Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6889–6894
Marseille, 11–16 May 2020

A French Corpus for Semantic Similarity

Rémi Cardon, Natalia Grabar
UMR 8163 STL, CNRS, Université de Lille
Domaine du Pont de bois
59653 Villeneuve d'Ascq CEDEX, France
{remi.cardon, natalia.grabar}@univ-lille.fr

Abstract

Semantic similarity is an area of Natural Language Processing that is useful for several downstream applications, such as machine translation, natural language generation, information retrieval, or question answering. The task consists in assessing the extent to which two sentences express or do not express the same meaning. To do so, corpora with graded pairs of sentences are required. The grade is positioned on a given scale, usually going from 0 (completely unrelated) to 5 (equivalent semantics). In this work, we introduce such a corpus for French, the first that we know of. It is comprised of 1,010 sentence pairs with grades from five annotators. We describe the annotation process, analyse these data, and perform a few experiments for the automatic grading of semantic similarity.

Keywords: semantic similarity, manual annotation, French language, regression

1. Introduction

Semantic textual similarity is a subtask of Natural Language Processing. At the level of sentences, the task consists in evaluating to what extent two sentences express the same meaning. This task is useful for several applications, such as machine translation, text summarization, information retrieval, natural language generation, or text simplification (Wieting et al., 2019; Vadapalli et al., 2017; Yasui et al., 2019; Kajiwara and Komachi, 2016). The computing of the semantic textual similarity requires corpora with annotated pairs of sentences. The annotation is most of the time performed on a continuous scale where scores range from 0 (the sentences express completely unrelated meanings) to 5 (the meaning is exactly the same in both sentences). Several challenges dedicated to semantic textual similarity (STS) have been held within the SemEval evaluation campaign between 2012 and 2017. STS provides the research community with bilingual and monolingual data. In our work, we are interested in monolingual semantic similarity. In relation with the monolingual semantic similarity, data from a few languages (English, Spanish and Arabic) have been exploited (Cer et al., 2017) and made available for the research community. The overall STS benchmark data for English[1], with data taken from editions held from 2012 to 2017, contains 8,628 sentence pairs, while only 250 sentence pairs were proposed for Spanish and for Arabic. Besides, similar data are also proposed for Portuguese through the ASSIN workshop (Feitosa and Pinheiro, 2017) dataset, which is composed of 10,000 pairs – 5,000 for Brazilian Portuguese and 5,000 for European Portuguese. All those datasets are taken from general language and various sources : news articles, forum posts and video subtitles. Yet, there is no similar data in French.

In our work, we introduce a semantic textual similarity corpus for French. We first describe the data that have been used and the annotation process, then we present the resulting resource. We also describe an experiment that shows an attempt at reproducing the annotation automatically.

2. Corpus and Annotation Process

In this section, we first present the data provided to the annotators. We then describe the annotation process and analyse the annotation criteria defined by the annotators.

2.1. Data Processed

The same batch with 1,010 sentence pairs was provided to five annotators. The sentence pairs are issued from a general language corpus containing sentences extracted from Wikipedia [2] and Vikidia [3] articles, and from texts related to the medical field. In this last case, the sentences are extracted from the CLEAR corpus (Grabar and Cardon, 2018), which includes information about drugs, medical literature reviews, and medicine-related articles from Wikipedia and Vikidia. The purpose of this corpus is to propose comparable contents which are distinguished by their technicality: technical and difficult to understand texts are paired with the corresponding simple or simplified texts. This is another factor that distinguishes our dataset from the existing datasets in other languages mentioned in section 1.. The candidate pairs of sentences were generated automatically while building a classification method(Cardon and Grabar, 2019) and then validated and selected manually. That method is similar to the one described in section 4.1.. The main difference is that it is based on the Random Forest classifier algorithm, whereas below we use it as a Regressor. In the work presented in this paper, the goal is to retain sentence pairs pertaining to various degrees of similarity in order to be able to train a model to assign values on a continuous scale instead of binary values (aligned or not aligned).

Hence, the semantic similarity between sentences within a given pair is due to their technicality and to the complexity of their contents, which can be lexical, syntactic or semantic. Here is an example from the CLEAR corpus, with an English translation :

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

[2] `https://fr.wikipedia.org/`

[3] `https://fr.vikidia.org`

	A1	A2	A3	A4	A5
0.5		A few identical segments			
1	Same topic, loose relation	One summarizes the other	Little shared information	Inference can be drawn	Almost unrelated meaning
1.5		Incomplete main information on one side and extra information missing			
2	Same topic, different information	Incomplete main information on one side	Same function, little shared information	Intermediate level	Same subject, different information
2.5		Same meaning, radically different expression			
3	Same topic, loosely shared information	Same meaning, different expression	Extra information on one side	Main concept of one sentence is missing in the other one	Extra information on one side
3.5		Same meaning, paraphrases are found			
4	Almost same content, additional information on one side	Same meaning, slight rephrasing	Same function and almost same information	Additional information on one side	One slight difference in the delivered information
4.5		Same meaning, slight syntactic difference			

Table 1: Annotation criteria defined by the annotators

1. *Les effets graves intéressant les systèmes hépatique et/ou dermatologique ainsi que les réactions d'hypersensibilité imposent l'arrêt du traitement.* (*Severe effects affecting the liver and/or dermatological systems and hypersensitivity reactions require discontinuation of treatment.*)

2. *Le traitement doit être arrêté en cas de réaction allergique généralisée, éruption cutanée ou altérations de la fonction du foie.* (*Treatment should be discontinued in the event of a generalized allergic reaction, rash or impaired liver function.*)

2.2. Annotation Process

The five annotators involved have received higher education: two of them are trained in Natural Language Processing, one is a medical practitioner. Except one, all annotators are native French speakers. The authors were not part of the annotators. The annotation guidelines provided to the annotators were very simple and short:

- to assign a score of 0 when the sentences are completely unrelated,
- to assign a score of 5 when the sentences mean the same,
- to come up with their own scale and criteria for the intermediate values,
- to define a short description of the annotation criteria.

We prefered not to bias the manual annotations with some *a priori* criteria, such as

1. use the score n for sentence pairs with syntactic modifications,
2. use the score m for sentence pairs with lexical modifications, etc.

Indeed, our motivation was to exploit the linguistic competence of the annotators and to compare their semantic sensitivity and judgements. We assume also that, in this way, the overall semantic scores should better represent the semantic similarity between the sentences.
The annotators estimated that the annotation of the 1,010 pairs of sentences took between seven and fifteen hours.

2.3. Scales and Annotation Criteria according to the Annotators

The scales and criteria that were used by the annotators can be seen in Table 1. We can observe differences and similarities between the various annotation principles provided by the annotators:

- Except one, all the annotators assigned integer scores [0, 1, 2, 3, 4, 5] to the pairs of sentences. One annotator also used intermediary scores [0.5, 1.5, 2.5, 3.5, 4.5];

- The A3 annotator considered that he took the sentences strictly as they were given, which means that the unknown context was considered as non-existent. That implies for example that pronouns in one sentence were never assumed to be referring to an element explicitly mentioned in the other sentence, increasing the likelihood of dissimilarity;

- The scales from A2 and A3 are much more conservative than the other three. Yet they greatly differ from one another. A2 is the only annotator who focused on phrasing. In order to assign the highest score according to their scale, the two sentences have to be identical. The scale given by A3 is more similar to the other ones but it is conservative because of the strict view related to context not being assumed;

- For specific grades, 2, 3 and 4 are quite similar for all the annotators but A2: 2 involves that the sentences have something that differentiates them, but they deal with the same subject. 3 implies each time that there is shared information but that one sentence expresses information that is not found in the other one, and 4 implies that the information is ”almost” or ”slightly” the same;

- It is more difficult to analyse the relationship between the descriptions for grade 1. A1 and A4 both mention something in common, the domain, or grounds for inference, but also state that nothing more can reinforce the link between the two. A3 and A5 focus on the lack of relation between the sentences.

To summarize, we can see that the annotators paid attention to several criteria when deciding about the semantic relatedness of the sentences:

- intersection of the meaning, such as missing information, incomplete information or extra words on either side,

- use of paraphrases and different expressions,

- possibility to do textual inference.

We observe also that the completeness of information is the most frequently used criteria by all the annotators.

2.4. Global scores

Using all the scores from the five annotators, we computed two more values :

- The average score for each pair, rounded (”Avg” further down);

- The most frequent score out of the five for each pair (”Vote” further down).

3. Analysis of the Annotations

In this section, we further analyse the annotations: their breakdown by score and the correlation of the scores from the five annotators.

3.1. Breakdown by Score

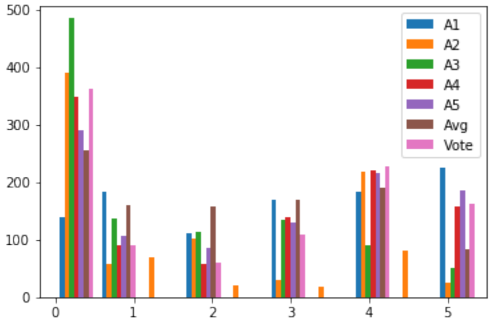

Figure 1: Breakdown by category and annotator

Figure 1 shows the breakdown by score and annotator. The x-axis shows the different scores and the y-axis shows the number of pairs. The isolated bars are due to the scale used by A2, which is the only one that included .5 values. We also indicate figures for Avg and Vote.
We can observe that the 0 score is the most used by every annotator but one (A1). The annotator A3 assigned the 0 score to almost half the pairs, which is coherent with the annotation criteria of this annotator, who did not assume context for coreference and thus had the most conservative approach.
We can also see that every annotator but one (A1) used 4 more often than 5. This can be explained by the nature of the sentence pairs. As stated in section 2.1., the source corpus is aimed towards simplification and the sentence pairs come from document pairs where one is more technical than the other one. In consequence, it can be expected that there are more almost identical sentences than entirely identical ones, as the texts are not written for the same audience and thus do not deliver the exact same information in the exact same way.
Looking at Avg and Vote, we observe that grades 3 and 4 are the most consistent overall. Grade 2 seems to be the most inconsistent, with an average that is way above the individual counts, and a vote that is low.

3.2. Correlation Coefficients

We computed the Krippendorff’s α (Krippendorff, 1970) to evaluate the global correlation coefficient of the annotations. The α value for the five annotators is 0.69. This value is above the generally observed threshold which is considered as reliable (α = 0.67). Yet, this score is quite low. When we take the average and the vote scores into account for the computation, the α value goes up to 0.77, which is a sign that putting all the annotations together significantly improves the data reliability. In order to explore those results more deeply, we computed the correlation between pairs of annotators.

	A1	A2	A3	A4	A5
A1	1.0	0.77	0.72	0.84	0.81
A2	0.77	1.0	0.64	0.75	0.74
A3	0.72	0.64	1.0	0.75	0.70
A4	0.84	0.75	0.75	1.0	0.80
A5	0.81	0.74	0.70	0.80	1.0

Table 2: Pearson's correlation coefficients between the annotators

Table 2 shows the Pearson correlation(Kirch, 2008) for every combination of two annotators. The observations we can make are consistent with figure 1 and the criteria described in section 2.3..:

- The lowest correlation coefficient (0.64) occurs between A2 and A3: A2 is the annotator who used steps of .5 in his scale and A3 relied on annotation principles that had him assign 0 to almost half the pairs. Hence, those two annotators applied annotation scales and criteria that differ the most from the other ones.
- The correlation coefficients between the other three annotators (A1, A4 and A5) are the highest: 0.84 for A1 and A4, 0.81 for A1 and A5 and 0.80 for A4 and A5.
- The other associations range between 0.70 and 0.77.

Globally, the correlation coefficients show a satisfying reliability for the dataset, with variations according to the different scales that were used. We see that the two scales that stand out have the lowest correlation coefficient with each other, but at the same time they have a good correlation coefficient with the other three. Those other three have strong coefficients with one another.

4. Experiments

In order to study how the resulting corpus can be exploited, we ran an experiment to check how accurately we could automatically reproduce the annotations. In this section, we first describe the automatic approach for scoring the pairs of sentences and then the results obtained.

4.1. Automatic Approach for Scoring the Pairs of Sentences

We exploited a previously proposed method dedicated to the detection of parallel sentences in comparable corpora (Cardon and Grabar, 2019). Yet, in order to predict values on a continuous scale, the Random Forest Regressor is exploited instead of the classifier. We compute and use several sets of features, mainly obtained from the lexical and sublexical content of the sentences, their word-based similarity, and the corpus-suggested similarity from word embeddings:

1. *Number of common non-stopwords*. This feature permits to compute the basic lexical overlap between specialized and simplified versions of sentences (Barzilay and Elhadad, 2003). It concentrates on non-lexical content of sentences;
2. *Percentage of words from one sentence included in the other sentence, computed in both directions*. This features represents possible lexical and semantic inclusion relations between the sentences;
3. *Sentence length difference between specialized and simplified sentences*. This feature assumes that simplification may imply stable association with the sentence length;
4. *Average length difference in words between specialized and simplified sentences*. This feature is similar to the previous one but takes into account average difference in sentence length;
5. *Total number of common bigrams and trigrams*. This feature is computed on character ngrams. The assumption is that, at the sub-word level, some sequences of characters may be meaningful for the alignment of sentences if they are shared by them;
6. *Word-based similarity measure exploits three scores (cosine, Dice and Jaccard)*. This feature provides a more sophisticated indication on word overlap between two sentences. Weight assigned to each word is set to 1;
7. *Character-based minimal edit distance* (Levenshtein, 1966). This is a classical computation of edit distance. It takes into account basic edit operations (insertion, deletion and substitution) at the level of characters. The cost of each operation is set to 1;
8. *Word-based minimal edit distance* (Levenshtein, 1966). This feature is computed with words as units within sentence. It takes into account the same three edit operations with the same cost set to 1. This feature permits to compute the cost of lexical transformation of one sentence into another;
9. *WAVG*. This features uses word embeddings. The word vectors of each sentence are averaged, and the similarity score is calculated by comparing the two resulting sentence vectors (Stajner et al., 2018);
10. *CWASA*. This feature is the continuous word alignment-based similarity analysis, as described in (Franco-Salvador et al., 2016).

For the last two features, we trained the embeddings on the CLEAR corpus using word2vec (Mikolov et al., 2013), and the scores are computed using the CATS tool (Stajner et al., 2018).

4.2. Results

We ran the experiment for every annotator. We also ran the experiment for Avg and Vote. We randomly split the data into 90% for training and 10% for testing. As there are small variations on each run due to random splitting, each reported score represents the average over twenty runs.
Table 3 shows the results obtained when scoring the pairs of sentences tackled as the regression task. For the annotators, the correlation coefficients range from 0.73 (A2) to

A1	A2	A3	A4	A5	Avg	Vote
0.82	0.73	0.80	0.79	0.78	0.87	0.78

Table 3: Pearson's correlation coefficient on regression experiments

0.82 (A1). This shows that the various scales can be automatically reproduced, and even if there are important difference between them, the annotations can be considered to be coherent.

The most engaging observation is that the best results (0.87) are obtained on the average scores. This may mean that the average scores and collective perception of the semantic similarity remain coherent despite the differences observed during the annotation process.

Interestingly, the result for Vote is the mean of the scores for the five annotators individually.

5. Conclusion

We introduced a corpus annotated for semantic textual similarity for French. Currently, this kind of data is indeed missing in French. The corpus is composed of 1,010 sentence pairs that come from comparable corpora aimed towards text simplification. More precisely, the original texts come from the CLEAR corpus and from Wikipedia and Vikidia articles. The corpus comes with grades manually assigned by five annotators. Together with the scores, the annotators provided the annotation scheme they adopted. We performed an analysis of the resulting data and showed that there are discrepancies in the scores that have been assigned. Those discrepancies can be explained with different annotation factors. We then used these data to automatically predict the scores of the pairs of sentences. This set of experiments shows that the scores can be quite well reproduced with automatic approaches. This indicates that the manually created data are reliable and can be used for a variety of experiments where semantic textual similarity is of interest. At the time of publication, the dataset is being used in an NLP challenge and will be made available for the research community.

6. Acknowledgements

We are grateful to our annotators for their valuable work. We would also like to thank the reviewers for their helpful comments. This work was funded by the French National Agency for Research (ANR) as part of the *CLEAR* project (*Communication*, *Literacy*, *Education*, *Accessibility*, *Readability*), ANR-17-CE19-0016-01.

7. Bibliographical References

Barzilay, R. and Elhadad, N. (2003). Sentence alignment for monolingual comparable corpora. In *EMNLP*, pages 25–32.

Cardon, R. and Grabar, N. (2019). Parallel sentence retrieval from comparable corpora for biomedical text simplification. In *Proceedings of Recent Advances in Natural Language Processing*, pages 168–177, Varna, Bulgaria, september.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August. Association for Computational Linguistics.

Feitosa, D. and Pinheiro, V. (2017). Análise de medidas de similaridade semântica na tarefa de reconhecimento de implicação textual (analysis of semantic similarity measures in the recognition of textual entailment task)[in Portuguese]. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 161–170, Uberlândia, Brazil, October. Sociedade Brasileira de Computação.

Franco-Salvador, M., Gupta, P., Rosso, P., and Banchs, R. E. (2016). Cross-language plagiarism detection over continuous-space and knowledge graph-based representations of language. *Knowledge-Based Systems*, 111:87–99.

Grabar, N. and Cardon, R. (2018). CLEAR – Simple Corpus for Medical French. In *Workshop on Automatic Text Adaption (ATA)*, pages 1–11, Tilburg, Netherlands.

Kajiwara, T. and Komachi, M. (2016). Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Wilhelm Kirch, editor, (2008). *Pearson's Correlation Coefficient*, pages 1090–1091. Springer Netherlands, Dordrecht.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics. Doklady*, 707(10).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Stajner, S., Franco-Salvador, M., Ponzetto, S. P., and Rosso, P. (2018). Cats: A tool for customised alignment of text simplification corpora. In *Proceedings of the 11th Language Resources and Evaluation Conference, LREC 2018, Miyazaki, Japan, May 7-12*.

Vadapalli, R., J Kurisinkel, L., Gupta, M., and Varma, V. (2017). SSAS: Semantic similarity for abstractive summarization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 198–203, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Wieting, J., Berg-Kirkpatrick, T., Gimpel, K., and Neubig, G. (2019). Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy, July. Association for Computational Linguistics.

Yasui, G., Tsuruoka, Y., and Nagata, M. (2019). Using semantic similarity as reward for reinforcement learning in sentence generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 400–406, Florence, Italy, July. Association for Computational Linguistics.