CLISTER: A Corpus for Semantic Textual Similarity in French Clinical Narratives

LREC 2022

Nicolas Hiebel, Olivier Ferret, Karën Fort, Aurélie Névéol

June 2022

Contributions

- **Definition of similarity guided by linguistic and clinical criteria**
- **CLISTER: a new STS[1] corpus of 1,000 sentence pairs**
- **Evaluation of a baseline STS model on CLISTER**
- **Comparison with an existing STS corpus in French**

[1] Semantic Textual Similarity

Semantic Textual Similarity (STS)

For clarity, examples are given in English

Semantic Textual Similarity (STS): evaluating the proximity between two pieces of text

(1) a. *There was no post-void residual urine.*
b. *There were no notable cardiovascular risk factors.*

$\Longrightarrow$ How similar are those two sentences on a scale of 0 to 5?

There is a need for diversity in STS corpora

Semantic Textual Similarity is hard to define

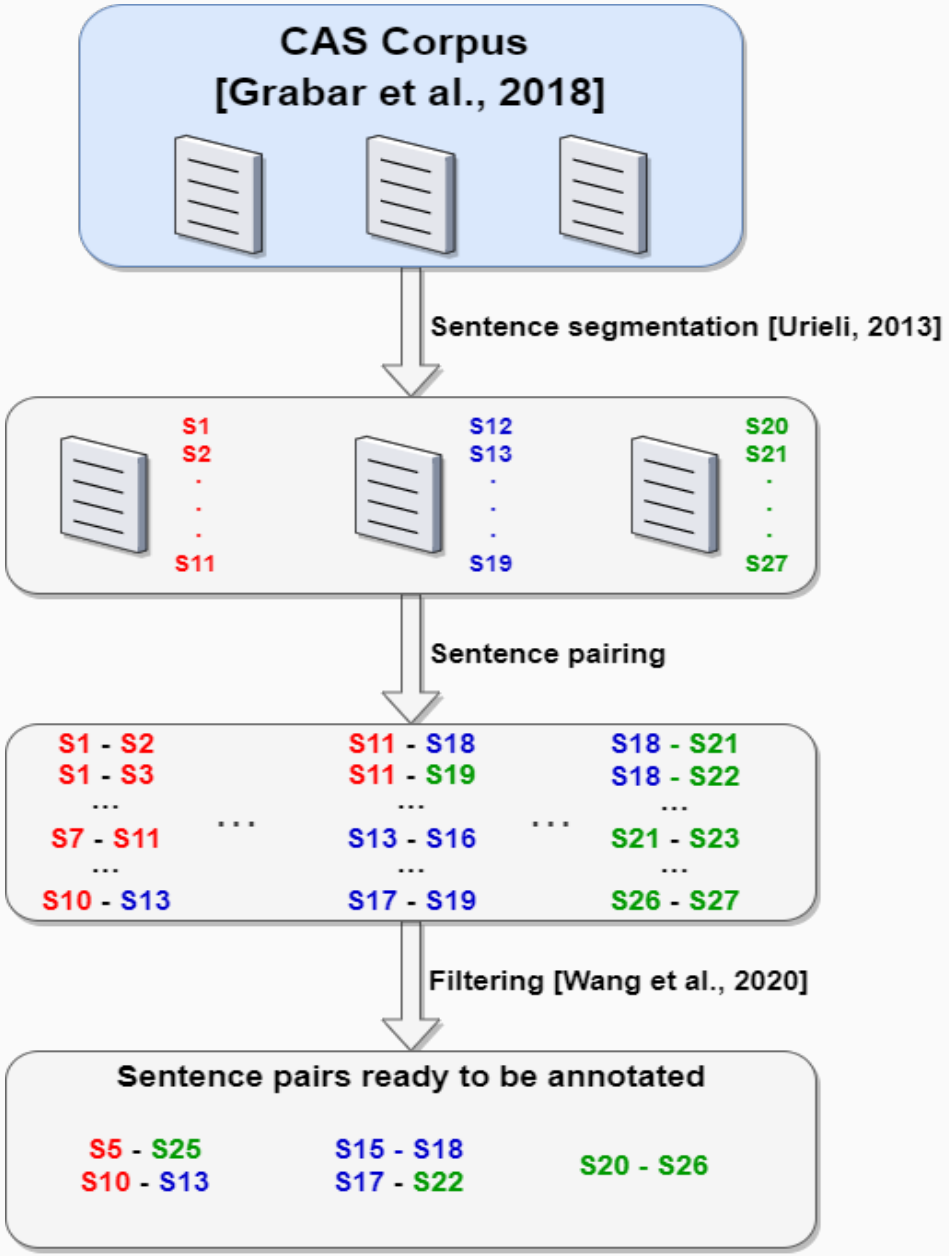
Difficulty of assessing similarity

(2) a. The patient's vitals were stable after the surgery
b. The patient's vitals were not stable after the surgery

(3) a. The patient was a 60 year old woman.
b. The patient was a 36 year old woman.

(4) a. The patient was disease free at 12-month follow-up.
b. The patient was in good health at 27-month follow-up.

Building the corpus - Sentence pair selection



Building the corpus - Annotation guidelines

- **Surface similarity**
  Concerns grammatical words or words unrelated to the domain
  There was no A / There was no B
- **Semantic similarity**
  Concerns medical concepts. The closer the concepts, the higher the similarity
  The exam A revealed... / The exam C revealed...
  $\Longrightarrow$ Two different exams. Are they medically related? Do they inspect the same body part?
- **Clinical compatibility**
  Can the sentences refer to the same clinical case?
  The scan results were negative / The scan results were positive

Building the corpus - Annotation guidelines

On a scale from 0 (completely dissimilar) to 5 (completely similar)

Similarity score 0

a. There was no post-void residual urine.

b. There were no notable cardiovascular risk factors.

$\Longrightarrow$ Only surface similarity

Similarity score 2

a. Head CT scan was negative.

b. Radionuclide scan was negative.

$\Longrightarrow$ Medical concepts with low semantic similarity

$\Longrightarrow$ Difficulty: how close are the two exams?

Building the corpus - Annotation guidelines

Similarity score 4

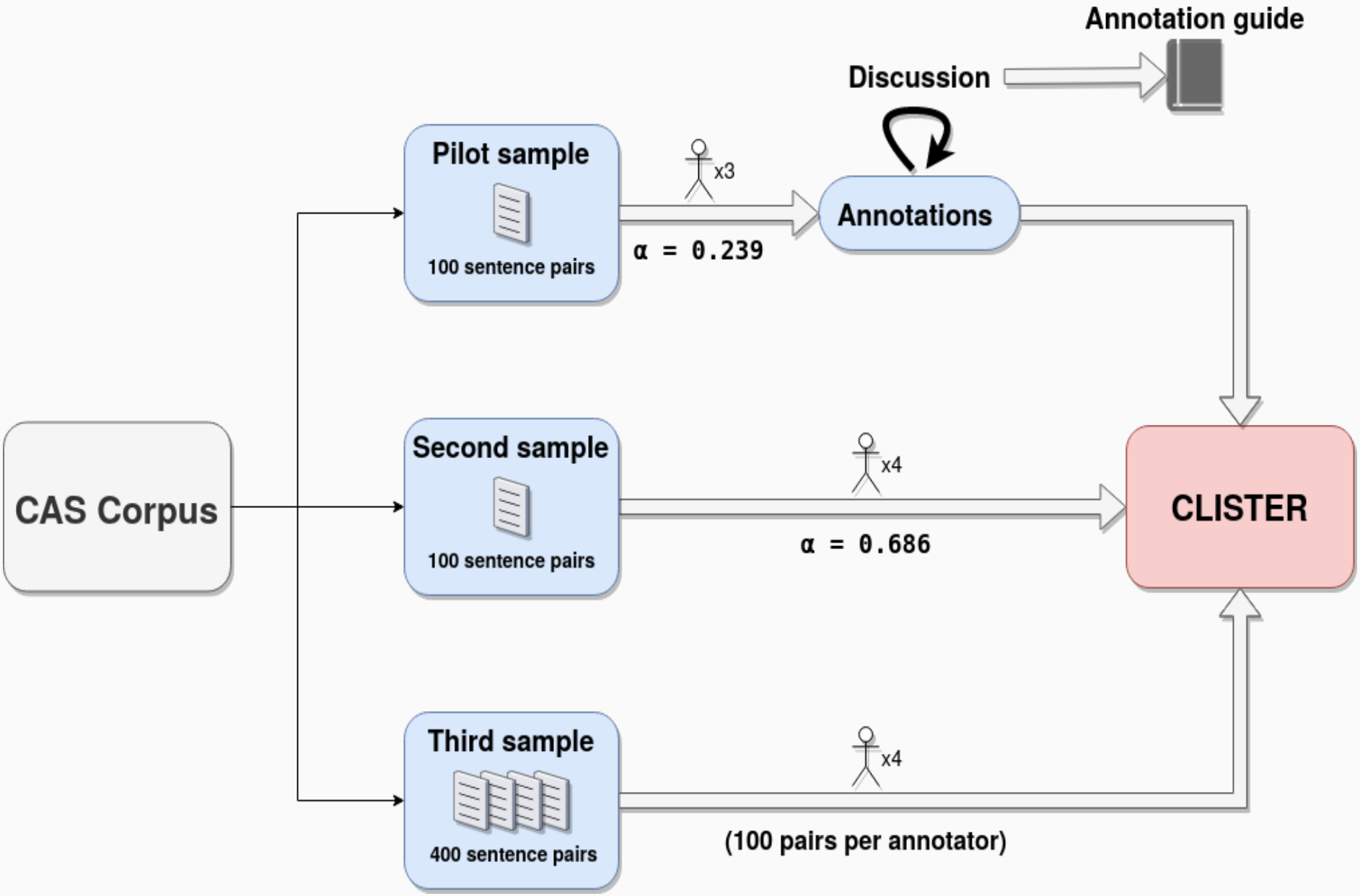
a. The patient was disease free at 12-month follow-up.
b. The patient was in good health at 27-month follow-up.

$\Longrightarrow$ High semantic similarity

$\Longrightarrow$ High clinical compatibility

$\Longrightarrow$ No contradiction between the differences

Building the corpus - Annotation steps



Building the corpus - Balancing categories

Objective: having a fair representation of the extreme scores

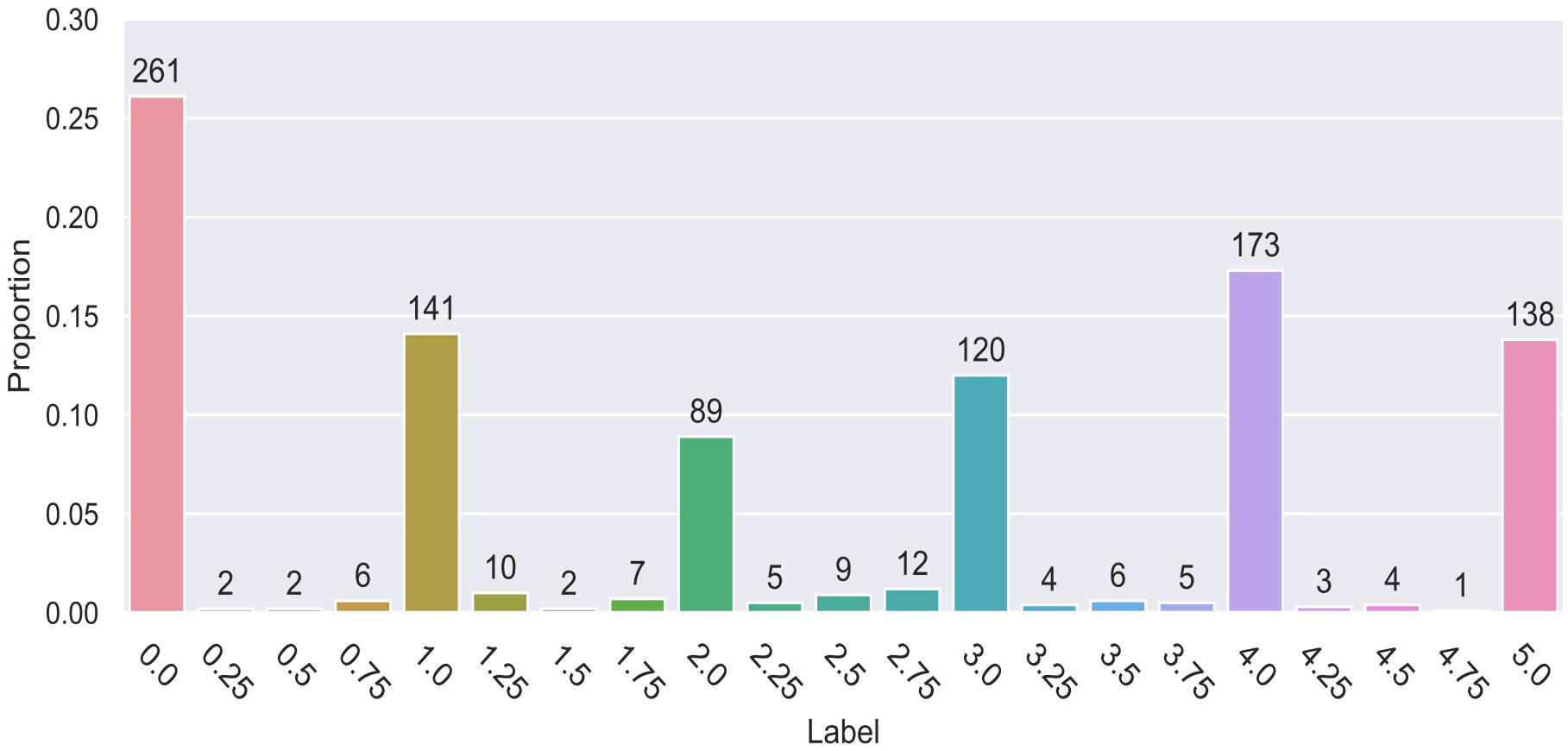
Solution: retrieve sentence pairs of similarity 0, 4 and 5.

Method: automatically retrieve candidate pairs with low / high similarity using Sentence-BERT[2] embeddings + manual verification

$\Longrightarrow$ 400 sentence pairs added

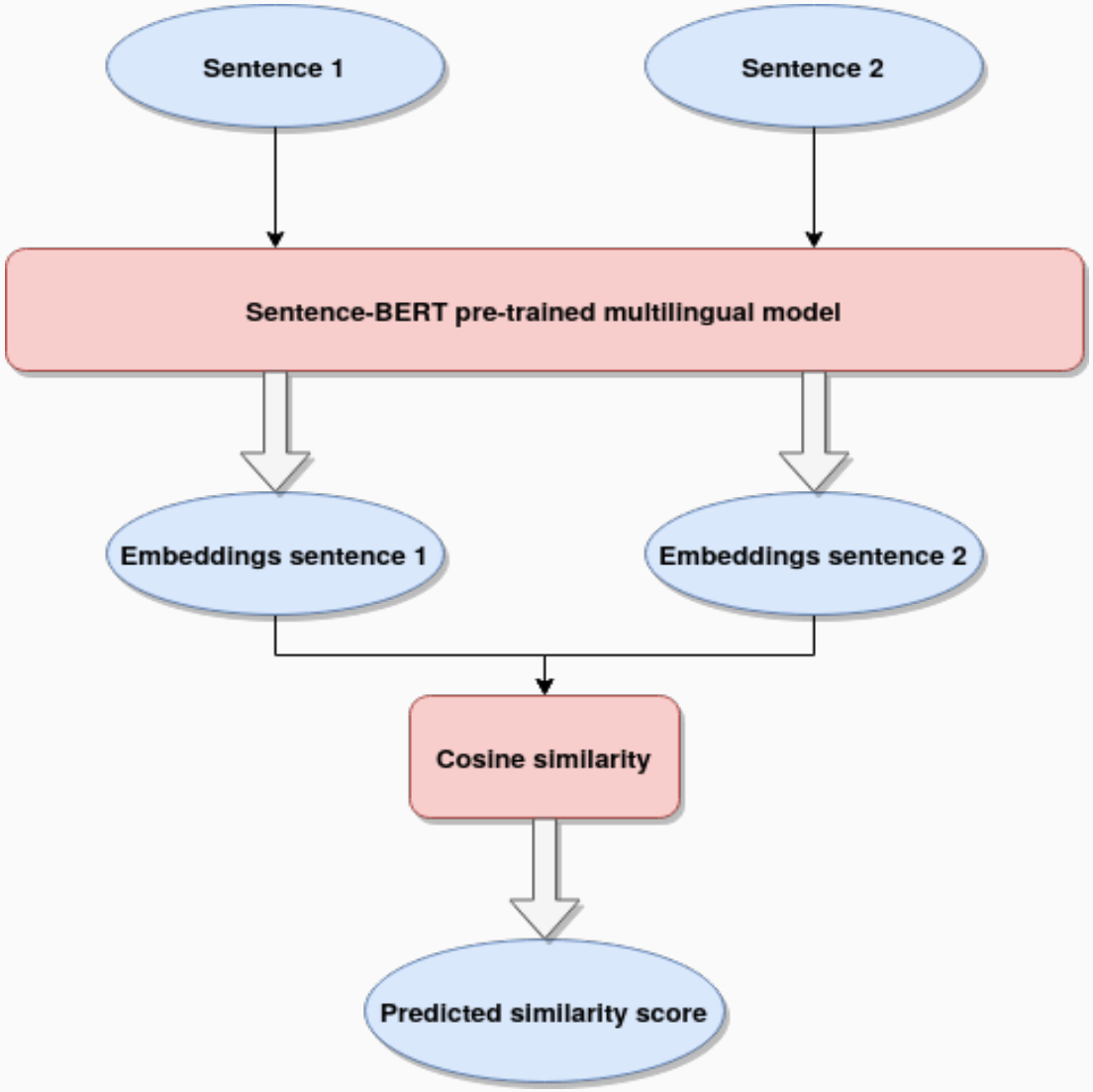
[2] [Reimers and Gurevych, 2019]

Resulting corpus



- **Sentence pairs**: 1,000
- **Total tokens**: 30,942
- **Mean sentence length**: 15.34 tokens ($\pm$9.2)

Extrinsic Evaluation of Semantic Textual Similarity



Comparison with DEFT STS

A French Corpus for Semantic Similarity[3] (DEFT STS)

$\Longrightarrow$ sentence pairs from encyclopedias, drug leaflets, medical literature reviews

Corpus	Tokens	Pairs	Mean Sentence Length
DEFT STS	53,633	1,010	26.52($\pm$12.4)
CLISTER	30,942	1,000	15.34($\pm$9.2)

Statistics on the two STS corpora.

[3] [Cardon and Grabar, 2020]

Comparison with DEFT STS[4]

Training Data	Test Data	EDRM	Spearman
None	CLISTER	0.7149	0.7547
CLISTER	CLISTER	**0.8410**	**0.8670**
DEFT STS	CLISTER	0.7084	0.7471
CLISTER + DEFT STS	CLISTER	0.8326	0.8659
None	DEFT STS	0.6505	0.7304
CLISTER	DEFT STS	0.6205	0.6906
DEFT STS	DEFT STS	**0.7926**	**0.8343**
CLISTER + DEFT STS	DEFT STS	0.7883	0.8266

[4] Train/Test repartition : 600/410 for DEFT STS, 600/400 for CLISTER

Conclusion & Perspectives

Conclusion

- New STS corpus of 1,000 sentence pairs in the clinical domain in French[5]
- A state-of-the-art sentence embedding model was able to capture our definition of similarity
- CLISTER and DEFT STS are complementary

Perspectives

- Investigate what makes the two French STS corpora different
- Use CLISTER to train a Sentence-BERT model and experiment the retrieval of similar sentences in clinical corpus

[5] https://gitlab.inria.fr/codeine/clister

Thank you !

CODEINE ANR-20-CE23-0026-01