CONCORDIA: COmputing semaNtic sentenCes for fRench Clinical Documents sImilArity

Khadim Dramé[1,2], Gorgoumack Sambe[1,2] and Gayo Diallo[3]
[1]*Université Assane Seck de Ziguinchor, Ziguinchor, Senegal*
[2]*Laboratoire d'Informatique et d'Ingénierie pour l'Innovation, Ziguinchor, Senegal*
[3]*SISTM - INRIA, BPH INSERM 1219, Univ. Bordeaux, Bordeaux, France*

Keywords: Text Duplicatoin, Semantic Sentence Similarity, Multilayer Perceptron, French Clinical Notes.

Abstract: Detecting similar sentences or paragraphs is a key issue when dealing with texts duplication. This is particularly the case for instance in the clinical domain for identifying the same multi-occurring events. Due to lack of resources, this task is a key challenge for French clinical documents. In this paper, we introduce *CONCORDIA*, a semantic similarity computing approach between sentences within French clinical texts based on supervised machine learning algorithms. After briefly reviewing various semantic textual similarity measures reported in the literature, we describe the approach, which relies on Random Forest, Multilayer Perceptron and Linear Regression algorithms to build supervised models. These models are thereafter used to determine the degree of semantic similarity between clinical sentences. *CONCORDIA* is evaluated using the Spearman correlation and EDRM classical evaluation metrics on standard benchmarks provided in the context of the Text Mining DEFT 2020 challenge based. According to the official DEFT 2020 challenge results, the *CONCORDIA* Multilayer Perceptron based algorithm achieves the best performances compared to all the other participating systems, reaching an EDRM of 0.8217.

1 INTRODUCTION

Computing semantic similarity between different sentences is a challenging task and raises very important issues in many Natural Language Processing (NLP) applications. Semantic similarity is used in various topics including question answering, plagiarism detection, machine translation and automatic text summarization. (Cer et al., 2017; P and Shaji, 2019). Various semantic sentences similarity approaches have been proposed in the literature (Cer et al., 2017; Agirre et al., 2015; Chandrasekaran and Mago, 2021). Some commonly used approaches exploit the lexical and syntactic features of sentences; common characters, tokens/words or terms between the source and the target sentences is usually exploited. Some other approaches attempt to take into account synonymy issues and to capture semantics of sentences using external semantic resources or statistical methods (Chen et al., 2020). In recent evaluation campaigns such as SemEval, supervised learning approaches have achieved the most effective performance for measuring semantic similarity between sentences in general (Cer et al., 2017; Agirre et al., 2016) and in the clinical domain in particular (Rastegar-Mojarad et al., 2018; Soğancıoğlu et al., 2017).

However, in the French clinical domain, due to the use of domain specific language and the lack of resources, computing effectively the semantic similarity between sentences is a challenging and open research problem. Similarly to international evaluation campaigns such as SemEval (Cer et al., 2017) and BioCreative/OHNLP (Rastegar-Mojarad et al., 2018), the DEFT 2020 (DÉfi Fouille de Textes (Text Mining Challenge)) challenge, which aims to promote the development of methods and applications in NLP, addresses this issue (Cardon et al., 2020) and provides standard benchmarks (Grabar et al., 2018; Grabar and Cardon, 2018).

In the current work, we propose *CONCORDIA*, a system based on a set of supervised methods which rely on classical machine learning (ML) algorithms (Random Forest (RF), Multilayer Perceptron (MLP) and Linear Regression (LR)) to determine semantic similarity between sentence pairs within French clinical notes. Annotated clinical data sets provided by the organizers of DEFT 2020 are used for assessing the performance of the proposed methods. According

Dramé, K., Sambe, G. and Diallo, G.
CONCORDIA: COmputing semaNtic sentenCes for fRench Clinical Documents sImilArity.
DOI: 10.5220/0010687500003058
In *Proceedings of the 17th International Conference on Web Information Systems and Technologies (WEBIST 2021)*, pages 77-83
ISBN: 978-989-758-536-4; ISSN: 2184-3252

to the performance and comparison from the official DEFT 2020 results, the MLP based *CONCORDIA* method outperforms all the other participating systems in the task 1 (which constitutes in 15 systems of 5 teams), reaching an EDRM of 0.8217. In addition, according to the Spearman correlation, the *CONCORDIA* LR and MLP based method got the best performance, respectively 0.7769 and 0.7691.

The remainder of the paper is structured as follows. First, the proposed methods for measuring semantic similarity are described in Section 2. Then, the official results of these methods on standard benchmarks are reported in Section 3 and discussed in Section 4. Conclusion and future work are finally presented in Section 5.

2 METHODS

This section describes the approach followed with *CONCORDIA*. Overall, it operates as follows. Sentence pairs are first represented by a set of features. Then, machine learning algorithms are used to build models. For feature engineering, various semantic text similarity measures are explored including token-based, character-based, vector-based measures and particularly the one using word embedding. The most significative measures are then combined to support supervised methods. An overview of the proposed approach is shown in Figure 1.

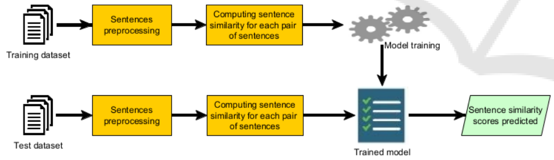
Sentences preprocessing
Computing sentence similarity for each pair of sentences
Model training
Training dataset
Sentences preprocessing
Computing sentence similarity for each pair of sentences
Sentence similarity scores predicted
Test dataset
Trained model

Figure 1: Overview of the proposed approach.

2.1 Feature Extraction

2.1.1 Token-based Similarity Measures

In this approach, each sentence is represented by a set of tokens. The degree of similarity between two sentences depends on the number of common tokens into these sentences.

The Jaccard similarity measure (Jaccard, 1912) of two sentences is the number of common tokens over the number of tokens in both sentences. Given two sentences $S1$ and S2, X and Y respectively the sets of tokens of S1 and S2, the Jaccard similarity is defined as follows:

$$sim_{Jaccard}(S1,S2) = \frac{|X \cap Y|}{|X \cup Y|} \quad (1)$$

The Dice similarity measure (Dice, 1945) of two sentences is two times the number of common tokens over the sum of cardinalities of the two sentences. Given two sentences S1 and S2, X and Y respectively the sets of tokens of S1 and S2, the Dice similarity is defined as :

$$sim_{Dice}(S1,S2) = \frac{2 \times |X \cap Y|}{|X| + |Y|} \quad (2)$$

The Ochiai similarity measure (Ochiai, 1957) of two sentences is the number of common tokens over the square root of the sum of cardinalities of the two sentences. Given two sentences S1 and S2, X and Y respectively the sets of tokens of S1 and S2, the Ochiai similarity is defined as:

$$sim_{Ochiai}(S1,S2) = \frac{|X \cap Y|}{\sqrt{|X| \times |Y|}} \quad (3)$$

The Manhattan distance measures the distance between two sentences by summing the differences of token frequencies in these sentences. Given two sentences S1 and S2, n the total number of tokens in both sentences and Xi and Yi respectively the frequencies of token i in S1 and S2, the Manhattan distance is defined as:

$$d_{Manhattan}(S1,S2) = \sum_{i=1}^{n} |X_i - Y_i| \quad (4)$$

2.1.2 Character-based Similarity Measures

The Q-gram similarity (Ukkonen, 1992) is a character-based measure widely used in approximate string matching. Each sentence is broken into substrings of length Q (Q-grams). Then, the similarity between two sentences is computed using the matches between their corresponding Q-grams.

The Levenshtein distance (Levenshtein, 1965) is an edit distance which computes the minimal number of required operations (character edits) to convert one string into another. These operations are insertions, substitutions, and deletions.

2.1.3 Vector-based Similarity Measures

The Term Frequency - Inverse Document Frequency (TD-IDF) weighting scheme (Jones, 2004) is commonly used in information retrieval and text mining for representing textual documents as vectors. In this model, each document is represented by a weighted real value vector. Then, the cosine measure is used to compute similarity between documents. Formally, let

$C = \{d_1, d_2, \ldots, d_n\}$, a collection of n documents, $T = \{t_1, t_2, \ldots, t_m\}$, the set of terms appearing in the documents of the collection and the documents d_i and d_j being represented respectively by the weighted vectors $d_i = (w_1^i, w_2^i, \ldots, w_m^i)$ and $d_j = (w_1^j, w_2^j, \ldots, w_m^j)$, their cosine similarity is defined as:

$$Sim_{COS}(d_i, d_j) = \frac{\sum_{k=1}^{m} w_k^i w_k^j}{\sqrt{\sum_{k=1}^{m} (w_k^i)^2}\sqrt{\sum_{k=1}^{m} (w_k^j)^2}} \quad (5)$$

where w_k^l is the weight (TF.IDF value) of the term t_k in the document d_l. In the context of this work, the considered documents are sentences.

The word embedding (word2vec) model (Mikolov et al., 2013), on the other hand, allows to build distributed semantic vector representations of words from large unlabeled text data. It is an unsupervised and neural network-based model that requires large amount of data to construct word vectors. Two main approaches are used to training, the continuous bag of words (CBOW) and the skip gram model. The former predicts a word based on its context words while the latter predicts the context words using a word. Considering the context word, the word2vec model can effectively capture semantic relations between words. This model is extended to sentences for learning vector representations of sentences (Le and Mikolov, 2014). Like the TF.IDF scheme, the cosine measure is used to compute the sentence similarity.

Before applying token-based, vector-based and Q-gram similarity algorithms, pre-processing consisting of converting sentences into lower cases is performed. Then, the pre-processed sentences are tokenized using the regular expression tokenizers of the Natural Language Toolkit (NLTK) (Bird and Loper, 2004). Therafter, the punctuation marks (dot, comma, colon, ...) and stopwords are removed.

2.2 Proposed Models

The proposed methods rely on the similarity measures described in the previous section. For feature selection, combinations of different similarity measures (which constitute the features) are experimented. These supervised methods require a labelled training set consisting of a collection of sentence pairs with their assigned similarity scores. First, each sentence pair, which is an instance, is represented by a set of features. Then, classical machine learning algorithms are used to build the models, which are thereafter used to determine the similarity between unlabelled sentence pairs. Several machine learning algorithms are experimented but the Random Forest (RF) and the Multilayer Perceptron (MLP), which yield the best performances in the validation set, are retained. In addition, we propose a Linear Regression (LR) model taking as inputs the predicted similarity scores of both models and the average score of the different similarity measures.

3 EXPERIMENTS

In order to assess the proposed sentence similarity methods, we used benchmarks of French clinical datasets (Grabar et al., 2018; Grabar and Cardon, 2018) provided by the organizers of the DEFT 2020 challenge. The EDRM (Accuracy in relative distance to the average solution) and the Spearman correlation coefficient are used as the official evaluation metrics (Cardon et al., 2020). We additionally used the Pearson correlation and the Accuracy metrics. The Pearson correlation is commonly used in semantic text similarity evaluation, while the accuracy measure enables to determine the correctly predicted similarity scores.

3.1 Datasets

In the DEFT 2020 challenge, the organizers provided annotated clinical datasets for the different tasks (Grabar et al., 2018; Grabar and Cardon, 2018).

For the task 1, the objective is to determine sentence pairs similarity. A labeled training set of 600 sentence pairs and a test set of 410 pairs are made available. Each sentence pair is manually annotated with a value indicating their degree of similarity. The datasets are annotated independently by five human experts that assess the similarity scores between sentences ranging from 0 (completely dissimilar) to 5 (semantically equivalent). Then, scores resulting from the majority vote are used as the reference annotations. Table 1 shows examples of sentence pairs in the training set with their similarity scores. The distribution of the similarity scores in the training set is illustrated in Figure 2.

In our experiments, this training set is partitioned into two datasets: a training set of 450 and a validation set of 150 sentence pairs. The validation set was used for feature selection but also for machine learning models comparison.

3.2 Results

The *CONCORDIA* proposed approach is experimented with different combinations of similarity measures as features for building the models. For each

Table 1: Examples of annotated sentence pairs.

Sentence 1	Sentence 2	Score
La plupart des biberons d'étain sont de type balustre à tétine vissée sur pied (ou piédouche).	On a ensuite fait des biberons en étain et en fer blanc.	0
La proportion de résidents ayant des prothèses dentaires allait de 62 % à 87%.	Dans toutes les études, la plupart des participants avaient des dentiers (entre 62 % et 87 %).	1
Les essais contrôlés randomisés, les essais cas-témoins et les études de cohorte comprenant des enfants et des adultes soumis à n'importe quelle intervention pour l'hématome aigu de l'oreille.	Nous avons recherché des essais portant sur des adultes ou des enfants ayant subi un hématome.	2
Les agents de déplétion du fibrinogène réduisent le fibrinogène présent dans le plasma sanguin, la viscosité du sang et améliorent donc le flux sanguin.	Ils réduisent également l'épaisseur du sang (ou la viscosité), ce qui permet d'améliorer le flux sanguin jusqu'au cerveau.	3
Refermez le flacon immédiatement après utilisation.	Refermez l'embout du flacon avec le bouchon immédiatement après utilisation.	4
La dose d'entretien recommandée est également de 7,5 mg par jour.	La posologie usuelle est de 7,5 mg de chlorhydrate de moexipril par jour.	5

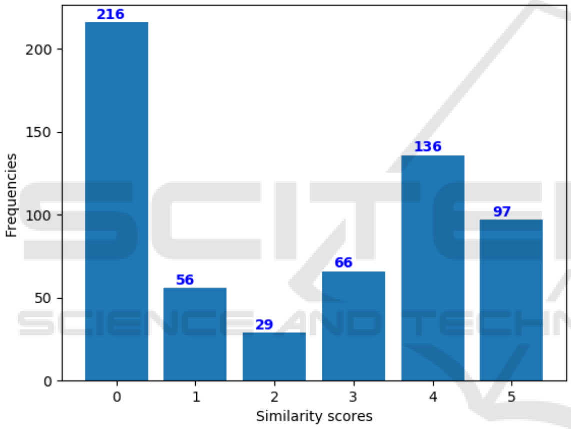

Figure 2: Distribution of similarity scores in the training set.

model, the results of the best combination are reported. The results of the proposed methods on the validation set (please see section 3.1) are presented in Table 2. According to the Pearson correlation, the LR model using outputs of the two other models as inputs got the best performance with a score of 0.8262. The MLP model slightly outperforms the RF one, while the latter yielded the highest accuracy with 0.6364.

Table 2: Results of the proposed models over the validation dataset.

Models	Pearson Correlation	Accuracy
RF model	0.8114	**0.6364**
MLP model	0.8132	0.6010
LR model	**0.8266**	-

Thereafter, the models were generated on the entire training set using the best combinations of features, which yielded the best results in the validation set. Table 3 shows the official *CONCORDIA* results during the DEFT 2020 challenge (Cardon et al., 2020). According to the EDRM, the MLP model got significantly better results. We also note that the RF model performed better than the LR model, which combines the similarity scores of the two other models. However, the latter yielded the highest Spearman correlation over the official test set. Compared to the other participating systems in the task 1 challenge, the proposed MLP model got the best performance (reaching an EDRM of 0.8217) (Cardon et al., 2020). Overall, *CONCORDIA* got higher scores than the average EDRM (0.7617). In addition, the two *CONCORDIA* best learning models, respectively MLP and RF, obtained an EDRM greater than (for MLP) or equal to (for RF) the median score (0.7947). According to the Spearman correlation, the LR-based learning model and MLP got the best performance (respectively 0.7769 and 0.7691) out of all the other methods presented at the task 1 of the DEFT 2020 challenge.

Table 3: Results of the proposed models over the official test set of the DEFT 2020.

Models	EDRM	Spearman Correlation
RF model	0.7947	0.7528
MLP model	**0.8217**	0.7691
LR model	0.7755	**0.7769**

4 DISCUSSION

4.1 Feature Importance

In order to estimate the relevance of the explored features for predicting the similarity between sentence pairs, the Pearson correlation score of each feature is computed over the entire training dataset (please see Table 4). The findings show that the 3-gram and 4-gram similarity measures obtained the best correlation scores (respectively, 0.7894 and 0.7854). They slightly outperformed the semantic similarity measure based on the word embedding (0.7746) and the 5-gram similarity (0.7734). In addition, we note that the Dice, Ochiai and TF.IDF based similarity measures performed well with correlation scores of over 0.76. Among the explored features, the Levenshtein similarity was the less important feature (with a correlation score of 0.7283) followed by the Jaccard similarity (0.7354) and the Manhattan distance (0.7354). These results are consistent with those of the related work (Chen et al., 2020; Soğancıoğlu et al., 2017) although the word embedding based measure got the highest Pearson correlation score in (Soğancıoğlu et al., 2017).

Table 4: Importance of each feature according to the Pearson correlation over the entire training dataset.

Feature	Pearson Correlation
Jaccard similarity	0.7354
Dice similarity	0.7644
Ochiai similarity	0.7630
Manhattan distance	0.7354
Q-gram similarity (Q=3)	0.7894
Q-gram similarity (Q=4)	0.7854
Q-gram similarity (Q=5)	0.7734
Levenshtein similarity	0.7283
TF-IDF similarity	0.7639
Word2vec similarity	0.7746

Using of together all these various similarity measures as features to support supervised methods did not yield the expected results. Therefore, some combinations of different similarity measures were experimented. The best performance (described in Section 3) was achieved with the following features: Dice, Ochiai, 3-gram, 4-gram, and Levenshtein similarity measures. These results show that these similarity measures complement each other and their combination in supervised methods allows improving the performances.

4.2 Analysis of the Results

The evaluation of the *CONCORDIA* semantic similarity approach over the DEFT 2020 dataset shows its effectiveness for this task. The results also show the relevance of the measures used to capture semantic similarity between different French sentences. In addition, all the *CONCORDIA*'s learning strategies allow to correctly estimate the semantic similarity between most of the sentence pairs of the official dataset.

However, extensive analysis of the results reveals limitations of these methods in predicting semantic similarity of some sentence pairs. The similarity measures used (Dice, Ochiai, Q-gram, Levenshtein) struggle to capture the semantics of sentences. Therefore, our methods fail to correctly predict similarity scores for sentences having similar terms, but which are semantically not equivalent. For example, for sentence pair 118 (id=118) in the test set, all methods estimated that the two sentences are roughly equivalent (with a similarity score of 4) while they are completely dissimilar according to the human experts (with similarity score of 0). On the other hand, our methods are limited in predicting the semantic similarity of sentences that are semantically equivalent but use different terms. For example, the sentences of pair 127 (id=127) are considered completely dissimilar (with a similarity score of 0) while they are roughly equivalent according to the human experts (with a similarity score of 4). To address these limitations, we proposed a semantic similarity measure based on words embedding. But the combination of this semantic measure with the above similarity measures in supervised methods did not allow to increase the performances.

We also performed an analysis of predictions errors of the proposed models according to the similarity classes. To this end, the Mean Squared Error (MSE) metric was used. Figure 3 shows the obtained results from the different models over the official test set of the DEFT 2020 challenge. Overall, the LR model significantly made fewer errors. Moreover, the MLP model performed slightly better than the RF model in all classes except class 4. These findings are consistent with the official results (Table 3) based on the Spearman correlation. They also show that the RF and MLP models made fewer errors in predicting classes 5 and 0 and more errors in predicting classes 2 and 3. We equally note that the proposed methods, especially the RF model and the MLP model, fail to predict the least representative classes (1, 2 and 3) in the training dataset. Indeed, in the official test dataset, classes 1 and 2 are respectively 37 and 28. The RF model does not predict any value in both classes, while the MLP model predicts only 9 values of the class 1.

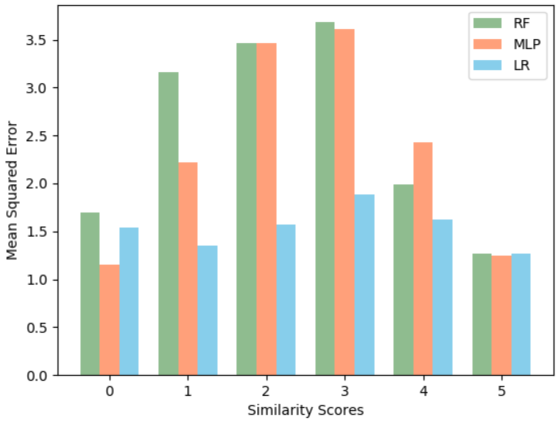

Figure 3: The Mean Squared Error of the proposed models according to the similarity classes over the test set.

4.3 Comparison with Other Participating Systems in DEFT 2020

At the task 1 of the DEFT 2020 challenge, most of the presented methods are based on similarity metrics or distances computation between sentences (Euclidean, Jaccard, Manhattan distances), with vector representations. These distances are then used as features to train machine learning models (Logistic Regression, Random Forest, etc.) (Cardon et al., 2020). Models of multilingual word embeddings derived from BERT (Bidirectional Encoder Representations from Transformers), in particular Sentence M-BERT and MUSE (Multilingual Universal Sentence Encoder) are also proposed but their performance is limited on this task.

If we compare *CONCORDIA* with the latter ones, it uses more advanced features to determine the degree of similarity between sentences. In addition, instead of combining all the explored similarity measures as features, we use a feature selection method to improve the performance of the machine learning models. Furthermore, it is based on classical ML algorithms for computing semantic sentence similarity.

According to the official DEFT 2020 challenge results, our MLP based method outperforms all the methods proposed at the task 1 of the challenge. This finding shows the effectiveness and the relevance of our promising proposition for measuring semantic similarity between sentences in the French clinical domain.

5 CONCLUSION AND FUTURE WORK

In this paper, we described the *CONCORDIA* approach based on supervised methods for computing semantic similarity between sentences in the French clinical domain. Three learning strategies have been proposed: RF, MLP and LR. The evaluation on a French standard dataset, from an established international challenge, showed that the MLP based strategy of *CONCORDIA* yielded the best results. Overall, the proposed approach achieved the best performances compared to the other best proposed methods which are presented during the DEFT 2020 challenge.

As of future work, to improve the performances of the approach, we plan to exploit additional features and similarity measures, especially those capable to capture the sentence semantics itself. A first experiment with word embedding on medium corpus did not allow to improve the results. Using a larger corpus could increase the performance. In addition, to overcome the limitation related to semantics, we plan to use more specialized biomedical resources, such as the UMLS (Unified Medical Language System) Metathesaurus. The latter contains various semantic resources, some of which are available in French (MeSH, Snomed CT, ICD 10, etc.). These resources could enable the exploitation of synonyms and semantic relations for computing the similarity between clinical sentences.

ACKNOWLEDGEMENTS

The authors would like to thank the DEFT 2020 challenge organizers who provided the evaluation dataset used in this study.

REFERENCES

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., and Wiebe, J. (2015). SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016). SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Bird, S. and Loper, E. (2004). NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Cardon, R., Grabar, N., Grouin, C., and Hamon, T. (2020). Presentation of the DEFT 2020 Challenge : open domain textual similarity and precise information extraction from clinical cases. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 1–13, Nancy, France. ATALA et AFCP.

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. Association for Computational Linguistics.

Chandrasekaran, D. and Mago, V. (2021). Evolution of Semantic Similarity—A Survey. *ACM Comput. Surv.*, 54(2). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Chen, Q., Du, J., Kim, S., Wilbur, W. J., and Lu, Z. (2020). Deep learning with sentence embeddings pretrained on biomedical corpora improves the performance of finding similar sentences in electronic medical records. *BMC Medical Informatics and Decision Making*, 20(1):73.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302.

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. Association for Computational Linguistics.

Grabar, N., Claveau, V., and Dalloux, C. (2018). CAS: French Corpus with Clinical Cases. In Lavelli, A., Minard, A.-L., and Rinaldi, F., editors, *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, Louhi@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 122–128. Association for Computational Linguistics.

Jaccard, P. (1912). The Distribution of the Flora in the Alpine Zone.1. *New Phytologist*, 11(2):37–50. _eprint: https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.1912.tb05611.x.

Jones, K. S. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*. Publisher: Emerald Group Publishing Limited.

Le, Q. V. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *arXiv:1405.4053 [cs]*. arXiv: 1405.4053.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]*. arXiv: 1310.4546.

Ochiai, A. (1957). Zoogeographical studies on the soleoid fishes found in Japan and its neighbouring regions-II. *Bull. Jpn. Soc. scient. Fish.*, 22:526–530.

P, S. and Shaji, A. P. (2019). A Survey on Semantic Similarity. In *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pages 1–8.

Rastegar-Mojarad, M., Liu, S., Wang, Y., Afzal, N., Wang, L., Shen, F., Fu, S., and Liu, H. (2018). BioCreative/OHNLP Challenge 2018. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, page 575, New York, NY, USA. Association for Computing Machinery.

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

Ukkonen, E. (1992). Approximate string-matching with q-grams and maximal matches. *Theoretical Computer Science*, 92(1):191–211.