Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

CLISTER : un corpus pour la similarité sémantique textuelle dans des cas cliniques en français*

Nicolas Hiebel[1] Olivier Ferret[2] Karën Fort[3] Aurélie Névéol[1]
(1) Universite Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(2) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
(3) Sorbonne Université / Université de Lorraine, CNRS, Inria, LORIA, 54506, Vandœuvre-lès-Nancy, France
[1]prenom.nom@lisn.upsaclay.fr, [2]olivier.ferret@cea.fr, [3]karen.fort@loria.fr

* Cet article est une adaptation en français de l'article (Hiebel *et al.*, 2022) publié en anglais à la conférence LREC 2022.

RÉSUMÉ

Le TAL repose sur la disponibilité de corpus annotés pour l'entraînement et l'évaluation de modèles. Il existe très peu de ressources pour la similarité sémantique dans le domaine clinique en français. Dans cette étude, nous proposons une définition de la similarité guidée par l'analyse clinique et l'appliquons au développement d'un nouveau corpus partagé de 1 000 paires de phrases annotées manuellement en scores de similarité. Nous évaluons ensuite le corpus par des expériences de mesure automatique de similarité. Nous montrons ainsi qu'un modèle de plongements de phrases peut capturer la similarité avec des performances à l'état de l'art sur le corpus DEFT STS (Spearman=0,8343). Nous montrons également que le contenu du corpus CLISTER est complémentaire de celui de DEFT STS.

ABSTRACT

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

Natural Language Processing relies on the availability of annotated corpora for training and evaluating models. There are very few resources for semantic similarity in the clinical domain in French. Herein, we introduce a definition of similarity guided by clinical facts and apply it to the development of a new shared corpus of 1,000 sentence pairs manually annotated with similarity scores. We 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 data set (Spearman=0.8343). We also show that CLISTER is complementary to DEFT STS.

MOTS-CLÉS : Similarité sémantique, Développement de corpus, Texte clinique, Français.

KEYWORDS: Semantic Similarity, Corpus Development, Clinical Text, French.

1 Introduction

La similarité sémantique est un problème du traitement automatique des langues visant à évaluer la proximité sémantique d'énoncés. La notion de similarité peut ainsi s'appliquer à la détection de paraphrases ou au résumé de texte. Ce problème a fait l'objet de nombreux travaux et campagnes d'évaluation, comme SemEval (Cer *et al.*, 2017) ou le Défi Fouilles de Textes (DEFT) (Cardon *et al.*, 2020).

Les travaux sur le calcul automatique de similarité au niveau phrastique s'appuient sur des ressources

comme les corpus STS Benchmark (Cer *et al.*, 2017) ou SICK (Marelli *et al.*, 2014) qui contiennent des paires de phrases annotées avec un degré de similarité (STS Benchmark) ou une relation de similarité (SICK). À notre connaissance, il n'existe pas de corpus de similarité en français à l'exception du corpus DEFT STS (Cardon & Grabar, 2020) et du corpus PAWS-X (Yang *et al.*, 2019), constitué de paraphrases traduites. Par ailleurs, la notion de similarité est difficile à définir et les corpus existants ne sont pas nécessairement accompagnés d'une définition de cette notion. Dans les domaines biomédical et médical qui nous occupent ici, plusieurs corpus de similarité sont disponibles en anglais. Le corpus BIOSSES (Soğancıoğlu *et al.*, 2017) contient 100 paires de phrases issues du challenge TAC sur le résumé de textes biomédicaux. Les critères de similarité définis dans ce corpus reposent sur la capacité des annotateurs à identifier les informations qui peuvent être considérées comme « importantes ». Les paires de phrases sont annotées avec un score de similarité sur une échelle allant de 0 à 4. Le corpus MedSTS (Wang *et al.*, 2020a) utilisé dans l'évaluation n2c2/OHNLP (Wang *et al.*, 2020b) contient 2 054 paires de phrases médicales en anglais annotées par deux experts du domaine médical. Ce corpus reprend la définition proposée dans BIOSSES en ajoutant un degré supplémentaire sur l'échelle de similarité (0 à 5). Le corpus français de similarité sémantique (Cardon & Grabar, 2020) utilisé dans DEFT 2020 contient 1 010 paires de phrases provenant du corpus CLEAR (Grabar & Cardon, 2018). La définition de la similarité utilisée dans ce corpus repose sur l'intuition des annotateurs.

Ces travaux montrent que l'annotation en similarité sémantique dans un domaine de spécialité est une tâche difficile. Définir des critères précis pour donner une définition de la similarité spécifique au domaine peut néanmoins permettre de créer une ressource de qualité. Ainsi, nous présentons une contribution au domaine de la similarité sémantique selon les axes suivants :

— une définition de la similarité guidée par des critères linguistiques et cliniques ;
— un nouveau corpus de 1 000 paires de phrases cliniques annotées avec un score de similarité ;
— une évaluation du corpus à l'aide d'un modèle de calcul de similarité offrant des performances à l'état de l'art ainsi qu'une comparaison avec le corpus DEFT STS existant.

2 Construction d'un corpus de similarité sémantique de phrases

2.1 Corpus source et sélection de paires de phrases candidates à l'annotation

Pour ce travail, nous nous sommes appuyés sur le corpus CAS (Grabar *et al.*, 2018), un corpus médical français contenant des descriptions de cas cliniques [1]. Le corpus a été découpé en phrases à l'aide de Talismane (Urieli, 2013). Pour sélectionner les phrases candidates à l'annotation, nous avons repris la méthode proposée par Wang *et al.* (2020b) en considérant que la sélection aléatoire de paires de phrases conduirait à un très fort déséquilibre vers des paires non similaires. Nous avons ainsi combiné deux métriques (distance de Levenshtein et similarité cosinus calculée sur les tokens) et sélectionné les paires de phrases dont la moyenne des métriques se situait au-delà du seuil de $0{,}45$.

2.2 Processus d'annotation

Définition des critères de similarité Nous avons défini la similarité en la décomposant selon trois dimensions de nature linguistique et clinique. En se concentrant sur ces dimensions, les annotateurs

1. Disponible auprès des organisateurs de DEFT : https://deft.limsi.fr/2020/

peuvent ainsi être plus cohérents. Nous présentons ici ces trois dimensions par ordre croissant de similarité.

Similarité surfacique : elle concerne la similitude structurelle. Cette similarité est fondée sur les mots grammaticaux ou les mots qui ne sont pas liés au domaine. Deux phrases présentant une similarité de surface peuvent ainsi être syntaxiquement proches mais sémantiquement éloignées.

Similarité sémantique : celle-ci s'appuie sur les concepts médicaux. Plus les concepts d'une phrase sont proches des concepts d'une autre phrase, plus la similarité de ces phrases est élevée. Ces concepts peuvent faire référence à des médicaments, des maladies, des procédures, etc.

Compatibilité clinique : au-delà d'une similarité fondée sur le simple partage de concepts, la compatibilité clinique évalue si les phrases d'une paire peuvent se référer au même cas clinique.

Pour évaluer les deux derniers critères, les annotateurs ont pu avoir recours à des ressources externes comme la base multilingue *Unified Medical Language System* (UMLS) (Lindberg *et al.*, 1993).

Échelle de notation et exemples Comme dans d'autres corpus de la littérature (Cer *et al.*, 2017; Wang *et al.*, 2020a), nous utilisons une échelle de score de 0 (similarité minimum) à 5 (similarité maximum). Étant donnés les critères de sélection des paires de phrases candidates, même les paires du corpus ayant un score de similarité de 0 présentent une certaine similarité surfacique.

L'exemple (1) présente une paire de phrases ayant un score de similarité de 0. Ce score correspond à des paires de phrases n'ayant qu'une similarité surfacique. Les deux phrases de l'exemple (1) ont une structure similaire, commençant par « *Il n'y avait / n'avait pas de...* ». Cependant, la partie restante des deux phrases n'est absolument pas liée.

(1) a. Il n'y avait pas de résidu post-mictionnel.
b. Il n'avait pas de facteurs de risque cardiovasculaire notable.

L'exemple (2) présente une paire de phrases avec un score de similarité de 1. Ce niveau est associé aux paires de phrases caractérisées par une similarité de surface concernant au plus une entité médicale. Les deux phrases de l'exemple (2) ont une structure commune, un certain type d'examen médical révélant un symptôme. Mais cette structure mise à part, les deux phrases ne sont pas liées.

(2) a. L'examen physique révélait une légère sensibilité de la fosse lombaire droite.
b. L'examen O.R.L. retrouvait une légère surdité de perception.

On peut observer sur l'exemple (3) une paire de phrases avec un score de 2, correspondant aux paires contenant des concepts médicaux faiblement similaires sur le plan sémantique mais n'ayant aucune compatibilité clinique. Typiquement, les phrases d'une paire peuvent concerner une maladie, une procédure ou un médicament. Les deux phrases de (3) font ainsi référence à un examen d'imagerie, qui est différent d'une phrase à l'autre mais aboutit à un diagnostic commun. Cet exemple montre également les difficultés potentielles de l'annotation dans un domaine technique car juger de la proximité de ces deux examens n'est pas forcément évident.

(3) a. La TDM cérébrale n'a pas révélé d'anomalie.
b. La scintigraphie n'a pas montré d'anomalie.

Le score de similarité 3 correspond aux paires de phrases présentant une similarité sémantique sur plusieurs concepts médicaux les rendant partiellement compatibles sur le plan clinique. Dans l'exemple (4), les deux phrases concernent la présence d'une tumeur conduisant à une exploration chirurgicale. Sur ces éléments, les deux phrases sont cliniquement compatibles mais les tumeurs que les phrases décrivent ne sont pas situées au même endroit, ce qui n'est pas cliniquement compatible. La phrase (4-a) contient également une description plus précise des symptômes.

(4) a. Devant cet aspect non spécifique d'une tumeur rétropéritonéale isolée entrainant des signes digestifs importants, une exploration chirurgicale était décidée.
b. Devant ce tableau de tumeur rénale, l'indication d'une exploration chirurgicale était posée.

Le score de similarité de 4 associé à l'exemple (5) correspond quant à lui aux paires de phrases présentant une similarité sémantique et une compatibilité clinique élevées. On remarque que la phrase (5-a) inclut des informations non présentes dans (5-b) (sexe du patient, caractère chronique de la pathologie dénoté par le terme « rémission »). Ces éléments ne créent toutefois pas de contradiction entre les énoncés.

(5) a. La patiente est en rémission complète avec un recul de 12 mois.
b. L'évolution était bonne avec un recul de 27 mois.

Le score de similarité 5 correspond aux paires de phrases présentant une similarité sémantique élevée et une compatibilité clinique totale. Les phrases ont globalement le même sens mais l'une peut être plus spécifique que l'autre. Nous faisons ici la différence entre être plus spécifique et contenir plus d'informations. Les phrases de l'exemple (6) présentent une grande similarité. La seule différence est la précision ajoutée sur les marqueurs tumoraux dans la phrase (6-a) mais les phrases sont équivalentes.

(6) a. Les marqueurs tumoraux (CA 15.3 et ACE) étaient normaux.
b. Les marqueurs tumoraux sériques étaient normaux.

2.3 Équilibrer les catégories dans le corpus

Afin d'obtenir une bonne représentation des catégories extrêmes et pour augmenter la taille du corpus en vue d'entraîner des modèles de calcul de similarité, nous avons étendu le corpus en ajoutant semi-automatiquement des paires de phrases avec des scores de 0, 4 et 5.

Pour le score 0, nous avons collecté 210 paires de phrases aléatoirement en partant de l'hypothèse que deux phrases sélectionnées aléatoirement ont très peu de chances d'être similaires. La similarité entre ces phrases a été calculée en utilisant une version pré-entraînée et multilingue de SENTENCE-BERT (Reimers & Gurevych, 2019) et les paires ayant les scores les moins faibles ont été vérifiées manuellement. Sur les paires vérifiées, il a été estimé que deux paires auraient pu avoir un score de similarité de 1. Celles-ci ont été supprimées et remplacées.

Les paires de phrases présentant une forte similarité ont aussi été sélectionnées à l'aide de SENTENCE-BERT et de la bibliothèque Faiss (Johnson *et al.*, 2021). Les plongements de phrases ont été calculés pour toutes les phrases du corpus et une matrice de similarité a été créée. Nous avons

annoté manuellement les paires les mieux classées en obtenant ainsi 190 paires supplémentaires ayant un score de similarité de 4 ou 5.

3 Annotation manuelle de la similarité sémantique de phrases

3.1 Processus d'annotation

Quatre annotateurs avec une expertise complémentaire dans les représentations sémantiques, la création de ressources annotées et l'annotation de textes biomédicaux sont intervenus.

Au cours d'un premier cycle d'annotation, trois annotateurs ont annoté un échantillon de 100 paires de phrases indépendamment, sans définition de catégories. Cette première étape avait pour objectif de tester l'intuition des annotateurs et d'apprécier la difficulté de la tâche d'annotation. Nous avons calculé l'accord inter-annotateur en utilisant l'α de Krippendorff (Krippendorff, 2013). Pour cet échantillon l'α obtenu était faible ($0{,}239$), ce qui montre un désaccord important entre les annotateurs.

Ce premier échantillon annoté a servi de support à la discussion et à la conception du guide d'annotation. L'annotation finale des paires de phrases de cet échantillon a fait l'objet d'un consensus entre les annotateurs. Un deuxième échantillon de 100 paires de phrases a ensuite été annoté. Les trois annotateurs initiaux ont été rejoints par un quatrième annotateur, qui n'a pas participé à la discussion lors du premier échantillon. Outre l'ensemble supplémentaire d'annotations, cette nouvelle contribution a permis d'évaluer la qualité du guide d'annotation, qui doit être compréhensible pour les annotateurs autres que les auteurs.

L'accord inter-annotateur a été calculé pour le deuxième échantillon. Il s'est avéré suffisant (α de Krippendorff de $0{,}689$), de sorte que chaque annotateur a travaillé indépendamment sur autre échantillon de 100 paires de phrases. Au total, 600 paires de phrases ont été annotées manuellement.

3.2 Statistiques globales

Le corpus final de 1 000 paires de phrases annotées contient 30 942 tokens (ponctuation incluse), selon la sortie de l'analyseur Talismane. La figure 1 montre le nombre de paires de phrases pour chaque catégorie. Grâce à la sélection semi-automatique des paires de phrases représentant une similarité élevée et faible, les scores les plus élevés et les plus bas sont désormais prédominants.

4 Évaluation extrinsèque de la similarité sémantique textuelle

4.1 Calcul de la similarité de phrases

Lithgow-Serrano *et al.* (2019) et Lara-Clares *et al.* (2021) proposent un état de l'art dans le domaine biomédical des méthodes de calcul de similarité sémantique ainsi que des corpus disponibles pour l'anglais. On retiendra notamment que les méthodes de calcul de similarité peuvent s'appuyer sur la comparaison au niveau des chaînes de caractères, comme la distance de Levenshtein (Levenshtein, 1966). D'autres méthodes proposent une représentation vectorielle des énoncés avec divers degrés de

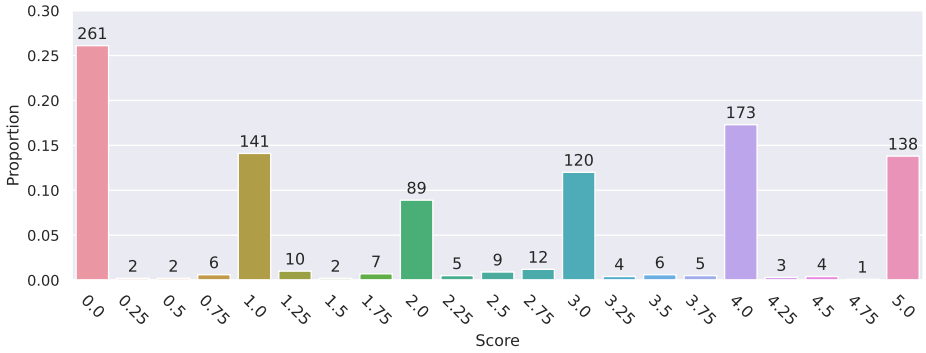

FIGURE 1 – Distribution du nombre de paires de phrases pour chaque score de similarité dans le corpus final (1 000 paires). Ces scores sont des moyennes sur les différents annotateurs.

représentation sémantique (par exemple, TF-IDF ou des plongements de phrases). Ces représentations peuvent ensuite être comparées à l'aide de métriques vectorielles usuelles, comme la similarité cosinus. Dans notre travail, nous avons utilisé le modèle SENTENCE-BERT, qui permet de produire des plongements de phrases de taille fixe grâce à une architecture de type réseau siamois s'appuyant sur le modèle de langue BERT (Devlin *et al.*, 2019).

4.2 Expériences sur CLISTER

Ces expériences visent à valider que notre définition de la similarité peut être apprise par un modèle entraîné puis testé sur CLISTER.

Les 1 000 paires de phrases ont été réparties entre entraînement et test (600/400) en conservant dans le jeu de test l'échantillon annoté par consensus et en divisant aléatoirement le reste des données entre ces deux ensembles. Les expériences ont été menées à l'aide d'un modèle pré-entraîné multilingue de SENTENCE-BERT [2] couvrant 15 langues, dont le français [3] (Reimers & Gurevych, 2020). Le modèle évalué est le produit de l'ajustement (*fine-tuning*) de ce modèle multilingue sur la partie d'entraînement de CLISTER. Globalement, les meilleurs résultats ont été obtenus avec 5 itérations (*epochs*) et 10 étapes d'échauffement (*warmup steps*). Nous présentons ici les valeurs moyennes obtenues sur trois itérations avec ces paramètres.

Le tableau 1 présente l'évaluation de la similarité sur les données de test de CLISTER ($CLISTER_{400}$), à l'aide de la corrélation de Spearman et de l'EDRM (Exactitude en Distance Relative à la Solution Moyenne), une métrique proposée dans DEFT pour une évaluation nuancée de la similarité des phrases (Cardon *et al.*, 2020). L'ajustement du modèle sur les données d'entraînement de CLISTER permet une amélioration des performances (EDRM : $+0{,}1261$, Spearman : $+0{,}1123$). L'utilisation de SENTENCE-BERT dans le processus de sélection de paires de phrases pour les catégories extrêmes peut introduire un biais dans l'évaluation puisque les prédictions du modèle s'aligneront sur la référence pour ces paires de phrases. Nous avons donc restreint cette évaluation au sous-ensemble de test annoté par consensus ($CLISTER_{100}$), donc non influencé par la sélection semi-automatique. Avec cette restriction, les valeurs d'EDRM ($0{,}6323$ vs $0{,}7149$) et de Spearman

2. https://www.sbert.net/docs/pretrained_models.html#multi-lingual-models
3. *distiluse-base-multilingual-cased-v1*

Données d'ajustement	Données de test	EDRM	Spearman
Aucune	$CLISTER_{100}$	0,6323	0,3794
CLISTER	$CLISTER_{100}$	0,8240	0,7340
Aucune	$CLISTER_{400}$	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
Aucune	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
Aucune	$CLISTER_{400}$ + 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 1 – Résultats des expériences de similarité sémantique avec différentes combinaisons de corpus d'ajustement et de test en utilisant le modèle multilingue de SENTENCE-BERT.

(0,3794 vs 0,7547) sont plus faibles lorsque le modèle n'est pas ajusté. Avec ajustement, on observe une augmentation des valeurs d'EDRM (+0,1917) et de Spearman (+0,3546).

Ces améliorations de performance suggèrent que le modèle peut être adapté à notre définition de la similarité avec une quantité modeste de données d'entraînement.

4.3 Évaluations croisées : expériences sur CLISTER et DEFT STS

Ces expériences permettent d'évaluer la contribution du nouveau corpus pour le calcul de similarité. Les deux corpus étant comparables en taille et répartition entraînement/test, une comparaison directe est pertinente. Le tableau 1 présente les résultats de ces expériences avec différentes combinaisons de données d'ajustement et d'ensembles de test en utilisant le modèle multilingue pré-entraîné de SENTENCE-BERT [4]. On observe de meilleurs résultats sur $CLISTER_{400}$ et DEFT STS en ajustant le modèle sur les données d'entraînement correspondantes. À l'inverse, les résultats obtenus en ajustant sur un corpus et en testant sur l'autre sont moins bons que ceux obtenus sans ajustement. Enfin, le modèle ajusté sur les deux corpus d'entraînement présente des performances légèrement inférieures à celles obtenues en ajustant sur les données d'entraînement correspondantes aux données de test. Lors de l'évaluation sur la combinaison des jeux de test $CLISTER_{400}$ et DEFT STS, les ajustements sur les corpus individuels vont chacun améliorer les performances du modèle sans ajustement. Les meilleurs résultats sont obtenus en ajustant sur les deux corpus d'entraînement.

Ces résultats montrent que le corpus DEFT STS et le corpus CLISTER que nous avons créé sont très différents. Cette différence de performance et cette non-compatibilité entre les deux corpus peuvent être liées à la nature des données (cliniques pour CLISTER, encyclopédiques pour DEFT STS) et/ou à la définition de la similarité qui sous-tend les scores de similarité dans les corpus.

4. *distiluse-base-multilingual-cased-v1*

4.4 Comparaison avec les résultats de DEFT

Les résultats de nos expériences utilisant DEFT STS pour l'entraînement et le test peuvent être directement comparés à ceux des systèmes soumis à DEFT. Les meilleures performances pour cette tâche ont été obtenues par une méthode représentant les paires de phrases par des caractéristiques de similarité exploitant un large éventail de scores de similarité et entraînant un classifieur ensembliste sur cette représentation (Dramé *et al.*, 2020). La méthode utilisée dans nos expériences obtient pour sa part une corrélation de Spearman plus élevée ($0,8343$ vs. $0,7769$) mais un EDRM plus faible ($0,8217$ vs. $0,7926$) en produisant une unique représentation de chaque phrase.

5 Conclusion

Nous avons présenté un nouveau corpus de 1 000 paires de phrases dans le domaine clinique en français, annotées en degrés de similarité selon une définition s'appuyant sur des critères linguistiques et cliniques. Les évaluations réalisées montrent la contribution de cette nouvelle ressource pour la tâche de calcul automatique de similarité. Le corpus est librement disponible en ligne : https://gitlab.inria.fr/codeine/clister.

La disponibilité d'un corpus de similarité dans le domaine clinique peut aider à la construction de modèles de recherche d'information dans ce domaine. Dans nos travaux à venir, nous prévoyons ainsi d'utiliser le corpus CLISTER afin d'entraîner un modèle SENTENCE-BERT pour la recherche de phrases similaires dans un corpus clinique et d'évaluer ainsi la confidentialité de phrases issues de documents cliniques vis-à-vis de la problématique de la réidentification possible des patients.

Remerciements

Nous remercions Natalia Grabar pour sa réactivité à nous fournir le corpus de paires de phrases utilisé pour la tâche DEFT (DEFT STS). Ce travail a été réalisé dans le cadre d'un projet de l'Agence Nationale de la Recherche, CODEINE (artificial text COrpus DEsIgNed Ethically), ANR-20-CE23-0026-01.

Références

CARDON R. & GRABAR N. (2020). A French corpus for semantic similarity. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 6889–6894, Marseille, France : European Language Resources Association.

CARDON R., GRABAR N., GROUIN C. & 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*, p. 1–13, Nancy, France : Association pour le Traitement Automatique des Langues. Presentation of the DEFT 2020 Challenge : open domain textual similarity and precise information extraction from clinical cases.

CER D., DIAB M., AGIRRE E. E., LOPEZ-GAZPIO I. & 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), p. 1 – 14, Vancouver, Canada : Steven Bethard and Marine Carpuat and Marianna Apidianaki and Saif M. Mohammad and Daniel Cer and David Jurgens. HAL : hal-01560674.

DEVLIN J., CHANG M.-W., LEE K. & 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)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DRAMÉ K., SAMBE G., DIOP I. & 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*, p. 49–54, Nancy, France : Association pour le Traitement Automatique des Langues. Supervised approach to compute semantic similarity between sentence pairs.

GRABAR N. & CARDON R. (2018). CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, p. 3–9, Tilburg, the Netherlands : Association for Computational Linguistics. DOI : 10.18653/v1/W18-7002.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, p. 122–128, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5614.

HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2022). CLISTER : A Corpus for Semantic Textual Similarity in French Clinical Narratives. In *13^{th} Language Resources and Evaluation Conference (LREC 2022)*, Marseille, France.

JOHNSON J., DOUZE M. & 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*. Thousand Oaks, CA : Sage, 3rd edition édition.

LARA-CLARES A., LASTRA-DÍAZ J. J. & 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.

LINDBERG D., HUMPHREYS B. & MCCRAY A. (1993). The unified medical language system. *Yearb Med Inform*, **1**, 41–51.

LITHGOW-SERRANO O., GAMA-CASTRO S., ISHIDA-GUTIÉRREZ C., MEJÍA-ALMONTE C., TIERRAFRÍA V. H., MARTÍNEZ-LUNA S., SANTOS-ZAVALETA A., VELÁZQUEZ-RAMÍREZ D. & COLLADO-VIDES J. (2019). Similarity corpus on microbial transcriptional regulation. *Journal of biomedical semantics*, **10**(1), 1–14.

MARELLI M., MENINI S., BARONI M., BENTIVOGLI L., BERNARDI R. & ZAMPARELLI R. (2014). The SICK (Sentences Involving Compositional Knowledge) dataset for relatedness and entailment. DOI : 10.5281/zenodo.2787612.

REIMERS N. & GUREVYCH I. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *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*, p. 3980–3990 : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

REIMERS N. & 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)*, p. 4512–4525, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.365.

SOĞANCIOĞLU G., ÖZTÜRK H. & ÖZGÜR A. (2017). BIOSSES : a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, **33**(14), i49–i58. DOI : 10.1093/bioinformatics/btx238.

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

WANG Y., AFZAL N., FU S., WANG L., SHEN F., RASTEGAR-MOJARAD M. & LIU H. (2020a). MedSTS : a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, **54**. DOI : 10.1007/s10579-018-9431-1.

WANG Y., FU S., SHEN F., HENRY S., UZUNER O. & LIU H. (2020b). The 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity : Overview. *JMIR medical informatics*, **8**(11). DOI : 10.2196/23375.

YANG Y., ZHANG Y., TAR C. & BALDRIDGE J. (2019). PAWS-X : A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3687–3692, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1382.