Détection automatique de phrases parallèles dans un corpus biomédical comparable technique/simplifié

Rémi Cardon Natalia Grabar
CNRS, UMR 8163, F-59000 Lille, France
Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
remi.cardon@univ-lille.fr, Natalia.grabar@univ-lille.fr

RÉSUMÉ

Les phrases parallèles contiennent des informations identiques ou très proches sémantiquement et offrent des indications importantes sur le fonctionnement de la langue. Lorsque les phrases sont différenciées par leur registre (comme expert *vs.* non-expert), elles peuvent être exploitées pour la simplification automatique de textes. Le but de la simplification automatique est d'améliorer la compréhension de textes. Par exemple, dans le domaine biomédical, la simplification peut permettre aux patients de mieux comprendre les textes relatifs à leur santé. Il existe cependant très peu de ressources pour la simplification en français. Nous proposons donc d'exploiter des corpus comparables, différenciés par leur technicité, pour y détecter des phrases parallèles et les aligner. Les données de référence sont créées manuellement et montrent un accord inter-annotateur de 0,76. Nous expérimentons sur des données équilibrées et déséquilibrées. La F-mesure sur les données équilibrées atteint jusqu'à 0,94. Sur les données déséquilibrées, les résultats sont plus faibles (jusqu'à 0,92 de F-mesure) mais restent compétitifs lorsque les modèles sont entraînés sur les données équilibrées.

ABSTRACT

Automatic detection of parallel sentences in comparable biomedical corpora

Parallel sentences provide identical or semantically similar information which gives important clues on language. When sentences vary by their register (like expert *vs* non-expert), they can be exploited for the automatic text simplification. The aim of text simplification is to improve the understanding of texts. For instance, in the biomedical field, simplification may permit patients to understand better medical texts in relation to their health. Yet, there is currently very few resources for the simplification of French texts. We propose to exploit comparable corpora, which are distinguished by their technicality, to detect parallel sentences and to align them. The reference data are created manually and show 0.76 inter-annotator agreement. We perform experiments on balanced and imbalanced data. The results on balanced data reach up to 0.94 F-measure. On imbalanced data, the results are lower (up to 0.92 F-measure) but remain competitive when using classification models trained on balanced data.

MOTS-CLÉS : Simplification, classification, similarité, phrases parallèles, corpus comparables, domaine médical.

KEYWORDS: Simplification, classification, similarity, parallel sentences, comparable corpora, medical domain.

1 Introduction

Les phrases parallèles possèdent une sémantique similaire et la véhiculent d'une manière qui peut varier selon un axe donné. Typiquement, les phrases parallèles sont collectées dans deux langues différentes et correspondent à des traductions. Dans la langue générale, le corpus Europarl (Koehn, 2005) contient de telles phrases dans plusieurs paires de langues. Cependant, l'axe à partir duquel on observe le parallélisme peut se trouver à d'autres niveaux, comme le registre de langue expert *vs.* non-expert. La paire de phrases ci-après illustre la différence de technicité entre deux phrases :

— Expert : *Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation*
— Non-expert : *Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou ralentir le transit intestinal*

Les paires de phrases parallèles fournissent des informations utiles sur le lexique, les structures syntaxiques, les traits stylistiques, etc. propres à un registre donné. Ainsi, les paires collectées dans différentes langues servent à la traduction automatique, alors que les paires ayant une technicité différente peuvent servir à la simplification automatique. L'objectif de la simplification consiste à produire une version simplifiée d'un texte, afin d'éliminer, de restructurer ou de remplacer les segments difficiles, tout en maintenant le sens. La simplification peut s'occuper de différents aspects, comme le lexique, la syntaxe, la sémantique, la pragmatique ou encore la structure d'un document.

La simplification automatique de textes peut être une étape de pré-traitement pour les applications de TAL ou pour produire des versions adaptées de textes pour différents utilisateurs humains. Par exemple, les documents simplifiés peuvent être destinés aux enfants (Vu *et al.*, 2014), aux étrangers ou personnes mal alphabétisées (Paetzold & Specia, 2016), aux personnes souffrant de pathologies mentales ou neurodégénératives (Chen *et al.*, 2016), ou au grand public qui cherche à comprendre des documents spécialisés (Leroy *et al.*, 2013). Notre objectif est de préparer les ressources nécessaires pour la création de documents médicaux simplifiés pour le grand public. Il a en effet été montré que les documents médicaux sont souvent difficiles à comprendre par les patients et leurs familles (AMA, 1999; Mcgray, 2005; Rudd, 2013), ce qui peut avoir une influence négative sur le processus de soins. Plus particulièrement, nous proposons de détecter des phrases parallèles et de les aligner. Comme il n'existe pas de textes parallèles en français qui soient différenciés par leur technicité, nous proposons d'exploiter le corpus comparable CLEAR (Grabar & Cardon, 2018), où les textes traitent des mêmes sujets mais diffèrent par leur technicité. Nous étudions également l'influence du déséquilibre sur la difficulté de la tâche, qui est en effet une caractéristique naturelle des données textuelles.

Il existe plusieurs travaux en détection et alignement de phrases parallèles au sein de corpus comparables bilingues pour les besoins de la traduction automatique. Différentes méthodes sont exploitées pour cela, comme des systèmes de recherche d'information cross-langue (Utiyama & Isahara, 2003; Munteanu & Marcu, 2006), des arbres d'alignement de séquences (Munteanu & Marcu, 2002) ou des traductions automatiques mutuelles (Yang & Li, 2003; Munteanu & Marcu, 2005; Kumano *et al.*, 2007; Abdul-Rauf & Schwenk, 2009). Souvent, il est nécessaire d'effectuer ensuite un filtrage pour la sélection de bonnes propositions. En ce qui concerne les travaux en alignement de phrases parallèles dans les corpus comparables monolingues, la difficulté principale est liée au faible chevauchement lexical entre les phrases. Récemment, cette tâche a gagné en popularité autour de la langue générale grâce aux tâches STS (semantic text similarity) et les compétitions *SemEval* (Agirre *et al.*, 2013, 2015, 2016) : pour une paire de phrases donnée, l'objectif consiste à prédire leur niveau de similarité sémantique et d'y attribuer un score allant de 0 (sémantique indépendante) à 5 (identité sémantique). Plusieurs types de méthodes sont proposés :

— Les *méthodes lexicales* se basent sur la similarité des mots ou des segments sublexicaux

(Madnani *et al.*, 2012). Nous trouvons parmi les descripteurs exploités : chevauchement lexical, longueur des phrases, distance d'édition des chaînes de caractères, nombres, entités nommées, la sous-chaîne commune la plus longue (Clough *et al.*, 2002; Zhang & Patrick, 2005; Qiu *et al.*, 2006; Nelken & Shieber, 2006; Zhu *et al.*, 2010) ;

— Les *méthodes basées sur des connaissances externes* utilisent des ressources comme WordNet (Miller *et al.*, 1993) ou PPDB (Ganitkevitch *et al.*, 2013). Parmi les descripteurs exploités se trouvent : recoupement avec des ressources externes, distance et intersection entre synsets, similarité sémantique entre les graphes, présence de synonymes, hyperonymes ou antonymes (Mihalcea *et al.*, 2006; Fernando & Stevenson, 2008; Lai & Hockenmaier, 2014) ;

— Les *méthodes basées sur la syntaxe* exploitent la modélisation syntaxique des phrases. Les descripteurs utilisés sont : catégories morphosyntaxiques, chevauchement syntaxique, dépendances syntaxiques, constituants, relations prédicatives, distance d'édition entre arbres syntaxiques (Wan *et al.*, 2006; Severyn *et al.*, 2013; Tai *et al.*, 2015; Tsubaki *et al.*, 2016) ;

— Les *méthodes basées sur les corpus* exploitent par exemple des méthodes distributionnelles, la LSA et les plongements lexicaux (Barzilay & Elhadad, 2003; Guo & Diab, 2012; Zhao *et al.*, 2014; Kiros *et al.*, 2015; He *et al.*, 2015; Mueller & Thyagarajan, 2016).

À notre connaissance, il n'existe pas de travaux en détection et alignement de phrases parallèles en domaine spécialisé, comme le domaine biomédical. Nous pensons que cela peut rendre la tâche d'alignement de phrases plus compliquée parce qu'il existe une variation lexicale importante entre le registre technique et simplifié, comme on peut le voir dans les exemples présentés au début de cette section. Dans ce qui suit, nous présentons d'abord les données linguistiques utilisées et les méthodes proposées. Ensuite nous présentons et discutons les résultats obtenus, avant de conclure avec des perspectives sur le travail à venir.

2 Données linguistiques

Nous utilisons le corpus comparable médical CLEAR disponible en ligne [1] qui contient trois sous-corpus comparables en français. Les documents de ces sous-corpus sont regroupés par paires : les textes de chaque paire traitent du même sujet mais varient par leur degré de technicité. Trois genres sont représentés : l'information sur les médicaments, les résumés de la littérature scientifique médicale et des articles encyclopédiques. Au total, ce corpus contient 16 190 paires de documents, avec plus de 15M d'occurrences de mots dans la version technique et 35M d'occurrences dans la version simplifiée.

Les données de référence sont créées manuellement à partir de 39 paires de textes sélectionnés aléatoirement. La référence contient les paires de phrases alignées, qui associent les contenus techniques et simplifiés. L'alignement est effectué par deux annotateurs indépendants selon ces critères :

1. exclure les paires de phrases identiques ou variant par la ponctuation ou mots grammaticaux ;
2. inclure les paires de phrases avec des variations morphologiques, comme dans : *Ne pas dépasser la posologie recommandée.* et *Ne dépassez pas la posologie recommandée.* ;
3. inclure les paires de phrases avec une sémantique équivalente, comme dans : *Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation.* et *Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou ralentir le transit intestinal.* ;
4. inclure les paires de phrases où une phrase est comprise dans l'autre, ce qui permet d'associer plusieurs phrases à une seule, comme dans : *C'est un organe fait de tissus membraneux et*

1. http://natalia.grabar.free.fr/resources.php#clear

musculaires, d'environ 10 à 15 mm de long, qui pend à la partie moyenne du voile du palais. et *Elle est constituée d' un tissu membraneux et musculaire.* ;

5. exclure les paires de phrases avec intersection sémantique, où chaque phrase contient de l'information qui lui est propre, en plus de la sémantique commune. Si l'accès à l'information commune aux deux phrases est intéressant en soi, cela complexifie grandement la tâche de reconnaissance des phrases parallèles. Par ailleurs, il devient plus difficile d'exploiter ce type de paires de phrases pour la préparation du lexique et de règles de simplification.

Suite au consensus, 663 paires de phrases alignées ont été produites. L'accord inter-annotateur est de 0,76 (Cohen, 1960). Le tableau 1 présente les données de référence obtenues : le nombre de documents, de phrases et d'occurrences de mots pour les sous-corpus et les registres technique et simplifié, de même que le taux d'alignement entre ces deux registres. Ces informations sont détaillées pour chaque sous-corpus : informations sur les médicaments (*Méd.*), littérature scientifique (*Sci.*) et articles encyclopédiques (*Wiki.*). Le taux d'alignement est indiqué car, dû aux principes de création des versions simplifiées, ces trois corpus ne montrent pas la même capacité à produire des paires de phrases parallèles. Le taux d'alignement correspond au rapport entre le nombre de phrases alignées et le nombre total de phrases dans un corpus donné. De manière peu surprenante, seule une petite fraction de toutes les paires possibles peut être alignée.

TABLE 1 – Taille des données de référence avec l'alignement consensuel : le nombre de documents, de phrases et d'occurrences de mots pour chaque sous-corpus et registre, et le taux d'alignement

		Expert				*Non-expert*				*Taux d'*	
		source		*aligné*		*source*		*aligné*		*alignement (%)*	
corpus	*nb docs*	*nb ph.*	*nb occ.*	*nb paires*	*nb occ.*	*nb ph.*	*nb occ.*	*nb paires*	*nb occ.*	*ex.*	*non-ex.*
Méd.	12*2	4 416	44 709	502	5 751	2 736	27 820	502	10 398	18	11
Sci.	13*2	553	8 854	112	3 166	263	4 688	112	3 306	20	43
Wiki.	14*2	2 494	36 002	49	1 100	238	2 659	49	853	2	21

3 Détection et alignement de phrases parallèles

Les documents sont d'abord segmentés en phrases en exploitant la ponctuation forte (*i.e. . ? ! ; :*). Nous avons aussi retiré, au sein de chaque sous-corpus, les phrases qui apparaissent au moins dans la moitié des documents (typiquement, les mentions légales et les titres de sections) et les phrases sans caractères alphabétiques. Cela réduit la combinatoire de phrases de 940 000 à 590 000 environ.

Nous abordons la détection et l'alignement automatique de phrases parallèles comme un problème de classification, où il s'agit d'assigner chaque paire de phrases analysées à l'une des deux catégories :

— *alignées* : les phrases sont parallèles et peuvent être alignées ;
— *non-alignées* : les phrases ne sont pas parallèles et ne peuvent pas être alignées.

Les données de référence fournissent 663 exemples positifs. Les exemples négatifs contiennent toutes les paires de phrases possibles, pour chaque paire de documents, en excluant les paires de phrases alignées, soit environ 590 000 paires de phrases au total. Nous avons effectué des tests d'alignement de phrases avec plusieurs classifieurs du module `scikit-learn` [2] : *Perceptron* (Rosenblatt, 1958),

2. `https://scikit-learn.org/stable/`

Multilayer Perceptron (MLP) (Rosenblatt, 1961), *Linear discriminant analysis* (LDA) (Fisher, 1936), *Quadratic discriminant analysis* (QDA) (Cover, 1965), *Stochastic gradient descent* (SGD), *Linear SVM* (Vapnik & Lerner, 1963). Les modèles obtenus avec le classification binaire issu de la régression logistique montrent les meilleurs résultats. Ce sont donc ces résultats que nous reportons dans la suite de la présentation.

Nos expériences sont basées sur plusieurs types de descripteurs calculés sur les textes non lemmatisés. Plusieurs combinaisons de ces descripteurs ont été testées (Cardon & Grabar, 2018), ce qui nous a permis de sélectionner les descripteurs les plus efficaces :

— Nombre de mots communs, à l'exception des mots vides. Ce descripteur permet de calculer l'intersection lexicale de base entre les versions techniques et simplifiées des phrases ;
— Pourcentage de mots d'une phrase inclus dans l'autre, dans les deux directions. Ce descripteur représente de possibles relations d'inclusion lexicale entre les phrases ;
— Différence de longueur entre les phrases. Ce descripteur vise les inclusions lexicales et suppose que la simplification peut impliquer une association stable avec la longueur des phrases ;
— Différence de la longueur moyenne des mots entre les phrases. Ce descripteur est similaire au précédent mais il prend en compte la différence moyenne de la longueur des phrases ;
— Nombre total de bigrammes et de trigrammes en communs (un descripteur pour chaque). Ce descripteur est calculé sur les n-grammes de caractères. La supposition est que, au niveau de caractères, certaines séquences plus petites que les mots peuvent aussi être partagées par les deux phrases et donc être significatives pour leur alignement ;
— Mesures de similarité (cosinus, Dice et Jaccard). Ce descripteur fournit une indication plus sophistiquée sur l'intersection lexicale des phrases. Le poids de chaque mot est de 1 ;
— Distance d'édition de Levenshtein (Levenshtein, 1966). La distance prend en compte les opérations d'édition de base (insertion, suppression et substitution) au niveau des caractères (acception classique) et au niveau des mots. Le coût de chaque opération est de 1.

Les paires alignées manuellement sont divisées en trois sous-ensembles :

E : 238 paires avec équivalence sémantique,
TIS : 237 paires où le contenu de la phrase technique est entièrement compris dans la phrase simplifiée. La phrase simplifiée contient donc une sémantique supplémentaire,
SIT : 112 paires où le contenu de la phrase simplifiée est entièrement compris dans la phrase technique. La phrase technique contient donc une sémantique supplémentaire.

Les paires de phrases avec d'inclusion (SIT et TIS) permettent de repérer les cas lorsque les phrases sont segmentées (une phrase technique est scindée en plusieurs phrases simples) ou fusionnées (plusieurs phrases techniques sont fusionnées en une seule phrase dans la version simplifiée du texte).

Pour chaque sous-ensemble, nous prenons d'abord autant de paires équilibrées que d'exemples négatifs, que nous sélectionnons aléatoirement. Ensuite nous augmentons progressivement le nombre de paires non alignées jusqu'à un ratio de 3000 :1, ratio proche de celui des données réelles. Pour chaque ensemble déséquilibré D ainsi constitué, nous faisons deux expériences :

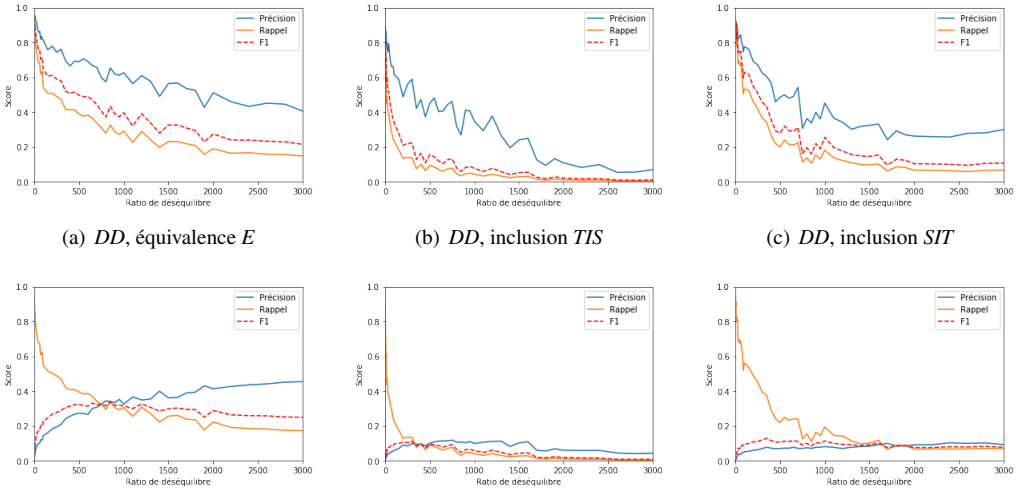
DD : Entraînement et test au sein de l'ensemble déséquilibré D ;
DR : Entraînement sur l'ensemble D et test sur les données réelles R. Notons que D est inclus dans R.

Pour l'évaluation, nous divisons les données en deux parties : deux tiers pour l'entraînement et un tiers pour le test. L'évaluation est effectuée en calculant le rappel, la précision et la F-mesure. Comme notre objectif vise la détection des paires alignées, les scores sont rapportés uniquement pour la classe des paires alignées. Une autre raison d'exclure les scores de paires non alignées est, qu'avec le déséquilibre qui augmente, cette classe négative obtient très rapidement des scores très elevés (>0,99).

Pour avoir une évaluation plus fiable, chaque expérience est effectuée cinquante fois et les résultats présentés correspondent donc aux valeurs moyennes de ces cinquante itérations. La différence d'une itération à l'autre est due au fait que l'ensemble des paires non alignées est différent à chaque fois car créé aléatoirement.

4 Présentation et discussion des résultats

FIGURE 1 – Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (*DD* et *DR*).



(a) *DD*, équivalence *E* (b) *DD*, inclusion *TIS* (c) *DD*, inclusion *SIT*

(d) *DR*, équivalence *E* (e) *DR*, inclusion *TIS* (f) *DR*, inclusion *SIT*

Nous présentons les résultats à la figure 1 : l'axe x représente l'augmentation du déséquilibre (seule la première position 1 correspond à des données équilibrées), alors que l'axe y représente les scores de précision, rappel et F-mesure. Les résultats pour les trois sous-ensembles sont présentés : équivalence (figures 1(a) et 1(d)), inclusion *TIS* (figures 1(b) et 1(e)) et inclusion *SIT* (figures 1(c) et 1(f)). La première ligne correspond aux résultats obtenus lorsque l'entraînement et le test sont effectués sur des données avec le même ratio de déséquilibre. La deuxième ligne correspond aux résultats obtenus par les mêmes modèles mais testés sur l'ensemble des données annotées manuellement.

Les paires équivalentes (figures 1(a) et 1(d)) sont plus faciles à catégoriser que les inclusions : d'une part, elles sont assez nombreuses et, d'autre part, elles doivent présenter des modèles de transformation plus stables. Les scores de précision et de rappel sont alors plus élevés à différents points de déséquilibre. Par exemple, au point 1 de la figure 1(a), la F-mesure est de 0,94 (0,96 de précision et 0,93 de rappel). Ce résultat est positif car les phrases équivalentes fournissent les informations les plus utiles et complètes pour décrire les transformations requises lors de la simplification. Avec les relations d'inclusion, au même point, nous avons 0,89 de F-mesure pour *TIS* (0,90 de précision et 0,89 de rappel) et 0,92 de F-mesure pour *SIT* (0,92 de précision et 0,93 de rappel). Nous supposons que des paires d'inclusion couvrent une grande variété de situations, ce qui est également plus difficile à modéliser. Nous prévoyons de faire des filtres supplémentaires pour mieux calibrer les résultats.

Nous voyons donc que les données équilibrées donnent de très bons résultats pour les trois jeux de données (*E, TIS* et *SIT*). En revanche, quand le déséquilibre est introduit, les performances sont réduites. Cela signifie que le déséquilibre crée de la confusion entre les paires alignables et non alignables. Cependant, le déséquilibre a un plus grand impact sur les paires qui relèvent de l'inclusion, que ce soit le sens *TIS* ou *SIT*. Une fois encore, il nous semble que cela est dû au fait que les cas d'inclusion sont beaucoup plus variés que les cas d'équivalence et sont en conséquence plus difficiles à circonscrire. Nos résultats indiquent que, quand on traite des données réelles, il vaut mieux effectuer la classification avec les modèles entraînés sur des données équilibrées. Un autre résultat intéressant est que la précision est plus élevée que le rappel. Cela s'observe particulièrement bien avec les expériences où l'entraînement et le test sont faits avec le même ratio de déséquilibre (figures 1(a), 1(b) et 1(c)). De manière générale, nous voyons que les résultats sont élevés lorsque l'on traite des données équilibrées. Cependant, comme le déséquilibre est une caractéristique naturelle des données que nous traitons, le travail à venir sera concentré sur la mise au point de descripteurs et filtres pour éliminer un maximum de phrases non alignables, *a priori* et/ou *a posteriori* de l'alignement.

5 Conclusion et perspectives

Nous proposons des expériences pour la détection et d'alignement de phrases parallèles dans des corpus comparables monolingues en français. L'aspect comparable se situe au niveau de la technicité des documents, qui met en regard des versions techniques et simplifiées des documents traitant du même sujet. Nous utilisons un corpus disponible CLEAR) qui relève du domaine biomédical. Plusieurs expériences sont menées : trois jeux de données (paires équivalentes et inclusion du sens d'une phrase dans l'autre) et données équilibrées et déséquilibrées. Sur les données équilibrées, nous atteignons une F-mesure de 0,94, avec un bon équilibre entre la précision et le rappel. Sur les données déséquilibrées, nous obtenons jusqu'à 0,89 et 0,92 de F-mesure. Les résultats restent meilleurs quand des modèles entraînés sur des données équilibrées sont utilisés. En l'état, cette méthode ne permet pas de générer un corpus de phrases parallèles de façon complètement automatique. Cependant, pour les paires équivalentes, le travail manuel est grandement allégé : nous atteignons un ratio d'environ 40 % de paires alignées correctement dans la sortie de notre classifieur sur de nouvelles données, alors que ce ratio est d'environ 0,025 % dans les données brutes (environ 4 000 paires non alignées pour 1 paire alignée). À l'avenir, nous prévoyons d'exploiter les meilleurs modèles générés pour enrichir le corpus de phrases parallèles. Les scores de rappel peuvent correspondre aux mesures de référence pour choisir le meilleur classifieur. Une attention particulière sera apportée au filtrage des données *a priori* et/ou *a posteriori* de l'alignement. D'autres pistes de travail concernent l'exploitation d'autres descripteurs et méthodes pour l'alignement de phrases. Comme la description du corpus le montre, la distance lexicale entre les phrases techniques et simplifiées est assez élevée. En conséquence, nous comptons nous tourner vers l'utilisation de plongements lexicaux ainsi que vers l'exploitation de connaissances externes pour pallier cette difficulté.

Remerciements

La présente publication s'inscrit dans le projet *CLEAR (Communication, Literacy, Education, Accessibility, Readability)* financé par l'ANR sous la référence ANR-17-CE19-0016-01. Nous remercions les relecteurs pour leurs remarques constructives.

Références

ABDUL-RAUF S. & SCHWENK H. (2009). On the use of comparable corpora to improve SMT performance. In *European Chapter of the ACL*, p. 16–23.

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. & WIEBE J. (2015). SemEval-2015 task 2 : Semantic textual similarity, english, spanish and pilot on interpretability. In *SemEval 2015*, p. 252–263.

AGIRRE E., BANEA C., CER D., DIAB M., GONZALEZ-AGIRRE A., MIHALCEA R., RIGAU G. & WIEBE J. (2016). SemEval-2016 task 1 : Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval 2016*, p. 497–511.

AGIRRE E., CER D., DIAB M., GONZALEZ-AGIRRE A. & GUO W. (2013). *sem 2013 shared task : Semantic textual similarity. In **SEM*, p. 32–43.

AMA (1999). Health literacy : report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. *JAMA*, **281**(6), 552–7.

BARZILAY R. & ELHADAD N. (2003). Sentence alignment for monolingual comparable corpora. In *EMNLP*, p. 25–32.

CARDON R. & GRABAR N. (2018). Identification of parallel sentences in comparable monolingual corpora from different registers. In *LOUHI 2018*, p. 1–11.

CHEN P., ROCHFORD J., KENNEDY D. N., DJAMASBI S., FAY P. & SCOTT W. (2016). Automatic text simplification for people with intellectual disabilities. In *AIST*, p. 1–9.

CLOUGH P., GAIZAUSKAS R., PIAO S. S. & WILKS Y. (2002). METER : Measuring text reuse. In *ACL*, p. 152–159.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

COVER T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, **14**(3), 326–334.

FERNANDO S. & STEVENSON M. (2008). A semantic similarity approach to paraphrase detection. In *Comp Ling UK*, p. 1–7.

FISHER R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**(2), 179–188.

GANITKEVITCH J., VAN DURME B. & CALLISON-BURCH C. (2013). PPDB : The paraphrase database. In *NAACL-HLT*, p. 758–764.

GRABAR N. & CARDON R. (2018). CLEAR – Simple Corpus for Medical French. In *Workshop on Automatic Text Adaption (ATA)*, p. 1–11.

GUO W. & DIAB M. (2012). Modeling sentences in the latent space. In *ACL*, p. 864–872.

HE H., GIMPEL K. & LIN J. (2015). Multi-perspective sentence similarity modeling with convolutional neural networks. In *EMNLP*, p. 1576–1586, Lisbon, Portugal.

KIROS R., ZHU Y., SALAKHUTDINOV R., ZEMEL R. S., TORRALBA A., URTASUN R. & FIDLER S. (2015). Skip-thought vectors. In *Neural Information Processing Systems (NIPS)*, p. 3294–3302.

KOEHN P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings : the tenth Machine Translation Summit*, p. 79–86, Phuket, Thailand : AAMT AAMT.

KUMANO T., TANAKA H. & TOKUNAGA T. (2007). Extracting phrasal alignments from comparable corpora by using joint probability SMT model. In *Int Conf on Theoretical and Methodological Issues in Machine Translation*.

LAI A. & HOCKENMAIER J. (2014). Illinois-LH : A denotational and distributional approach to semantics. In *Workshop on Semantic Evaluation (SemEval 2014)*, p. 239–334, Dublin, Ireland.

LEROY G., KAUCHAK D. & MOURADI O. (2013). A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *Int J Med Inform*, **82**(8), 717–730.

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

MADNANI N., TETREAULT J. & CHODOROW M. (2012). Re-examining machine translation metrics for paraphrase identification. In *NAACL-HLT*, p. 182–190.

MCGRAY A. (2005). Promoting health literacy. *J of Am Med Infor Ass*, **12**, 152–163.

MIHALCEA R., CORLEY C. & STRAPPARAVA C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, p. 1–6.

MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D. & MILLER K. (1993). *Introduction to WordNet : An On-line Lexical Database*. Rapport interne, WordNet.

MUELLER J. & THYAGARAJAN A. (2016). Siamese recurrent architectures for learning sentence similarity. In *AAAI Conference on Artificial Intelligence*, p. 2786–2792.

MUNTEANU D. S. & MARCU D. (2002). Processing comparable corpora with bilingual suffix trees. In *EMNLP*, p. 289–295.

MUNTEANU D. S. & MARCU D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, **31**(4), 477–504.

MUNTEANU D. S. & MARCU D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *COLING-ACL*, p. 81–88.

NELKEN R. & SHIEBER S. M. (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. In *EACL*, p. 161–168.

PAETZOLD G. H. & SPECIA L. (2016). Benchmarking lexical simplification systems. In *LREC*, p. 3074–3080.

QIU L., KAN M.-Y. & CHUA T.-S. (2006). Paraphrase recognition via dissimilarity significance classification. In *Empirical Methods in Natural Language Processing*, p. 18–26, Sydney, Australia.

ROSENBLATT F. (1958). The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**(6), 386–408.

ROSENBLATT F. (1961). *Principles of Neurodynamics : Perceptrons and the Theory of Brain Mechanisms*. Washington DC : Spartan Books.

RUDD E. (2013). Needed action in health literacy. *J Health Psychol*, **18**(8), 1004–10.

SEVERYN A., NICOSIA M. & MOSCHITTI A. (2013). Learning semantic textual similarity with structural representations. In *Annual Meeting of the Association for Computational Linguistics*, p. 714–718.

TAI K. S., SOCHER R. & MANNING C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Annual Meeting of the Association for Computational Linguistics*, p. 1556–1566, Beijing, China.

TSUBAKI M., DUH K., SHIMBO M. & MATSUMOTO Y. (2016). Non-linear similarity learning for compositionality. In *AAAI Conference on Artificial Intelligence*, p. 2828–2834.

UTIYAMA M. & ISAHARA H. (2003). Reliable measures for aligning Japanese-English news articles and sentences. In *Annual Meeting of the Association for Computational Linguistics*, p. 72–79.

VAPNIK V. & LERNER A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, **24**, 709–715.

VU T. T., TRAN G. B. & PHAM S. B. (2014). Learning to simplify children stories with limited data. In L. . SPRINGER, Ed., *Intelligent Information and Database Systems*, p. 31–41.

WAN S., DRAS M., DALE R. & PARIS C. (2006). Using dependency-based features to take the "para-farce" out of paraphrase. In *Australasian Language Technology Workshop*, p. 131–138.

YANG C. C. & LI K. W. (2003). Automatic construction of English/Chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, **54**(8), 730–742.

ZHANG Y. & PATRICK J. (2005). Paraphrase identification by text canonicalization. In *Australasian Language Technology Workshop*, p. 160–166.

ZHAO J., ZHU T. T. & LAN M. (2014). ECNU : One stone two birds : Ensemble of heterogenous measures for semantic relatedness and textual entailment. In *Workshop on Semantic Evaluation (SemEval 2014)*, p. 271–277.

ZHU Z., BERNHARD D. & GUREVYCH I. (2010). A monolingual tree-based translation model for sentence simplification. In *COLING 2010*, p. 1353–1361.