Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4306–4315
Marseille, 20-25 June 2022

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

Nicolas Hiebel[†], Olivier Ferret[‡], Karën Fort[⋆], Aurélie Névéol[†]
[†]Université Paris Saclay, CNRS, LISN, France
[‡]Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[⋆]Sorbonne Université / Université de Lorraine, CNRS, Inria, LORIA, France
[†]firstname.lastname@lisn.upsaclay.fr, [‡]olivier.ferret@cea.fr, [⋆]karen.fort@loria.fr

Abstract

Modern Natural Language Processing relies on the availability of annotated corpora for training and evaluating models. Such resources are scarce, especially for specialized domains in languages other than English. In particular, there are very few resources for semantic similarity in the clinical domain in French. This can be useful for many biomedical natural language processing applications, including text generation. We introduce a definition of similarity that is guided by clinical facts and apply it to the development of a new French corpus of 1,000 sentence pairs manually annotated according to similarity scores. This new sentence similarity corpus is made freely available to the community. We further evaluate the corpus through experiments of automatic similarity measurement. We show that a model of sentence embeddings can capture similarity with state-of-the-art performance on the DEFT STS shared task evaluation data set (Spearman=0.8343). We also show that the CLISTER corpus is complementary to DEFT STS.

Keywords: Semantic Similarity, Corpus Development, Clinical Text, French

1. Introduction

Semantic Textual Similarity (STS) is a Natural Language Processing (NLP) task aiming at evaluating the proximity between two pieces of text. It has various applications, such as question answering or text summarization, and has its roots in earlier work about paraphrase (Dolan et al., 2004) and textual entailment (Dagan et al., 2010). It is a well-known NLP task that has been regularly studied through evaluation campaigns like SemEval since 2007 and challenges like the 2020 edition of the French challenge DEFT (*Défi Fouilles de Textes*) (Cardon et al., 2020). Work on the automatic detection of sentence similarity is supported by resources such as the STS Benchmark dataset (Cer et al., 2017) and the SICK dataset (Marelli et al., 2014), which include pairs of sentences annotated with a degree of similarity (STS Benchmark) or a relation label (SICK dataset). To our knowledge, there are no existing STS-oriented corpora in French except the French Corpus for Semantic Similarity (Cardon and Grabar, 2020) from the DEFT challenge we mentioned, and no STS corpora in the medical domain. Moreover, the notion of similarity is hard to define and existing corpora are not necessarily accompanied with guidelines providing an exact definition of the notion.

In English, several STS corpora in the medical domain exist. The BIOSSES corpus (Soğancıoğlu et al., 2017) contains 100 sentence pairs from the Text Analysis Conference track on biomedical summarization in English. Criteria for similarity are defined in this corpus but rely on the knowledge of the annotators for identifying information that can be considered as "important" or as "details". The sentence pairs are annotated with a similarity score on a scale going from 0 to 4.

The larger version of the MedSTS Corpus (Wang et al., 2020) used in the n2c2/OHNLP shared task (Wang et al., 2020) contains 2,054 medical sentence pairs in English annotated by two annotators who are experts in the medical domain. The similarity criteria in this corpus are defined following the ideas used for the BIOSSES corpus, but adding an extra degree on the similarity scale (0 to 5). The French Corpus for Semantic Similarity (Cardon and Grabar, 2020) used in the 2020 edition of DEFT contains 1,010 annotated sentence pairs taken from the CLEAR corpus (Grabar and Cardon, 2018). The definition of the similarity throughout the annotation process of this corpus relied on the intuition of the annotators. The work presented herein shows that annotating a sentence similarity corpus in a technical domain is a hard task, especially when annotators have variable levels of knowledge of the domain. Defining precise criteria to give a definition of similarity specific to the domain can significantly improve inter-annotator agreement.

The main contributions of this work are the following:

- we introduce a definition of similarity guided by linguistic and clinical criteria;
- we propose a new STS corpus, called CLISTER, composed of 1,000 sentence pairs from the clinical domain;
- we evaluate the newly created corpus for semantic similarity using a state-of-the-art sentence embedding model.
- we compare the corpus with an existing STS corpus from French encyclopedias by experimenting

with other semantic similarity tests on both corpora to test their compatibility.

2. Computing Similarity

Lithgow-Serrano et al. (2019) and Lara-Clares et al. (2021) review corpora and methods for computing sentence similarity in English. In this section, we briefly introduce elements that are most relevant to our work.
Working on similarity implies finding methods to compare the objects that are involved. For sentences, some methods can be applied directly on two strings, like the Levenshtein distance (Levenshtein, 1966).
Other methods aim at encoding the sentence into a vector of numbers and then use existing metrics that work efficiently on such vectors. A basic method to transform text sequences into vectors of numbers is to compute the TF-IDF of all the words within the sentence, using the vocabulary of the entire corpus.
In order to obtain a sentence representation that retains (some) semantic information, deep learning models can be trained to compute sentence embeddings. In our work, we used the Sentence-BERT model (Reimers and Gurevych, 2019). Based on the BERT language model (Devlin et al., 2019), the architecture of Sentence-BERT allows for adjustments of the BERT output in order to generate fixed-sized sentence embeddings.
Once the embeddings are obtained, a similarity score between the vectors can be computed using cosine similarity.

3. Building a Sentence Similarity Corpus

3.1. Source Corpus

For this work, we used the `CAS` corpus (Grabar et al., 2018), a French medical corpus containing clinical case descriptions. Clinical cases cover a variety of clinical information such as descriptions of the medical history of the patients, as well as treatments or follow-ups. These descriptions can apply to any medical disorder. The data in the corpus has been de-identified and the publication of documents is done with the written permission of patients. The corpus can be obtained from the DEFT 2020 shared task organizers[1]. The `CAS` corpus has also been linguistically and semantically annotated, but we did not use those annotations in our work. We chose to use this corpus as the basis for our work because it is one of the largest medical corpora available in French, containing data that is intended for the medical community and not simplified to help external understanding.

3.2. Selecting Sentence Pairs

The corpus was split into sentences using the Talismane parser (Urieli, 2013) for French, with default parameters. We made the hypothesis that two randomly selected sentences are unlikely to be semantically related. That means that we needed to find a way to filter sentence pairs so that the resulting corpus is not composed of a high majority of unrelated sentence pairs.
We drew our inspiration for the selection of sentence pairs from Wang et al. (2020), to which we added a filter. In this article, sentence pairs are selected using three metrics: the Ratcliff/Obershelp pattern-matching algorithm (Black, 2021) and the Levenshtein distance at character level, and the cosine similarity at token level. The mean score of the three metrics is computed and the sentence pairs with a mean score above a threshold of 0.45 are selected.
For our work, we thought it would be sufficient to keep one method at character level and one method at token level; so we only used the Levenshtein distance and the cosine similarity. We used the same threshold of 0.45 for the mean score of the two metrics to select sentence pairs.
First, we encoded our sentences according to a Bag-of-Words representation with a TF-IDF weighting scheme. We then used the cosine similarity to keep sentence pairs above a given threshold of 0.4, slightly smaller than our final threshold (0.45) in order to filter a good number of sentence pairs before the Levenshtein distance needs to be computed, which is a lot more costly than computing cosine similarity. Then, we kept only the pairs where the two sentences were close in number of tokens to exclude sentence pairs where a small sentence is included in a larger one and where the similarity of the remaining non-common part would be hard to evaluate.
We then computed the Levenshtein distance between the sentences of a pair. In the end, we kept the sentence pairs where the mean of the cosine similarity and the Levenshtein distance was above a threshold of 0.45, as was done in Wang et al. (2020).

3.3. Annotation Guidelines

3.3.1. Defining Criteria

We defined similarity by decomposing it into three linguistically and clinically driven dimensions. This helps the annotators focus on those aspects and therefore be more consistent. We divided the notion of similarity into three categories. We present here those categories, from the less important to the most important. Examples can be found in section 3.3.2.

Surface similarity The surface similarity concerns the structural similarity. This similarity is based on grammatical words or words that are not related to the domain. Two sentences that have a surface similarity can be syntactically close but semantically distant.

Semantic similarity In our corpus, semantic similarity concerns medical concepts. The closer the concepts are to one another, the higher the similarity. These concepts can refer to medications, diseases, procedures, and others.

Clinical compatibility Going further into the semantics, clinical compatibility is an assessment of whether

[1] `https://deft.limsi.fr/2020/`

sentences in a pair can refer to the same clinical case.

3.3.2. Scoring scale with examples

When annotating sentence pairs with a similarity score, it is common to use scores going from 0 for the minimum similarity score to 5 for the maximum score, as in the `STS Benchmark` (Cer et al., 2017) and `MedSTS` (Wang et al., 2020). We decided to use this scale for our corpus, using only round values.
Each score from 0 to 5 is defined as follows[2]:

Similarity score 0: For sentence pairs with only surface similarity, such as words non-specific to the medical domain or stop-words.

(1) a. Il n'y avait pas de résidu post-mictionnel. [*There was no post-void residual urine.*]
b. Il n'avait pas de facteurs de risque cardiovasculaire notable. [*There were no notable cardiovascular risk factors.*]

Both sentences in (1) have a similar structure, starting with "*There was / were no...*". However the remaining parts of the sentences are absolutely not related.

Similarity score 1: For sentence pairs with only surface similarity, concerning at most one medical entity.

(2) a. L'examen physique révélait une légère sensibilité de la fosse lombaire droite. [*Physical exam revealed mild tenderness in the right side and abdomen.*]
b. L'examen O.R.L. retrouvait une légère surdité de perception. [*HEENT exam revealed light sensorineural deafness.*]

Both sentences (2) share a common structure, with a certain type of medical exam revealing a symptom. Besides this structure, the two sentences are not related.

Similarity score 2: For sentence pairs containing medical concepts with low semantic similarity, but no clinical compatibility. Typically, sentences in a pair can concern a disease, a procedure, or a drug.

(3) a. La TDM cérébrale n'a pas révélé d'anomalie. [*Head CT scan was negative.*]
b. La scintigraphie n'a pas montré d'anomalie. [*Radionuclide scan was negative.*]

Both sentences in (3) are about a scan, which is different from one sentence to another but ends up with a common diagnosis. This example also shows the potential difficulties of annotating in a technical domain, because judging the proximity of those two scans is not necessarily obvious.

Similarity score 3: For sentence pairs with semantic similarity on several medical concepts making them partially clinically compatible.

(4) a. Devant cet aspect non spécifique d'une tumeur rétropéritonéale isolée entrainant des signes digestifs importants, une exploration chirurgicale était décidée. [*Based on the unusual aspect of a tumor located in the retroperitoneal space impacting the digestive system it was felt appropriate to take the patient to the OR for exploration.*]

b. Devant ce tableau de tumeur rénale, l'indication d'une exploration chirurgicale était posée. [*Clinical findings were consistent with a renal tumor so the patient was taken to the OR for kidney exploration.*]

Both sentences in (4) concern the presence of a tumor that leads to a surgical exploration. On those elements, the two sentences are clinically compatible, but the tumors the sentences describe are not located in the same place, which is not clinically compatible. Sentence (4-a) also contains a more precise description of symptoms.

Similarity score 4: For sentence pairs with high semantic similarity and clinical compatibility. One sentence may contain more information than the other may, and vice-versa.

(5) a. La patiente est en rémission complète avec un recul de 12 mois. [*The patient was disease free at 12-month follow-up.*]
b. L'évolution était bonne avec un recul de 27 mois. [*The patient was in good health at 27-month follow-up.*]

The sentences in (5) describe a similar situation and are clinically compatible. The main difference is the timing of the follow-up. There is also more information in sentence (5-a) with the sex of the patient (in French) and "*disease free*" is also more specific than "*in good health*". Even if sentence (5-b) does not present this information, there is no contradiction between them.

Similarity score 5: For sentence pairs with high semantic similarity and full clinical compatibility. The sentences have globally the same meaning, while one may be more specific than the other. Here, we differentiate between being more specific and containing more information.

(6) a. Les marqueurs tumoraux (CA 15.3 et ACE) étaient normaux. [*Tumor markers (CA 15.3 and CEA) were within the normal range.*]
b. Les marqueurs tumoraux sériques étaient normaux. [*Tumor marker levels were within the norm.*]

The sentences in (6) present high similarity. The only difference is the added precision on the tumor markers in sentence (6-a), but the sentences are equivalent.

[2]Examples from the corpus are presented in French, followed by a translation into English between brackets

3.4. Balancing Categories in the Corpus

Having a closer look at the annotations of the 600 sentence pairs selected initially (see Section 4), we observed that there were only a few sentence pairs annotated with extreme similarity scores of 0 and 5. However, we think it is important to have a fair representation of those extremes in order to have a reliable definition of similarity. We also aimed to increase the size of the corpus to a scale that would allow for the training and testing of statistical similarity methods. For those reasons, we decided to expand the corpus at lower cost by semi-automatically adding sentence pairs with scores of 0, 4, and 5.

Sentence pairs with a similarity score of 0 are the easiest to get, given the hypothesis that taking two random sentences in the corpus, they are very unlikely to be similar. We thus collected 210 random sentence pairs from the corpus, with the same length constraint that we used before, and annotated them with the score 0. To make sure that no similar sentence pairs are remaining, we computed the cosine similarity between the sentence embeddings obtained with SENTENCE-BERT (Reimers and Gurevych, 2019) for each pair. The highest similarity score for a pair was 0.453. From some experience working with SENTENCE-BERT, sentence pairs with a similarity score below 0.5 are already very unlikely to have any similarity besides surface similarity. We manually checked the sentences with a cosine similarity score above 0.3. About 15 pairs were reviewed, leading to the substitution of 2 pairs.

Sentence pairs with high similarity were selected using the FAISS library (Johnson et al., 2021). Sentence embeddings were computed for all sentences in the corpus, and a similarity matrix was created. We kept the highest-ranked pairs and we manually annotated pairs with a similarity score of 4 or 5, while removing the pairs that were not similar enough for the annotator's criteria. This process was done until 190 new pairs were retrieved.

4. Manual annotation of semantic textual similarity in French clinical text

The four authors of the paper annotated the corpus. We will name them here A1, A2, A3, and A4[3]. A1 has experience in the annotation of biomedical text, A2 in semantic representation, A3 is an early-stage researcher in NLP, and A4 has experience in the creation of annotated resources.

We led the first round of annotation where three annotators (A1, A2, and A3) annotated a sample of 100 sentence pairs without any discussion between them or guidelines. The goal of this first annotation step was to test the intuition of the annotators and get an idea of the difficulty of the annotation task.

Once annotated, this first sample was used as support for discussion and the conception of the annotation guidelines. The final annotation of the sentence pairs of this sample was obtained from a global consensus between the annotators. A second sample of 100 sentence pairs was then annotated. The three original annotators were joined by the fourth one (A4), who did not participate in the discussion during the first sample. Besides the additional set of annotations, this new contribution was useful to assess the quality of the annotation guidelines, which should be understandable to annotators other than the authors.

Inter-annotator agreement was computed for the second sample. It was found substantial, so each annotator was given another sample of 100 sentence pairs.

In total, 600 sentence pairs were manually annotated.

4.1. Pilot annotation

We present herein the results of the annotation of the first sample of 100 sentence pairs, which was done without discussion between annotators or guidelines. For the inter-annotator agreement, we computed Krippendorff's α (Krippendorff, 2013) using a Python library implementing the metric[4] with interval data type and default parameters. The α for this first sample was 0.239. This poor value shows a significant disagreement between the annotators. Figure 1 displays the number of sentence pairs for each score and each annotator. We can observe an important heterogeneity of the distribution for each annotator. In this sample, annotator 2 is by far the strictest of the three annotators, with more than 60 sentence pairs annotated with the minimum score 0 and an overall average score of 0.64. Annotators 1 and 3 are more balanced between scores but annotator 3 presents a much higher average score, 2.74, against 1.99 for annotator 1.

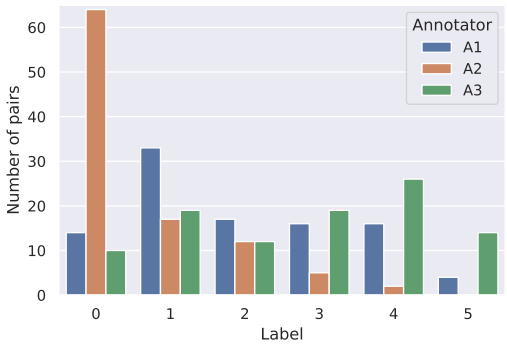

Figure 1: Distribution of annotated scores per annotator for the first sample (100 sentence pairs).

The high disparity in the first sample annotation shows that annotating clinical data with similarity based on the annotators' intuition is not trivial. Annotators can have a very distinct perception of the concept of simi-

[3]Not according to authors' order

[4]https://github.com/grrrr/krippendorff-alpha

larity. This is why during the discussion and the elaboration of the annotation guide, our goal was to define precise criteria to judge similarity in a clinical context. The annotators discussed each sentence pairs of the first sample and agreed on a consensual score.

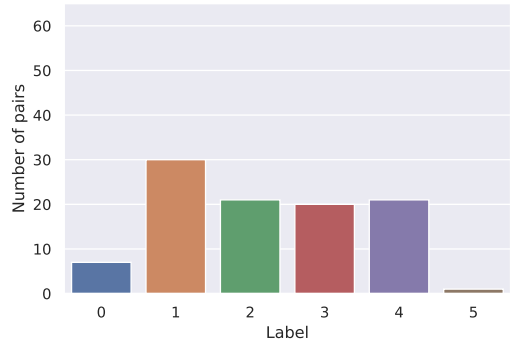

Figure 2: Distribution of annotated scores after consensus (100 sentence pairs).

The resulting distribution is displayed in Figure 2. The small number of pairs annotated with the lowest score of 0 probably means that the criteria we used to select candidate sentence pairs are often enough to achieve some similarity regarding the criteria we defined. At the same time, the fact that there is only one sentence pair annotated with the highest score of 5 probably means that such a degree of similarity between two sentences is not very common using our criteria.

4.2. Annotation consolidation

The results of the annotation of the second sample of 100 sentence pairs are shown in Figure 3. In this figure, annotators A1, A2, and A3 are the same as in sample 1, while annotator A4 is the new one.

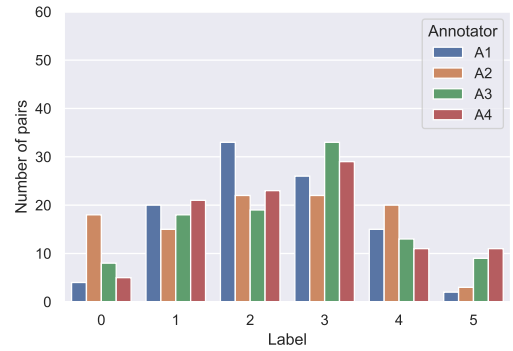

Figure 3: Distribution of annotated scores per annotator for the second sample (100 sentence pairs).

Overall, we can see a similar trend between Figure 3 and Figure 2, with a small number of sentence pairs annotated with the highest and lowest scores, 5 and 0, and most pairs in between. Only annotator A2 annotated an important number of sentence pairs with the score of 0.

On this sample, we obtained a Krippendorff's α of 0.686, which is much better than the α achieved on the first sample and seems satisfactory considering the complexity of the corpus[5].

For the final annotation of this sample, we decided to compute the mean similarity value for each pair. In the end, the mean score for this sample is 2.40 ($\pm$1.17). This mean score and the relatively small standard deviation show that most sentence pairs are annotated with scores between 1 and 4, therefore that the filters we used when selecting sentence pairs often ensure at least a small similarity in our criteria and that highly similar sentence pairs are rare.

We considered that the results obtained on the second sample were good enough to let the annotators independently annotate a new sample of 100 each. Table 1 shows the mean scores for the four new samples, which were annotated by a (different) single annotator. We can see that annotator A2 obtained a mean score very close to the mean score obtained on the second sample. The mean scores for annotators A1 and A4 are not very far, with a difference of approximately 0.40. Lastly, annotator A3 obtained a slightly lower score of 1.89. In general, mean scores are not very far from the tendency we observed in the previous sample.

Annotator	Mean score
A1	2.0 ($\pm$1.42)
A2	2.46 ($\pm$1.47)
A3	1.89 ($\pm$1.40)
A4	2.87 ($\pm$1.36)

Table 1: Mean scores obtained by annotators on their respective sample.

4.3. Global Statistics

The final corpus of 1,000 annotated sentence pairs, with an average length of 15.34 tokens ($\pm$9.2) per sentence, contains 30,942 tokens, including punctuation marks. Those numbers were computed from the TALISMANE parser output. Figure 4 shows the number of sentence pairs for each label in the corpus. With the semi-automatic selection of sentence pairs representing high/low similarity, the highest and lowest scores are now prevalent.

5. Extrinsic Evaluation of Semantic Textual Similarity

5.1. Evaluation Settings

In order to extrinsically evaluate the similarity defined in our corpus, we decided to test whether a model

[5]The semantics of inter-annotator agreement metrics is yet to be defined precisely (Mathet et al., 2012).

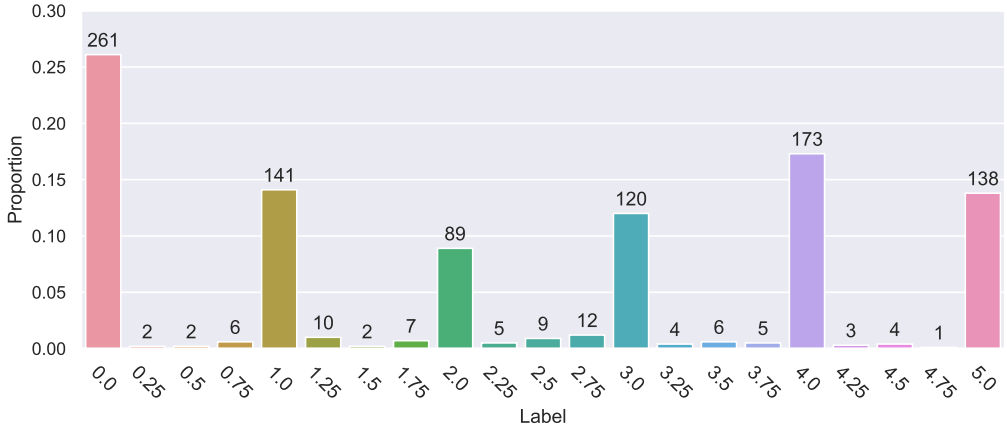

Figure 4: Proportion and number of sentence pairs for each similarity score in the final corpus (1,000 pairs).

trained on our data can account for our definition of similarity. For this experience, we used the sentence embedding model SENTENCE-BERT to build the representation of each sentence and predict the similarity between each pair of sentences using cosine similarity between the two resulting embeddings.

To perform this evaluation, we split our 1,000 sentence pairs into training and testing sets. We chose 600 sentence pairs for training and 400 sentence pairs for testing. Since we added a substantial amount of pairs semi-automatically, we made sure to have our full manually validated first sample in the test data. This ensures that the test set comprises the most reliable gold standard obtained through consensus annotations.

The rest of the data was split randomly while controlling that label distribution was consistent in the train and test partitions. As mentioned above, experiments were conducted using SENTENCE-BERT, a pre-trained multilingual model[6] covering 15 languages[7], including French (Reimers and Gurevych, 2020).

We used the basic training function of SENTENCE-BERT[8] by varying the number of epochs and the number of warm-up steps. We used default values for the rest of the parameters. Overall, the best results of the model were achieved with 5 epochs and 10 warm-up steps. The results we present in this paper are obtained with those parameters when fine-tuning SENTENCE-BERT. In order to get robust results, all experiments were repeated three times with shuffled training data. Therefore, we present here the mean values achieved with each configuration.

Two metrics were used for evaluation: Spearman correlation and EDRM (*Exactitude en Distance Relative à la Solution Moyenne* or "Accuracy as Relative Distance to Mean Solution"). EDRM is a metric created for the evaluation of the DEFT task for nuanced assessment of sentence similarity. This metric measures the mean distance between the true similarity scores and the predicted scores, taking into account the maximum distance possible from the true scores for each prediction. For example, if the true score is 0, the maximum distance is by predicting the score 5. The maximum distance here is 5 (5 - 0 = 5). If the true score is 2, the maximum distance is also achieved by predicting the score 0. However, the maximum distance here will be 3 (5 - 2 = 3). The formula for computing EDRM on a set of n sentence pairs, where h is the prediction and r is the reference score, is shown in equation 1.

$$EDRM = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{d(h_i, r_i)}{dmax(h_i, r_i)}\right) \quad (1)$$

Spearman's correlation is one of the standard metrics used to evaluate STS tasks. Both measures give a coefficient describing how much the variations of two variables are related, the variables here being the expected similarity score and the predicted similarity score.

5.2. Automatic semantic text similarity measurement in CLISTER

Table 2 presents the results of evaluating the similarity on our entire test set. When the model is used without fine-tuning, it achieves acceptable scores for EDRM (0.7149) and Spearman's correlation (0.7547). We observe a significant improvement for both EDRM and Spearman's correlation when the model is fine-tuned on the training data (+0.1261 for EDRM and +0.1123 for Spearman's correlation). It is encouraging to see that the model could adapt to an extent to our definition of similarity.

However, the use of SENTENCE-BERT in the process of selecting high-scoring sentence pairs added to the

[6] https://www.sbert.net/docs/pretrained_models.html#multi-lingual-models

[7] *distiluse-base-multilingual-cased-v1*

[8] https://www.sbert.net/docs/training/overview.html#sentence_transformers.SentenceTransformer.fit

Training Data	Test Data	EDRM	Spearman
None	$CLISTER_{100}$	0.6323	0.3794
CLISTER	$CLISTER_{100}$	0.8240	0.7340
None	$CLISTER_{400}$	0.7149	0.7547
CLISTER	$CLISTER_{400}$	**0.8410**	**0.8670**
DEFT STS	$CLISTER_{400}$	0.7084	0.7471
CLISTER + DEFT STS	$CLISTER_{400}$	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
None	$CLISTER_{400}$ + DEFT STS	0.6823	0.7449
CLISTER	$CLISTER_{400}$ + DEFT STS	0.7307	0.7474
DEFT STS	$CLISTER_{400}$ + DEFT STS	0.7519	0.8032
CLISTER + DEFT STS	$CLISTER_{400}$ + DEFT STS	**0.8123**	**0.8449**

Table 2: Results for similarity experiments with different combinations of training and testing sets using SENTENCE-BERT pre-trained multilingual model (*distiluse-base-multilingual-cased-v1*). Training was done with 5 epochs and 10 warm-up steps.

corpus could introduce a bias in the evaluation as the model's predictions will align with the reference for these sentence pairs.

We also experimented using a bias-free subset of the test set, namely the sample of sentences that were independently annotated by three annotators followed by the creation of consensus annotations ($CLISTER_{100}$). The results of the experiment are shown in Table 2.

The value for EDRM is significantly lower on this part of the testing data than on the whole test when the model is not fine-tuned (0.6323 against 0.7149), and Spearman's correlation drops drastically (0.3794 against 0.7547). When fine-tuning the model, the value for EDRM increases clearly ($+0.1917$) and is close to the value achieved on the whole test set. Spearman's correlation nearly doubles ($+0.3546$), but remains inferior to the value obtained on the complete test set. Those important improvements between the vanilla model and the fine-tuned model on this particular part of the test set show that the model was not necessarily familiar with our conception of similarity, but was able to learn from the training data to adapt to an extent, with a quite small amount of training data.

5.3. Automatic semantic text similarity measurement in CLISTER vs. DEFT STS

Cross corpus evaluations. We conducted cross corpus evaluations to assess the benefit of new data on STS performance.

To compare our corpus with the existing French STS corpus from the DEFT shared task, we experimented with the sentence embedding model SENTENCE-BERT on both datasets, crossing the training and testing data.

Table 2 presents the results of those experiments with different combinations of fine-tuning and testing sets using the SENTENCE-BERT pre-trained multilingual model (*distiluse-base-multilingual-cased-v1*).

Both corpora being similar in terms of number of sentence pairs (1,010 for DEFT STS, 1,000 for CLISTER) and of distribution across train set and test set (600/410 for DEFT STS, 600/400 for CLISTER), the results can be compared.

We kept the same parameters for SENTENCE-BERT as the ones we used in the experiments presented in Section 5. Table 2 presents the results of our experiments.

CLISTER and DEFT STS are complementary. When evaluating on the combination of the CLISTER test set, the results without fine-tuning the model and fine-tuning on the CLISTER training data are shown in Table 2. When fine-tuning on the DEFT STS training data, the values of EDRM and Spearman's correlation are slightly lower than the values achieved without training (-0.0065 for EDRM and -0.0076 for Spearman's correlation). This is the first indication that the similarity and/or the data type of DEFT STS are not necessarily compatible with CLISTER's similarity criteria and/or data type. When fine-tuning on the train sets of both corpora, the values of EDRM and Spearman's correlation are just slightly lower than the values achieved when fine-tuning only on CLISTER train set (-0.0084 for EDRM and -0.0011 for Spearman's correlation).

When evaluated on the combination of the DEFT STS test set, the model without fine-tuning achieved an EDRM of 0.6505 and a Spearman's correlation of 0.7304. Its performance decreases when it is fine-tuned on CLISTER train data, here again contributing to the

intuition that the corpora are quite different. When fine-tuning on the DEFT STS training data, the model achieves an EDRM of 0.7926 and a Spearman's correlation of 0.8343. For reference, during the shared task, the best results were 0.8217 for EDRM and 0.7769 for Spearman's correlation. Lastly, when fine-tuning on both training data, the results are slightly lower than the results when fine-tuning only on the DEFT STS test set (-0.0043 for EDRM and -0.0077 for Spearman's correlation).

When evaluating on the combination of the CLISTER and DEFT STS test sets, the results were logically in between the results obtained on each individual test set. Without fine-tuning, the values for EDRM and Spearman's correlation were lower than the values achieved when testing only on the CLISTER test set (-0.0326 for EDRM and -0.0098 for Spearman's correlation), and higher than the values achieved only on DEFT STS test set (-0.0318 for EDRM and -0.0145 for Spearman's correlation). For both cases, fine-tuning on only one of the two train sets and combining the two datasets' test sets decreases the achieved values. Training on CLISTER achieves an EDRM of 0.7307 and a Spearman's correlation of 0.7474, thus 0.1103 lower for EDRM and 0.1196 lower for Spearman's correlation than with testing on CLISTER only. Meanwhile, training on DEFT STS achieves an EDRM of 0.7519 and a Spearman's correlation of 0.8032, this time 0.0407 lower for EDRM and 0.0311 lower for Spearman's correlation than with testing on DEFT STS only. Once again, the two corpora seem to be complementary to one another. The overall best performance on this combination of test sets is achieved with training on both train sets, which reaches 0.8123 for EDRM and 0.8449 for Spearman's correlation.

We can also see in Table 2 that the DEFT STS test set gets lower results on average for the metrics we used.

Several elements of our experiments show that the DEFT STS corpus and the CLISTER corpus we created are quite different. SENTENCE-BERT performed globally better for the STS task on the CLISTER corpus than it does on the DEFT STS corpus. Moreover, fine-tuning on one corpus while testing on the other decreases the performance of SENTENCE-BERT, and fine-tuning on both corpora gets slightly lower results than fine-tuning only on the corpus corresponding to the test set.

The difference in performance and this non-compatibility between the two corpora can be related to the nature of the data (clinical for CLISTER, encyclopedic for DEFT STS) and/or to the definition of similarity underlying the similarity scores within the corpora.

5.4. Comparison to DEFT's results

The results of our experiments using DEFT STS for training and testing can be directly compared to those of systems submitted to the DEFT. The best performance in the task was obtained by a method representing sentence pairs by similarity features leveraging a wide range of similarity scores and training an ensemble classifier on this feature representation (Dramé et al., 2020). The method used in our experiments offers individual representation for sentences and yields a higher Spearman correlation (0.8343 vs. 0.7769) but lower EDRM (0.8217 vs. 0.7926).

6. Conclusion

We introduce in this paper a new STS corpus of 1,000 sentence pairs in the clinical domain in French. The annotation of the corpus followed a meticulous procedure with a good number of annotators (3 to 4). This procedure, combining discussions between annotators and the elaboration of a precise annotation guide, ensured the consistency of the annotation, with a correct inter-annotator agreement given the complexity of the corpus. We confirmed the quality of this corpus by experimenting with the assessment of sentence similarity using Sentence-BERT, a state-of-the-art sentence embedding model. We showed that a model trained on our data can capture our definition of similarity.

We compared our corpus with the DEFT STS corpus, another STS corpus in French with nearly the same amount of annotated sentence pairs (1,010). We used once again Sentence-BERT to assess sentence similarity on both corpora, separately and jointly, varying the data used to train the model. We showed that the model would not adapt to the similarity of a corpus when trained on the other, and training on both corpora does not improve the results. We concluded that the two corpora are complementary, but we have yet to determine if this complementarity comes from the different definitions of similarity (intuition for DEFT STS, criteria for CLISTER), from the type of data (encyclopedia for DEFT STS, clinical cases for CLISTER), or both.

The CLISTER corpus is freely available to the research community at https://gitlab.inria.fr/codeine/clister.

The availability of an STS corpus in the clinical domain can help build models for information retrieval in the domain. In the future, we plan to use the CLISTER corpus in order to train a Sentence-BERT model and experiment with the retrieval of similar sentences in clinical corpus to assess the privacy of specific clinical sentences with respect to individual patients. It would also be interesting to further investigate the source of the differences between the DEFT STS and the CLISTER corpora.

7. Acknowledgements

This work was supported by the French National Agency for Research under grant CODEINE (artificial text COrpus DEsIgNed Ethically) ANR-20-CE23-0026-01.

8. Bibliographical References

Black, P. E. (2021). Ratcliff/obershelp pattern recognition. *Dictionary of Algorithms and Data Structures*.

Cardon, R., Grabar, N., Grouin, C., and Hamon, T. (2020). Présentation de la campagne d'évaluation deft 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In *Actes de l'atelier Défi Fouille de Textes@JEP-TALN 2020 similarité sémantique et extraction d'information fine. Atelier DÉfi Fouille de Textes*, pages 1–13, Nancy, France. Association pour le Traitement Automatique des Langues.

Dagan, I., Dolan, B., Magnini, B., and Roth, D. (2010). Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 16(1):105—-105.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356, Geneva, Switzerland. COLING.

Dramé, K., Sambe, G., Diop, I., and Faty, L. (2020). Approche supervisée de calcul de similarité sémantique entre paires de phrases. In *Actes de l'atelier Défi Fouille de Textes@JEP-TALN 2020 similarité sémantique et extraction d'information fine. Atelier DÉfi Fouille de Textes*, pages 49–54, Nancy, France, 6. Association pour le Traitement Automatique des Langues.

Johnson, J., Douze, M., and Jégou, H. (2021). Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547.

Krippendorff, K. (2013). *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, 3rd edition edition.

Lara-Clares, A., Lastra-Díaz, J. J., and Garcia-Serrano, A. (2021). Protocol for a reproducible experimental survey on biomedical sentence similarity. *Plos one*, 16(3):e0248663.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, February.

Lithgow-Serrano, O., Gama-Castro, S., Ishida-Gutiérrez, C., Mejía-Almonte, C., Tierrafría, V. H., Martínez-Luna, S., Santos-Zavaleta, A., Velázquez-Ramírez, D., and Collado-Vides, J. (2019). Similarity corpus on microbial transcriptional regulation. *Journal of biomedical semantics*, 10(1):1–14.

Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S., and Zweigenbaum, P. (2012). Manual Corpus Annotation: Giving Meaning to the Evaluation Metrics. In *International Conference on Computational Linguistics*, pages 809–818, Mumbaï, India, December.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, et al., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online, November. Association for Computational Linguistics.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse II le Mirail.

Wang, Y., Fu, S., Shen, F., Henry, S., Uzuner, O., and Liu, H. (2020). The 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity: Overview. *JMIR medical informatics*, 8(11), November.

9. Language Resource References

Cardon, R. and Grabar, N. (2020). A French corpus for semantic similarity. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6889–6894, Marseille, France, May. European Language Resources Association.

Cer, D., Diab, M., Agirre, E. E., Lopez-Gazpio, I., and Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation. In *The 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 1 – 14, Vancouver, Canada, August.

Grabar, N. and Cardon, R. (2018). CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands, November. Association for Computational Linguistics.

Grabar, N., Claveau, V., and Dalloux, C. (2018). CAS: French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, Brussels, Belgium, October. Association for Computational Linguistics.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). The

SICK (Sentences Involving Compositional Knowledge) dataset for relatedness and entailment, May.

Soğancıoğlu, G., Öztürk, H., and Özgür, A. (2017). BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58, 07.

Wang, Y., Afzal, N., Fu, S., Wang, L., Shen, F., Rastegar-Mojarad, M., and Liu, H. (2020). Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54, 03.