** improving the **sensitivity of embeddings towards negations**.

2

Negation-focused Dataset

Negation-focused Dataset

CANNOT

Compilation of ANnotated, Negation-Oriented Text-pairs

CANNOT Dataset
Format

The dataset is given as a .tsv file with the following structure:

premise	hypothesis	label
A sentence.	An equivalent, non-negated sentence (paraphrased).	0
A sentence.	The sentence negated.	1

CANNOT Dataset

Construction

The dataset created by cleaning up and merging the following datasets:

- *Not another Negation Benchmark: The NaN-NLI Test Suite for Sub-clausal Negation [1]*
- *GLUE Diagnostic Dataset [2]*
- *Automated Fact-Checking of Claims from Wikipedia [3]*
- *From Group to Individual Labels Using Deep Features [4]*
- *It Is Not Easy To Detect Paraphrases: Analysing Semantic Similarity With Antonyms and Negation Using the New SemAntoNeg Benchmark [5]*

CANNOT Dataset
Construction

Once processed, the number of remaining samples in each of the datasets are:

Dataset	**Samples**
Not another Negation Benchmark	118
GLUE Diagnostic Dataset	154
Automated Fact-Checking of Claims from Wikipedia	14,970
From Group to Individual Labels Using Deep Features	2,110
It Is Not Easy To Detect Paraphrases	8,597
Total	**25,949**

CANNOT Dataset

Construction

Additionally:

For each negated sample, a pair of non-negated sentences has been added by **paraphrasing** them with the pre-trained model 🤗`tuner007/pegasus_paraphrase`.

CANNOT Dataset
Construction

Additionally:

For each negated sample, a pair of non-negated sentences has been added by **paraphrasing** them with the pre-trained model 🤗`tuner007/pegasus_paraphrase`.

The **swapped version** of each pair (premise ⇆ hypothesis) has been included.

CANNOT Dataset

Construction

Additionally:

For each negated sample, a pair of non-negated sentences has been added by **paraphrasing** them with the pre-trained model 🤗`tuner007/pegasus_paraphrase`.

The **swapped version** of each pair (premise ⇆ hypothesis) has been included.

Duplicates have been removed.

CANNOT Dataset

Construction

With this, the CANNOT dataset currently contains **77,376 samples**.

It is publicly available on GitHub and the HuggingFace Hub.

3

Rule-based Negator

Rule-based Negator

The data included in *From Group to Individual Labels Using Deep Features* is **not related to negation**.

It simply contains sentences labelled with positive or negative sentiment.

Rule-based Negator

In order to obtain a **high number of negated-pairs** in a **fast and cheap way**, we developed a rule-based negator.

Rule-based Negator

In order to obtain a **high number of negated-pairs** in a **fast and cheap way**, we developed a rule-based negator.

Negate: A Python module to negate sentences

Rule-based Negator

Usage

```
pip install -U negate
```

```python
from negate import Negator

negator = Negator()
sentence = "An apple a day, keeps the doctor away."
negated_sentence = negator.negate_sentence(sentence)
print(negated_sentence)  # "An apple a day, doesn't keep the doctor away."
```

Rule-based Negator

v1.0.0 – Current State

Works **correctly** for most sentences!

However, only **verbal negations** are supported.

Rule-based Negator

v1.0.0 – Current State

Works **correctly** for most sentences!

However, only **verbal negations** are supported.

Inversions (questions) coming soon!

4

Negation-aware evaluation metric

Rule-based Negator

Existing metrics are **insensitive towards negation**.

Reference: An apple a day, keeps the doctor away.

Candidate: An apple a day, **doesn't** keep the doctor away.

BERTScore: 0.98

COMET: 0.81

Rule-based Negator

Existing metrics are **insensitive towards negation**.

Reference: An apple a day, keeps the doctor away.

Candidate: An apple a day, **doesn't** keep the doctor away.

`BERTScore: 0.98`

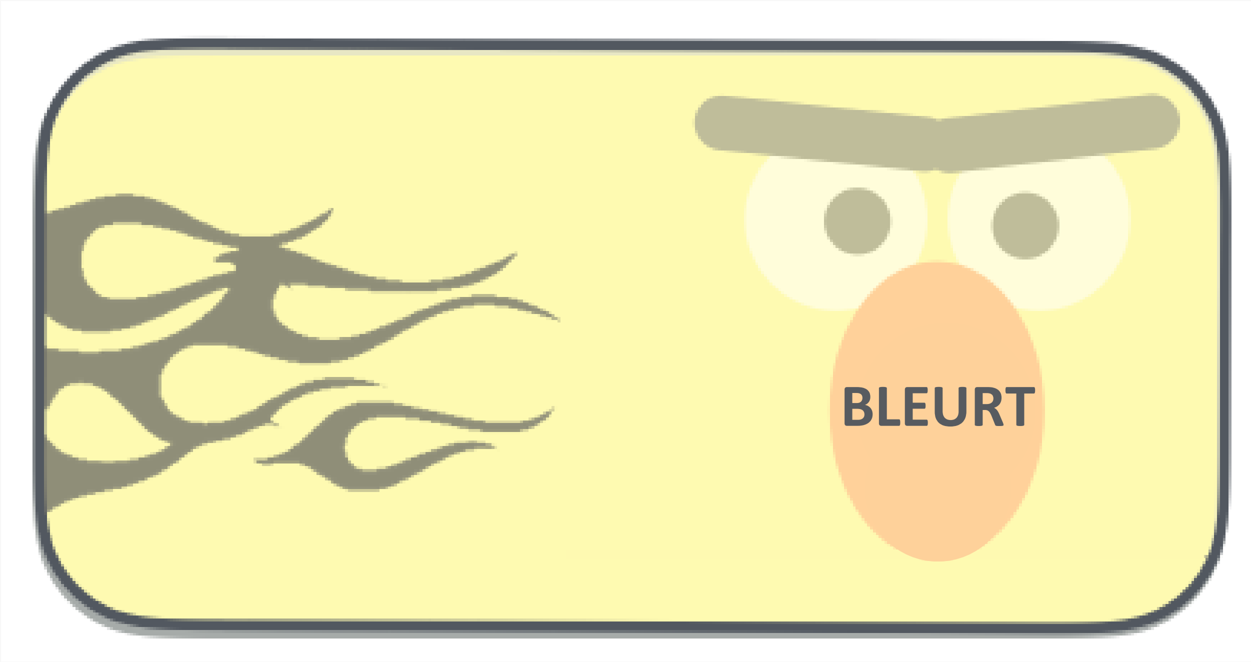
`COMET: 0.81`

We built such a negation-aware metric.

Fine-tuning BLEURT

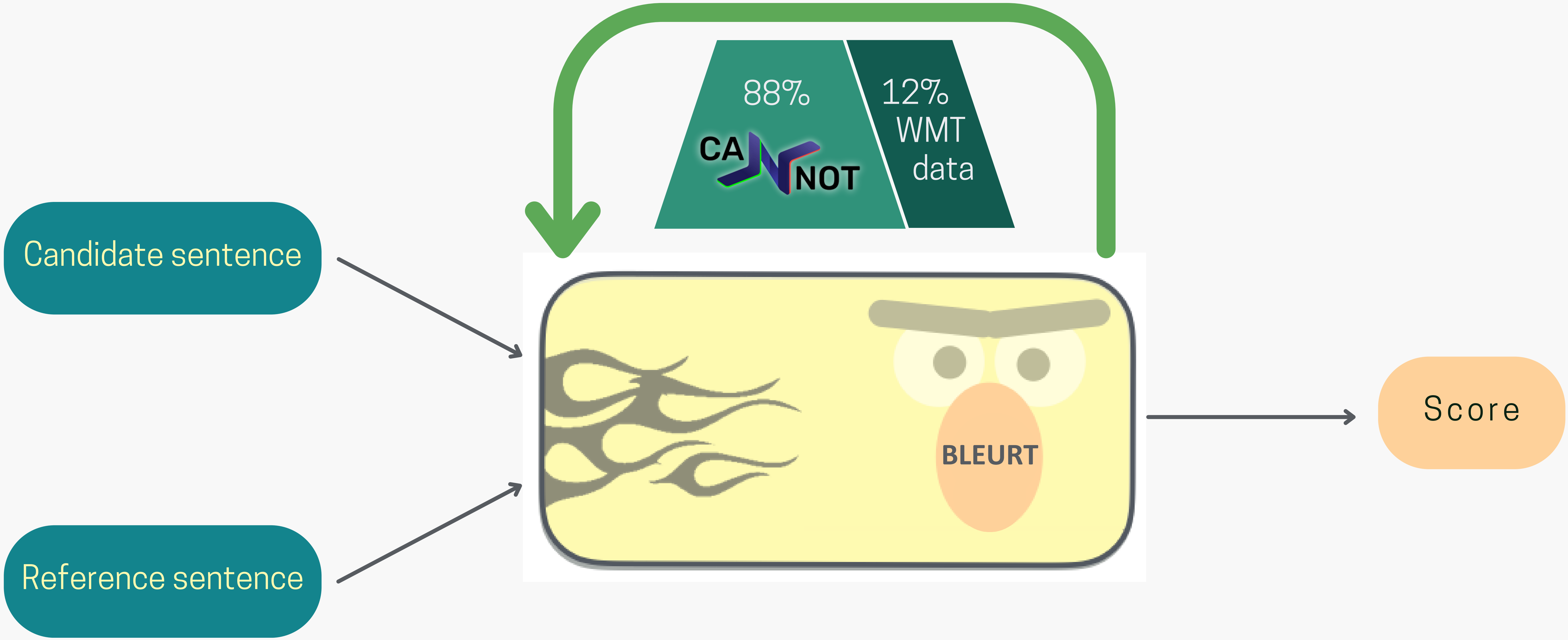
Candidate sentence

Reference sentence

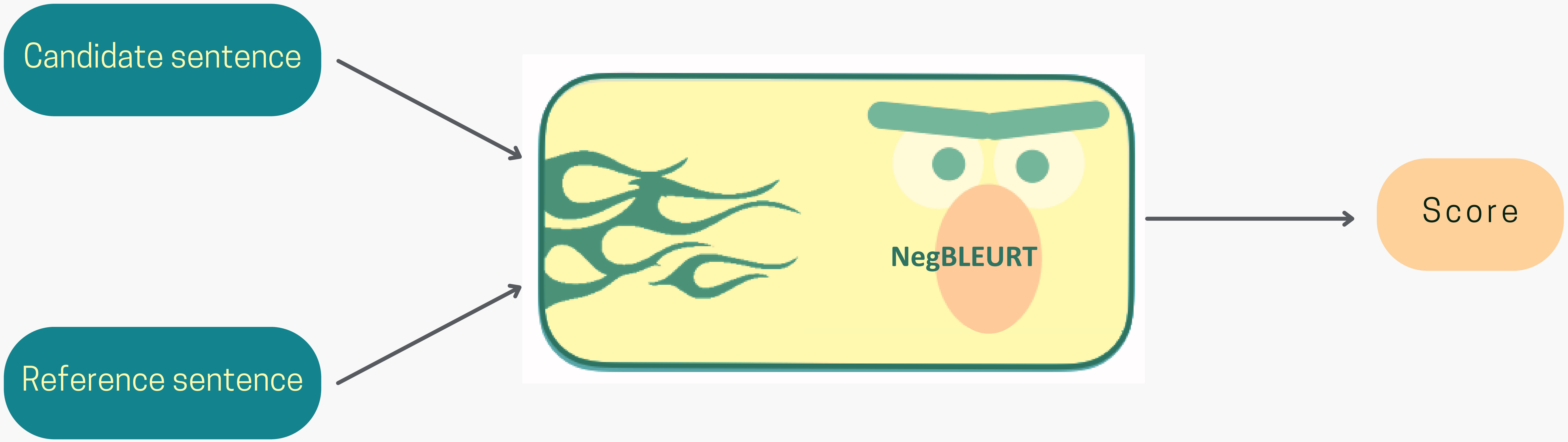


Score

Fine-tuning BLEURT



NegBLEURT



NegBLEURT

Reference: An apple a day, keeps the doctor away.

Candidate: An apple a day, **doesn't** keep the doctor away.

BERTScore: 0.98

COMET: 0.81

NegBLEURT: 0.45

5

Evaluation

DEMETR [6] metric benchmark

Perturbation	BARTScore*	BERTScore*	BLEURT20*	COMET*	PRISM*	NegBLEURT
base_shuffled	0.44	1.7	0.46	0.88	0.54	0.05
base_unrelated_trans	1.1	2.2	0.81	0.62	0.62	1.64
critical_addition	0.032	0.12	0.065	0.076	0.043	0.18
critical_antonym	0.043	0.15	0.088	0.098	0.044	0.38
critical_codemix	0.052	0.58	0.1	0.23	0.056	0.55
critical_gender	0.023	0.067	0.1	0.093	0.031	0.02
critical_ne_removed	0.14	0.31	0.15	0.17	0.084	0.17
critical_ne_replaced	0.18	0.37	0.2	0.18	0.12	0.38
critical_negation	0.058	0.21	0.15	0.15	0.053	0.93
critical_noun_removed	0.057	0.25	0.14	0.18	0.055	0.1
critical_numbers_replaced	0.07	0.044	0.052	0.01	0.046	0.09
critical_removed_adj_adv	0.045	0.1	0.047	0.052	0.034	0.07
critical_subj_removed	0.082	0.25	0.13	0.16	0.062	0.17
critical_verb_removed	0.031	0.21	0.12	0.17	0.043	0.09

DEMETR [6] metric benchmark

Perturbation	BARTScore*	BERTScore*	BLEURT20*	COMET*	PRISM*	NegBLEURT
base_shuffled	0.44	1.7	0.46	0.88	0.54	0.05
base_unrelated_trans	1.1	2.2	0.81	0.62	0.62	1.64
critical_addition	0.032	0.12	0.065	0.076	0.043	0.18
critical_antonym	0.043	0.15	0.088	0.098	0.044	0.38
critical_codemix	0.052	0.58	0.1	0.23	0.056	0.55
critical_gender	0.023	0.067	0.1	0.093	0.031	0.02
critical_ne_removed	0.14	0.31	0.15	0.17	0.084	0.17
critical_ne_replaced	0.18	0.37	0.2	0.18	0.12	0.38
critical_negation	0.058	0.21	0.15	0.15	0.053	0.93
critical_noun_removed	0.057	0.25	0.14	0.18	0.055	0.1
critical_numbers_replaced	0.07	0.044	0.052	0.01	0.046	0.09
critical_removed_adj_adv	0.045	0.1	0.047	0.052	0.034	0.07
critical_subj_removed	0.082	0.25	0.13	0.16	0.062	0.17
critical_verb_removed	0.031	0.21	0.12	0.17	0.043	0.09

DEMETR [6] metric benchmark

Perturbation	BARTScore*	BERTScore*	BLEURT20*	COMET*	PRISM*	NegBLEURT
base_shuffled	0.44	1.7	0.46	0.88	0.54	0.05
base_unrelated_trans	1.1	2.2	0.81	0.62	0.62	1.64
critical_addition	0.032	0.12	0.065	0.076	0.043	0.18
critical_antonym	0.043	0.15	0.088	0.098	0.044	0.38
critical_codemix	0.052	0.58	0.1	0.23	0.056	0.55
critical_gender	0.023	0.067	0.1	0.093	0.031	0.02
critical_ne_removed	0.14	0.31	0.15	0.17	0.084	0.17
critical_ne_replaced	0.18	0.37	0.2	0.18	0.12	0.38
critical_negation	0.058	0.21	0.15	0.15	0.053	0.93
critical_noun_removed	0.057	0.25	0.14	0.18	0.055	0.1
critical_numbers_replaced	0.07	0.044	0.052	0.01	0.046	0.09
critical_removed_adj_adv	0.045	0.1	0.047	0.052	0.034	0.07
critical_subj_removed	0.082	0.25	0.13	0.16	0.062	0.17
critical_verb_removed	0.031	0.21	0.12	0.17	0.043	0.09

6

Conclusion

Contribution

1

CANNOT dataset

2

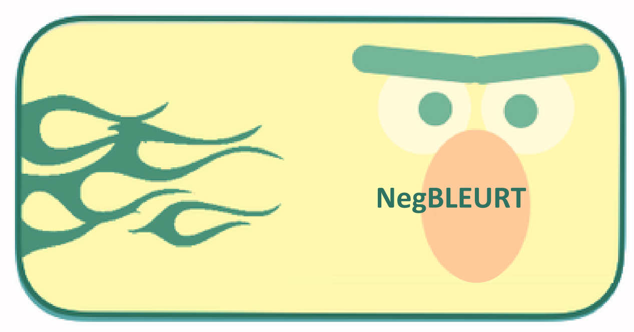
CANNOT dataset

Negate: A Python module to negate sentences

pypi v1.0.0 active issues 0

3

NegBLEURT metric



Spaces tum-nlp/negbleurt

Thank you for listening!

14th September 2023

Check out our code on GitHub!

Contact the authors!

Miriam Anschütz, Diego Miguel Lozano, Georg Groh

Technical University of Munich
School of Computation, Information and Technology
Research Group Social Computing

References

[1] Not another Negation Benchmark: The NaN-NLI Test Suite for Sub-clausal Negation (Truong et al., AACL-IJCNLP 2022)

[2] Transformers: State-of-the-Art Natural Language Processing (Wolf et al., EMNLP 2020)

[3] Automated Fact-Checking of Claims from Wikipedia (Sathe et al., LREC 2020)

[4] From group to individual labels using deep features (Kotzias et al., KDD 2015)

[5] It Is Not Easy To Detect Paraphrases: Analysing Semantic Similarity With Antonyms and Negation Using the New SemAntoNeg Benchmark (Vahtola et al., BlackboxNLP 2022)

[6] DEMETR: Diagnosing Evaluation Metrics for Translation (Karpinska et al., EMNLP 2022)