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)
Nancy, France, 08-19 juin 2020

Approche supervisée de calcul de similarité sémantique entre paires de phrases

Khadim Dramé[1, 2] Gorgoumack Sambe[1, 2] Ibrahima Diop[1, 2] Lamine Faty[1, 2]
(1) Université Assane Seck de Ziguinchor, Diabir, Ziguinchor, Sénégal
(2) Laboratoire d'Informatique et d'Ingénierie pour l'Innovation, Ziguinchor, Sénégal
khadim.drame@univ-zig.sn, gsambe@univ-zig.sn,
ibrahima.diop@univ-zig.sn, lamine.faty@univ-zig.sn

RÉSUMÉ

Ce papier décrit les méthodes que nous avons développées pour participer aux tâches 1 et 2 de l'édition 2020 du défi fouille de textes (DEFT 2020). Pour la première tâche, qui s'intéresse au calcul de scores de similarité sémantique entre paires de phrases, sur une échelle de 0 à 5, une approche supervisée où chaque paire de phrases est représentée par un ensemble d'attributs a été proposée. Des algorithmes classiques d'apprentissage automatique sont ensuite utilisés pour entrainer les modèles. Différentes mesures de similarité textuelle sont explorées et les plus pertinentes sont combinées pour supporter nos méthodes. Différentes combinaisons ont été testées et évaluées sur les données de test du DEFT 2020. Notre meilleur système qui s'appuie sur un modèle Random Forest a obtenu les meilleures performances sur la première tâche avec une EDRM de 0,8216.

ABSTRACT

Supervised approach to compute semantic similarity between sentence pairs.

In this paper, we present the methods that we developed to participate in tasks 1 and 2 of the 2020 edition of the french text mining challenge (DEFT 2020). For the first task, which focuses on semantic similarity computation between sentence pairs, a supervised approach where sentence pairs are first represented by a set of features has been proposed. Classical machine learning algorithms are then used to train the models. Different measures of textual similarity are explored and the most relevant are combined to support our methods. Different combinations were tested and evaluated on test data of the DEFT 2020. Our best system based on a Random Forest model performed best on the first task with an EDRM of 0,8216.

MOTS-CLÉS : similarité sémantique, phrases parallèles, méthodes supervisés, apprentissage automatique, forêts aléatoires, perceptron multicouche.

KEYWORDS: semantic similarity, parallel sentences, supervised methods, machine learning, Random Forest, Multilayer Perceptron.

1 Introduction

Le défi fouille de textes (DEFT) est une campagne d'évaluation visant à promouvoir le développement de méthodes et d'applications dans le domaine du traitement automatique de langues naturelles (TALN). Dans son édition de 2020, il continue à s'intéresser à l'analyse de cas

Atelier DÉfi Fouille de Textes, pages 49–54.
HAL : hal-02784741.

cliniques rédigés en Français; les tâches 1 et 2 portent sur la similarité sémantique entre phrases tandis que la troisième tâche s'intéresse à l'extraction d'information fine à partir de textes biomédicaux (Cardon et al., 2020).

La tâche 1 du défi consiste à déterminer le degré de similarité entre paires de phrases, sur une échelle de 0 à 5. La tâche 2, quant à elle, consiste, pour une phrase source donnée, à identifier à partir de phrases cibles fournies celle qui est parallèle à cette dernière. Ces questions se rapportent au calcul de similarité sémantique entre phrases.

Dans la littérature, ce problème a été largement exploré et différentes mesures de similarité sont proposées (Agirre et al., Agirre et al.,2015; 2016; Cer et al., 2017). Certaines approches couramment utilisées exploitent la structure syntaxique des phrases ; le nombre de tokens ou n-grams en commun entre la phrase source et la phrase cible est généralement calculé. D'autres tentent de prendre en compte les problèmes de synonymie et la sémantique des phrases en exploitant des ressources sémantiques ou des méthodes statistiques (Chen et al., 2020). Dans les dernières campagnes d'évaluation, les méthodes supervisées se sont montrées performantes pour mesurer la similarité sémantique entre phrases (Cer et al., 2017; Mojarad et al., 2018).

Les méthodes que nous proposons s'inscrivent dans cette dernière approche et utilisent les algorithmes classiques d'apprentissage automatique pour déterminer les scores de similarité entre les paires de phrases. Ce papier décrit les méthodes développées pour participer aux deux premières tâches du DEFT 2020. Le reste du papier est structuré comme suit : les méthodes de calcul de similarité sémantique sont présentées dans la section 2; les résultats obtenus sont décrits et discutés respectivement dans les sections 3 et 4.

2 Méthodes de calcul de similarité entre phrases

Dans cette section, nous décrivons les trois méthodes que nous avons développées pour le calcul de similarité entre paires de phrases.

Nous proposons une approche supervisée où chaque paire de phrases est représentée par un ensemble d'attributs. Différentes mesures de similarité sémantique sont explorées : les mesures de similarité basées sur les tokens (mesure de Dice (Dice, 1945), mesure de Ochiai (OCHIAI, 1957), mesure de Jaccard (Jaccard, 1912)), les mesures utilisant les séquences de caractères (Q-gram (Ukkonen, 1992)), la distance d'édition de Levenshtein (Levenshtein, 1966), les mesures basées sur la représentation vectorielle (TF.IDF (Jones, 2004) et les plongements lexicaux (Mikolov et al., 2013) combinées avec le cosinus).

Tout d'abord, chaque paire de phrases, qui est une instance, est représentée par un ensemble d'attributs, constitués par les scores des mesures de similarité citées ci-dessus. Ensuite, des algorithmes classiques d'apprentissage automatique sont utilisés pour entrainer les modèles, qui sont ensuite utilisés pour déterminer la similarité entre des paires de phrases non annotées.

Différents algorithmes d'apprentissage sont expérimentés mais seuls les résultats des modèles Random Forest (RF) et MultiLayer Perceptron (MLP) qui ont donné les meilleures performances sur les jeux de données d'entrainement sont soumis. De plus, nous avons développé un modèle de régression linéaire (LR) qui prend en entrée les scores de similarité de ces deux modèles et le score moyen des différentes mesures de similarité.

Une validation croisée est effectuée sur les jeux de données d'entrainement pour sélectionner les attributs les plus pertinents mais aussi pour déterminer les meilleures combinaisons. La combinaison de quatre mesures de similarité sémantique (Dice, Ochiai, Q-gram, Levenshtein) a donné les meilleures performances.

3 Evaluation

Dans cette section, nous allons d'abord présenter les jeux de données et les métriques utilisées pour évaluer les systèmes participants au DEFT 2019. Ensuite, les résultats de nos méthodes seront analysés et discutés.

3.1 Jeux de données

Pour chaque tâche, les organisateurs du DEFT 2020 ont fourni des jeux de données annotées (Grabar, Claveau & Dalloux, 2018; Grabar & Cardon, 2018). Pour la première tâche, un corpus d'entrainement constitué de 600 paires de phrases a été fourni avec, pour chaque paire, son score de similarité. Chaque paire de phrases est annotée manuellement avec un score indiquant le degré de similarité des phrases. Les données sont annotées indépendamment par deux experts qui attribuent des scores de similarité entre les paires de phrases allant de 0 (complètement différentes) à 5 (sémantiquement équivalentes). Ensuite, les annotations de référence font l'objet d'un accord entre les deux annotateurs. Le corpus de test est quant à lui constitué de 410 paires de phrases.

3.2 Mesures d'évaluation

L'exactitude en distance relative à la solution moyenne (EDRM) et la corrélation de Spearman sont utilisées pour mesurer les performances des systèmes participants à la tâche 1. Pour la deuxième tâche, la MAP (Mean Average Precision) est utilisée pour évaluer les résultats.

3.3 Résultats

Les résultats de nos différents systèmes participants à la tâche 1 sur les jeux de données de test officiels sont présentés dans TABLE 1. Nous remarquons que le système *uasz-run2*, utilisant le modèle perceptron multicouche (MLP), a obtenu des résultats largement meilleurs selon l'exactitude en distance relative à la solution moyenne. Notons également que le système *uasz-run1*, qui utilise le modèle Random Forest (RF), est plus performant que *uasz-run3*, qui lui combine les scores de similarité de ces deux modèles dans un modèle de régression linéaire. Nous avons également expérimenté une approche supervisée combinant les différentes mesures de similarité avec différents classifieurs (Naive Bayes, MultiLayer Perceptron, Support Vector Machine et Random Forest) mais les systèmes présentés ont obtenu les meilleurs résultats sur les jeux de données d'entrainement.

Systèmes	EDRM	Corrélation de Spearman
uasz-run1	0,7946	0,7527
uasz-run2	**0,8216**	0,7691
uasz-run3	0,7755	**0,7768**

TABLE 1 : Résultats de nos systèmes participants à la tâche 1 du DEFT 2020 sur les jeux de données de test officiels

Comparé aux différents systèmes participants à la tâche 1, *uasz-run2*, notre meilleur système, a obtenu les meilleurs résultats (avec une EDRM de 0,8216). Nous notons aussi que tous nos systèmes ont obtenu une EDRM dépassant la moyenne (0,762). Nos deux meilleurs systèmes, *uasz-run2* et *uasz-run1* ont également obtenu une EDRM supérieure (*uasz-run2*) ou égale (*uasz-run1*) à la médiane (0,795).

4 Discussion

L'évaluation de nos différentes méthodes de calcul de similarité sémantique sur les jeux de données du DEFT 2020 montre la bonne performance des algorithmes classiques d'apprentissage automatique pour cette tâche. Les résultats montrent également la pertinence des mesures de similarité utilisées pour capturer la similarité sémantique entre phrases. Les méthodes développées permettent toutes d'estimer correctement la similarité de la plupart des paires de phrases du corpus de test. Une analyse des résultats a permis de relever des limites de ces méthodes pour la prédiction des scores de similarités pour certaines paires de phrases. Les mesures de similarité utilisées (Dice, Ochiai, Q-gram, Levenshtein) ne prennent pas en compte la dimension sémantique des phrases ; par conséquent, nos méthodes peinent à prédire correctement les scores de similarité des phrases qui ont des structures syntaxiques similaires mais sont sémantiquement différentes. Par exemple, pour la paire de phrases 22 (id=22) du corpus de test, toutes les trois méthodes ont estimé les deux phrases similaires avec un score de similarité de 4 tandis que le degré de similarité fourni par les experts est de 1. De manière analogue, nos méthodes sont limitées pour la prédiction de la similarité des phrases sémantiquement proches mais utilisant des mots différents. Par exemple, les phrases de la paire 52 (id=52) sont considérées comme différentes avec un score de similarité de 0 tandis qu'elles sont similaires selon les experts (avec un degré de similarité de 4).

Nous remarquons également que les méthodes proposées, et particulièrement *uasz-run1* et *uasz-run2* utilisant respectivement le modèle Random Forest et le perceptron multicouche, peinent à prédire les classes les moins représentatives (1 et 2) dans le corpus d'entrainement. Dans le jeu de données officiel de test, les classes 1 et 2 sont respectivement 37 et 28. *uasz-run1* ne prédit aucune valeur de ces deux classes tandis que *uasz-run2* prédit seulement 9 valeurs dans la classe 1.

5 Conclusion

Dans ce papier, nous avons présenté les méthodes que notre équipe a développées pour participer aux tâches 1 et 2 du DEFT 2020. Trois méthodes de calcul de similarité sémantique entre phrases ont été proposées : une méthode utilisant le modèle Random Forest (RF), une autre utilisant le perceptron multicouche (MLP) et une dernière combinant les résultats de ces deux modèles. Les résultats officiels du DEFT 2020 montrent que notre méthode basée sur le perceptron multicouche a

obtenu les meilleures performances sur la tâche 1. Comparés aux différents systèmes participants, les deux autres méthodes ont donné des résultats encourageants. Nous envisageons d'exploiter d'autres mesures de similarité notamment celles permettant de capturer la sémantique des phrases afin d'améliorer les performances ; une première expérimentation avec les plongements lexicaux sur un corpus moyen n'a pas permis d'améliorer les résultats. Leur utilisation sur un corpus plus conséquent pourrait permettre d'augmenter les performances des systèmes.

Remerciements

Nous remercions les organisateurs du DEFT 2020.

Références

AGIRRE E., BANEA C., CARDIE C., CER D., DIAB M., AGIRRE A. G. & GUO W. (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)*, 252–263. Denver, Colorado: Association for Computational Linguistics. https://doi.org/10.18653/v1/S15-2045.

AGIRRE E., BANEA C., CER D., DIAB M., AGIRRE A. G., MIHALCEA R., RIGAU G. & 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)*, 497–511. San Diego, California: Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1081.

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 DEFT*.

CER D., DIAB M., AGIRRE E., GAZPIO I. L. & 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)*, 1–14. Vancouver, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-2001.

CHEN Q., DU J., KIM S., WILBUR W. J. & LU Z. (2020). Deep Learning with Sentence Embeddings Pre-Trained on Biomedical Corpora Improves the Performance of Finding Similar Sentences in Electronic Medical Records. *BMC Medical Informatics and Decision Making* 20 (1): 73. https://doi.org/10.1186/s12911-020-1044-0.

DICE L. R. (1945). "Measures of the Amount of Ecologic Association Between Species." *Ecology* 26 (3): 297–302. https://doi.org/10.2307/1932409.

GRABAR N. & CARDON R. (2018). CLEAR – Simple Corpus for Medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, 3–9. Tilburg, the Netherlands: Association for Computational Linguistics. https://doi.org/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*, 122–128. Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5614.

JACCARD P. (1912). The Distribution of the Flora in the Alpine Zone.1. *New Phytologist* 11 (2): 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

LEVENSHTEIN V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *SPhD* 10 (February): 707.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *ArXiv:1310.4546 [Cs, Stat]*, October. http://arxiv.org/abs/1310.4546.
OCHIAI A. (1957). Zoogeographical Studies on the Soleoid Fishes Found in Japan and Its Neighhouring Regions-II. *NIPPON SUISAN GAKKAISHI* 22 (9): 526–30. https://doi.org/10.2331/suisan.22.526.
MOJARAD M. R., LIU S., WANG Y., AFZAL N., WANG L., SHEN F., FU S. & LIU H. (2018). BioCreative/OHNLP Challenge 2018. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 575. BCB '18. Washington, DC, USA: Association for Computing Machinery. https://doi.org/10.1145/3233547.3233672.
JONES K. S. (2004). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation* 60 (5): 493–502. https://doi.org/10.1108/00220410410560573.
UKKONEN E. (1992). Approximate String-Matching with q-Grams and Maximal Matches. *Theoretical Computer Science* 92 (1): 191–211. https://doi.org/10.1016/0304-3975(92)90143-4.