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

Similarité sémantique entre phrases : apprentissage par transfert interlingue

Charles Teissèdre Thiziri Belkacem Maxime Arens
Synapse Développement, 7 Boulevard de la Gare – 31500 Toulouse
{thiziri.belkacem, maxime.arens, charles.teissedre}@synapse-fr.com

RÉSUMÉ

Dans cet article, nous décrivons une approche exploratoire pour entraîner des modèles de langue et résoudre des tâches d'appariement entre phrases issues de corpus en français et relevant du domaine médical. Nous montrons que, dans un contexte où les données d'entraînement sont en nombre restreint, il peut être intéressant d'opérer un apprentissage par transfert, d'une langue dont nous disposons de plus de ressources pour l'entraînement, vers une langue cible moins dotée de données d'entraînement (le français dans notre cas). Les résultats de nos expérimentations montrent que les modèles de langue multilingues sont capables de transférer des représentations d'une langue à l'autre de façon efficace pour résoudre des tâches de similarité sémantique telles que celles proposées dans le cadre de l'édition 2020 du Défi fouille de texte (DEFT).

ABSTRACT

Semantic Sentence Similarity : Multilingual Transfer Learning

In this paper, we describe an exploratory approach to train language models and solve sentence-matching tasks from French corpora in the medical field. We show that, in a context where training data are limited, it may be interesting to transfer learning from a language with more training resources to a target language with less training data (French in our case). The results of our experiments show that multilingual language models are able to transfer representations from one language to another efficiently to solve semantic similarity, tasks such as those proposed in the 2020 edition of the Text Mining Challenge (DEFT).

MOTS-CLÉS : Similarité Sémantique Textuelle, Modèles Neuronaux Multilingues, Apprentissage par transfert Interlingue.

KEYWORDS: Semantic Textual Similarity, Multilingual Neural Models, Cross-lingual Transfer Learning.

1 Introduction et motivation

Mesurer la similarité entre des textes est une tâche importante dans plusieurs applications du traitement des langues et de la recherche d'information (Baziz *et al.*, 2005; Manning *et al.*, 2010; Guo *et al.*, 2016; Kusner *et al.*, 2015). Dans cet article, nous présentons des travaux expérimentaux comparant différents modèles de calcul de similarité entre phrases, pour résoudre deux tâches proposées dans le cadre de l'édition 2020 du défi fouille de textes, DEFT 2020 [1] (Cardon *et al.*, 2020). La première

1. https://deft.limsi.fr/2020/

Atelier DÉfi Fouille de Textes, pages 97–107.
HAL : hal-02784746.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

tâche consiste à mesurer la similarité entre des paires de phrases sur une échelle allant de 0 à 5. Elle renvoie à une tâche d'appariement désormais classique proposée dans les campagnes SemEval de 2012 à 2017. La seconde tâche consiste à sélectionner une phrase cible similaire à une phrase source dans un ensemble de phrases candidates. Les corpus d'entraînement fournis pour ces deux tâches d'appariement entre phrases relèvent du domaine médical.

Les difficultés principales de ces deux tâches consiste dans le fait que l'on souhaite manipuler des représentations de phrases et que l'on dispose de données de faible volumétrie pour l'entraînement, en particulier pour la tâche 1 (le corpus d'entraînement contient 600 paires de phrases uniquement).

L'équipe de Synapse souhaitait à travers ce défi tester la capacité des modèles de langue multilingues, ré-entraînés avec peu de données dans une langue cible, à résoudre des problèmes d'appariement entre phrases, un problème central dans les systèmes de Question-Réponse et de recherche d'information couvrant plusieurs langues, dans un contexte où l'on dispose de peu voire même d'aucune donnée d'entraînement. L'enjeu sur un plan industriel est en effet de favoriser le développement d'applications multilingues, sans qu'il soit nécessaire de disposer à l'initialisation des systèmes, de ressources ou de données dans chacune des langues à traiter. Pour compenser le manque de données d'entraînement en français, nous avons ainsi utilisé un jeu de données de plus grande volumétrie en anglais provenant du benchmark STS [2] (STSBenchmark). Dans nos expérimentations, nous avons testé différents modèles de langue multilingues spécialisés pour générer des représentations de phrases. Nous montrons que ces modèles entraînés à partir des ressources disponibles en anglais sont en mesure de transférer des représentations latentes d'une langue à l'autre, et ainsi d'apprendre à résoudre la tâche en français. Afin d'en évaluer l'intérêt, nous comparons les résultats de cette approche à ceux d'approches concurrentes, obtenus par des méthodes d'apprentissage supervisées et des modèles d'apprentissage monolingues.

2 Les corpus et les modèles testés

2.1 Les corpus d'entraînement

Les corpus d'entraînement fournis pour les tâches 1 et 2 proviennent du projet CLEAR (Grabar & Cardon, 2018) et comprennent [3] des articles d'encyclopédie, des notices de médicaments et des résumés Cochrane, dont le contenu présente de grandes similarités d'un sous corpus à l'autre, ce qui permet de constituer des paires de phrases parallèles. Il s'agit ainsi de corpus relevant du domaine médical. Les annotations de référence ont été normalisées par consensus, après une double annotation indépendante.

La tâche 1 correspond à une transposition, sur des données en français, de la tâche de Similarité Sémantique Textuelle (Semantic Textual Similarity) telle que définie dans les campagnes SemEval (2012-2017) [4]. Étant donnés des couples de phrases, la tâche 1 invite les systèmes participants à retourner un score de similarité sur une échelle de valeurs graduées de 0 (phrases sémantiquement indépendantes) à 5 (phrases équivalentes). L'évaluation des différents systèmes mesure l'écart entre la valeur fournie et la valeur de référence. La figure 1 présente un extrait du corpus fourni pour la tâche 1. Le corpus d'entraînement associé à cette tâche contient 600 paires de phrases.

2. https://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark
3. https://deft.limsi.fr/2020/
4. http://alt.qcri.org/semeval2017/task1/

```
<paire id="2" vote="5">
<source>- En l'absence d'amélioration comme en cas de persistance des
symptômes, prendre un avis médical.
</source>
<cible>En l'absence d'amélioration comme en cas de persistance des
symptômes, prenez un avis médical.
</cible>
</paire>
```

FIGURE 1 – Exemple de données du corpus de la tâche 1.

```
<ensemble id="1" cible="2">
  <source>
    compte tenu des données disponibles , l' utilisation chez la femme
enceinte ou qui allaite est possible ponctuellement
  </source>
  <cible num="1">ce médicament est un laxatif utilisé par voie
orale</cible>
  <cible num="2">ce médicament , dans les conditions normales d'
utilisation , peut être utilisé ponctuellement pendant la grossesse et
l' allaitement</cible>
  <cible num="3">boîte de 1 flacon de 250 ml ou 500 ml</cible>
</ensemble>
```

FIGURE 2 – Exemple de données du corpus de la tâche 2.

La tâche 2 consiste à identifier les phrases parallèles d'une phrase source parmi un ensemble de phrases cibles. Un extrait du corpus associé à cette tâche est présenté dans la figure 2. Dans l'exemple illustré, la phrase cible numéro 2 est une phrase parallèle à la phrase source. Le corpus d'entraînement associé à cette tâche contient un peu plus de 1700 ensembles de phrases, chaque ensemble comprenant une phrase source et trois phrases cibles.

2.2 Les modèles testés

Dans cette section, nous décrivons brièvement les différents modèles utilisés pour résoudre les tâches auxquelles nous participions, ainsi que les motivations qui nous ont conduits à tester ces différents modèles. Pour les deux tâches, nous avons utilisé des méthodes d'apprentissage supervisé devant servir de modèles de référence (baseline), puis testé différents modèles de langue multilingues.

2.2.1 Des approches supervisées comme modèles de référence

Pour la tâche 1, le modèle supervisé utilisé comme méthode de référence est un modèle de similarité proposé par (Guo *et al.*, 2016), basé sur la pertinence (Deep Relevance Matching Model for ad-hoc retrieval ou DRMM). Les auteurs montrent que les méthodes d'apprentissage profond conçues pour l'appariement sémantique ne seraient pas bien adaptées à la recherche ad-hoc. Cette dernière concerne essentiellement l'appariement par pertinence plutôt que l'appariement sémantique. Basé sur cette différence, le modèle DRMM calcule les interactions mot-mot entre les séquences d'entrée représentées dans l'espace distribué (embedding), où chaque mot est représenté par un vecteur. DRMM utilise une similarité cosinus et calcule une matrice M qui est ensuite transformée en histogrammes [5] d'interaction qui sont calculés en utilisant les valeurs de similarité entre tout les termes des séquences de texte en entrées.

Les modèles supervisés qui nous ont servis de modèles de référence pour la tâche 2 proviennent de précédentes expériences sur un corpus maison en français contenant des paires de questions parallèles. Ces modèles présentent de bons résultats sur ce corpus permettant de capter efficacement la similarité sémantique entre des questions, un problème très similaire à celui de la tâche 2, pour laquelle il faut retrouver une phrase parallèle à la phrase source dans un ensemble de phrases cibles. Pour entraîner ces modèles d'apprentissage supervisé, nous avons utilisé un ensemble de caractéristiques (features) liées aux différentes séquences d'entrée. Ces caractéristiques sont ensuite combinées dans un modèle de classification entraîné en utilisant un algorithme de descente de gradient (SGD) (Bottou, 2010) et la méthode de gradient boost (XGB) (Chen & Guestrin, 2016).

2.2.2 Modèles de langue multilingues

Les modèles de langue testés sur la tâche 1 sont des dérivés du modèle BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al.*, 2018), dont une des principales innovations techniques est l'application aux modèles de langue d'un entraînement bidirectionnel des tranformateurs (Dehghani *et al.*, 2018), modèles neuronaux basés sur le mécanisme d'attention. L'entraînement bidirectionnel de BERT exploite le contexte de gauche à droite et de droite à gauche. Les résultats de BERT sur un grand nombre de tâches classiques du TAL montrent qu'un modèle de langue entraîné de manière bidirectionnelle peut avoir une perception plus profonde du contexte et du flux linguistique, comparé notamment aux modèles de langue basés sur un contexte unidirectionnel (Radford *et al.*, 2018).

BERT multilingue (M-BERT) est le pendant multilingue de BERT (Devlin *et al.*, 2018) pré-entraîné à partir de corpus monolingues dans 104 langues différentes. Dans leurs expérimentations, (Pires *et al.*, 2019) ont montré que M-BERT est efficace dans les applications d'apprentissage par transfert appliquées aux traitements des langues. Ce modèle est notamment performant dans des applications multilingues dites zéro-shot (zero-shot transfer), où seules les représentations spécifiques à une langue sont utilisées pour affiner le modèle dans une autre langue. Pré-entraîné sur de grands corpus de textes non annotés et facilement disponibles, le modèle BERT est affiné pour des tâches spécifiques sur de plus petites quantités de données qualifiées, en s'appuyant sur la structure du modèle induit pour faciliter la généralisation au-delà des données d'entraînement. Dans (Pires *et al.*, 2019), les

5. Cette méthode sert à traduire des vecteurs de différentes dimensions dont les valeurs sont dans l'intervalle $[-1; 1]$ à un ensemble de vecteurs de même dimensions et dont les valeurs sont des entiers, en se basant sur une taille prédéfinie des intervalles de valeurs.

auteurs montrent que le transfert inter-linguistique opère d'autant mieux lorsqu'il y a un important chevauchement lexical entre les langues et qu'elles sont typologiquement proches, par exemple, entre des langues de type sujet-verbe-objet (SVO) telles que l'anglais et le français, l'espagnol ou le bulgare.

Dérivé de M-BERT et de Sentence-BERT (Reimers & Gurevych, 2019), le modèle Sentence Multilingual BERT [6] (Sentence M-BERT) que nous avons testé (101 langues) est un encoder de phrases initialisé avec M-BERT et affiné sur le corpus anglais MultiNLI (Conneau *et al.*, 2018) et sur le corpus de développement multilingue XNLI (Williams *et al.*, 2018). Les représentations de phrases sont des moyennes de vecteurs correspondant aux différents symboles (tokens) des phrases d'entrée.

Multilingual Universal Sentence Encoder [7] (MUSE) (Chidambaram *et al.*, 2018) est un modèle de représentation de phrases basé sur la traduction pour traiter plusieurs langues différentes. La traduction est effectuée sur la base d'une tâche de tri, de sorte que les réponses codées obtiennent des représentations très similaires à celles des questions correspondantes à l'aide d'une fonction de produit. MUSE utilise une seule couche d'encodage et est entraîné dans un cadre multi-tâches, où des couches supplémentaires spécifiques à la tâche en cours de traitement sont utilisées en plus de l'encodeur unique. Dans leur travail (Chidambaram *et al.*, 2018), les auteurs ont proposé en fonction des usages plusieurs modèles multilingues qui couvrent 16 langues (dont l'anglais, le français, l'espagnol et l'allemand) dans un unique espace sémantique. Ces modèles obtiennent des performances compétitives avec l'état de l'art sur différentes tâches de traitement automatique des langues (classification de textes, regroupement de textes, recherche de similarités sémantiques).

Nous avons exploités ces différents modèles multilingues dans nos expérimentations, parce qu'ils paraissaient adaptés aux tâches 1 et 2, où il s'agit de mesurer la similarité entre des couples de phrases et sachant que nous souhaitions opérer un apprentissage par transfert de langue.

3 Expérimentations et résultats

Dans cette section, nous décrivons la démarche expérimentale que nous avons suivie pour résoudre les tâches d'appariement entre phrases, ainsi que les résultats de l'évaluation des différents modèles testés.

3.1 Tâche 1 : Similarité sémantique entre phrases

Conformément à la méthodologie d'évaluation proposée dans SemEval (Cer *et al.*, 2017), nous avons évalués et comparés les modèles testés lors de nos expérimentations au moyen de la mesure de Pearson, qui permet d'établir une corrélation entre les valeurs de similarité correctes (de référence) et celles obtenues de façon automatique par les différents modèles entraînés.

La difficulté principale de la tâche 1 est que les données d'entraînement sont peu nombreuses : le jeu de données fourni par les organisateurs comprend 600 couples de phrases pour 6 classes différentes, de 0 à 5. Les prédictions obtenues à partir du modèle de langue française CamemBERT (Martin *et al.*, 2020) sur les données fournies pour DEFT 2020 ont montré des résultats intéressants (0.77 sur le corpus de développement sans exploiter les données d'entraînement), mais nous souhaitions pouvoir exploiter

6. http://files.deeppavlov.ai/deeppavlov_data/bert/sentence_multi_cased_L-12_H-768_A-12.tar.gz
7. https://tfhub.dev/google/universal-sentence-encoder-multilingual/3

Synthèse des résultats - Tâche 1		
Modèle	Corpus Dev	Corpus Test
DRMM	0.55	-
MUSE	0.77	0.73
M-BERT + STS	0.82	0.74
Sentence M-BERT + STS	0.83	0.76

TABLE 1 – Performances, en termes de corrélation de Pearson, calculée pour les différents modèles concernant les prédictions pour la tâche 1, avec et sans entraînement sur les données STS.

des jeux de données plus volumineux, bien que n'étant pas disponibles en français, en particulier le corpus associé au STSBenchmark. D'où l'idée d'exploiter des modèles de langue multilingues et de les affiner avec ce jeu de données, leur permettant ainsi de générer des représentations de phrases et d'évaluer ensuite leurs performances sur des données en français.

Comme le montrent (Reimers & Gurevych, 2019), les modèles BERT (Devlin *et al.*, 2018) / RoBERTa (Liu *et al.*, 2019) / XLM-RoBERTa (Conneau *et al.*, 2019) ne produisent pas par défaut des représentations de phrases efficaces. Une bonne approche pour résoudre une tâche de similarité sémantique de textes consiste ainsi à affiner ces modèles pré-entraînés sur des jeux de données leur permettant d'améliorer leurs représentations de phrases. Le corpus anglais NLI (Natural Language Inference) (Bowman *et al.*, 2015) et son pendant multilingue MultiNLI (Williams *et al.*, 2018) peuvent être utilisés à cette fin. Ils comprennent des couples de phrases classés selon le type de relation qu'elles entretiennent (neutre, contradiction, inférence). Pour améliorer les modèles de langue, les paires de phrases sont passées à un transformeur qui génère des vecteurs de représentation de taille fixe. Ces représentations de phrases sont alors transmises à un classifieur qui prédit le label décrivant leur relation. Ceci permet de générer des représentations qui peuvent alors servir à d'autres tâches, en particulier, pour ce qui nous retient, une tâche mesurant la similarité sémantique entre des phrases. Le modèle Sentence Multilingual BERT (Sentence M-BERT) est un modèle ré-entraîné avec les corpus NLI. Le corpus STS (Cer *et al.*, 2017) peut être également utilisé dans cette optique, ce qui fait ici pleinement sens, puisque la tâche 1 est une transposition directe de la tâche de SemEval à laquelle le corpus STS est associé. Nous avons ainsi testé des modèles dérivés de BERT avec et sans ré-entraînement avec le corpus STS.

Le tableau 1 montre la synthèse des résultats obtenus par plusieurs des modèles que nous avons testés pour résoudre la tâche 1. STS fait référence à l'utilisation du corpus du STSBenchmark comme données d'entraînement pour affiner les modèles. Un sous ensemble des données du corpus DEFT 2020 associés à la tâche 1 (504 couples de phrases) sont utilisés lors de cette phase de ré-entraînement comme données d'évaluation pour guider le fine-tuning. Le corpus de développement qui a servi à l'évaluation des systèmes (Corpus Dev) comprenait 96 phrases, soit 16 de chaque classe.

Il est à noter que les résultats obtenus sur le corpus de développement issus du corpus d'entraînement ont été calculés à partir de mesures de similarité continues (nombres réels), alors que les résultats sur le corpus de test, à la demande des organisateurs, ont été calculés après discrétisation des mesures (entiers naturels).

Le modèle MUSE pour Universal Sentence Encoder Multilingue (Yang *et al.*, 2019), sans même être entraîné avec les données de DEFT ni avec celles du STSBenchmark (les données d'entraînement de DEFT n'ont été utilisées que pour l'évaluation) obtient des résultats légèrement inférieurs aux encoders de type BERT affinés pour générer des embeddings de phrases. Ce modèle est donc un

excellent candidat pour des approches multilingues dites zero-shot.

Les modèles qui présentent les meilleurs résultats sont ceux entraînés avec les données du STSBenchmark, ce qui dans un même mouvement les spécialise sur la tâche 1 et leur permet de générer des embeddings de phrases plus riches.

3.2 Tâche 2 : Sélection de phrases parallèles

Pour nos expérimentations, nous avons décomposé la tâche 2 en trois phases : (1) une tâche consistant à mesurer la similarité des phrases cibles avec la phrase source (2) une tâche d'ordonnancement selon la mesure de similarité et (3) une tâche de sélection de la phrase présentant la meilleure similarité.

Pour résoudre cette tâche, nous avons testés différents approches supervisées ainsi qu'une approche entièrement non supervisée, sans ré-entraînement.

Les approches supervisées testées sont deux modèles de classification, le modèle SGD (Bottou, 2010) et le modèle XGB (Chen & Guestrin, 2016). Ils ont été entraînés à partir de l'extraction de différentes caractéristiques des séquences d'entrée.

Les différentes features utilisées sont les suivantes :

— *Features Simples* : elles tiennent compte de la taille des phrases ramenée en nombre de caractères, en nombres de mots, en nombres de mots en commun, ainsi que de la fréquence d'apparition de ces phrases dans la collection.

— *Features Avancées* : elles renvoient à différents ratios qui tiennent compte de la similarité du début et de la fin des phrases, de la taille des phrases à comparer, de la taille de leur partie commune ainsi que des ratios de mots-clés en commun entre les deux phrases.

— *Word Embeddings* sont les représentations distribuées (plongement lexicaux) des mots. Nous avons utilisé le modèle word mover's distance (Kusner *et al.*, 2015) afin de calculer la distance entre les deux phrases en entrée et également calculer différentes distances entre les moyennes de vecteurs de mots de chaque phrase : Cosinus, City Block (Manhattan), Canberra, Euclidienne et Minkowski.

Nous avons combiné deux types de features, *simples* et *avancées* ainsi que les représentations distribuées des mots (word embeddings). Nous avons comparé les résultats de ces modèles au modèle MUSE (Chidambaram *et al.*, 2018), un modèle multilingue de représentation universel des phrases, utilisé pour le calcul des représentations des différentes phrases à l'entrée d'un classifieur. Le tableau 2 montre une synthèse des résultats obtenus par les différents modèles testés sur cette tâche, après ordonnancement des phrases cibles par similarité et extraction de la phrase présentant le plus de similarité avec la phrase source.

Il est à noter qu'à l'issue de la première phase où l'on établit une mesure de similarité entre phrases cibles et phrases source, il est possible de calculer la capacité des modèles à opérer une classification binaire (phrases équivalentes vs phrases indépendantes). Le tableau 3 présente les résultats obtenus par les différents modèles sur cette tâche de classification. Ceci permettrait de traiter les cas où plusieurs phrases cibles seraient parallèles à la phrase source et les cas où aucune d'entre elles ne le seraient, alors que ces informations sont perdues avec la méthode de sélection du meilleur candidat. Le corpus d'entraînement ne semblait toutefois pas contenir de cas de ce type, bien si l'énoncé de la tâche 2 semblait prévoir ces cas.

Synthèse des résultats - Tâche 2		
Méthode	Modèle	P@1
Features simples + TF-IDF	SGD_Clf1	0.975
Embeddings + Features avancées	SGD_Clf2	0.975
Embeddings + Features avancées	XGB_Clf1	0.975
Multilingue (apprentissage par transfert)	MUSE	1.0

TABLE 2 – Comparaison des modèles sur la tâche 2

Synthèse des résultats intermédiaires			
Représentation des entrées	Modèle	Précision Moyenne	Précision Pondérée
Features simples + TF-IDF	SGD_Clf1	0.88	0.88
Word Embed. + Features avancées	SGD_Clf2	0.91	0.91
Word Embed. + Features avancées	XGB_Clf1	0.94	0.94
Multilingue (apprentissage par transfert)	MUSE	0.94	0.94

TABLE 3 – Comparaison des modèles pour la classification en équivalence

3.3 Résultats d'évaluation

Dans cette section, nous présentons les résultats de l'évaluation officielle après soumission des trois meilleurs systèmes (modèles) évalués dans les sections précédentes. Le tableau 4 montre les résultats d'évaluation de chacun des modèles retenus par tâche, en comparaison aux résultats de référence (performances maximales) d'après l'évaluation des différents modèles participants à l'atelier. Nous remarquons que nos systèmes présentent des niveaux de performances opposés pour les deux tâches, comparés aux autres systèmes participants.

Pour la tâche 1, les modèles que nous avons testés sont encore loin des performances des meilleurs modèles concurrents évalués dans cet atelier (Sentence-M-BERT + STS, est près de 17% moins performant par rapport au système le plus performant), mais ils montrent néanmoins qu'il est possible d'obtenir des résultats intéressants dans l'optique de produire des applications multilingues dans un contexte où l'on dispose de peu voire d'aucune donnée d'entraînement pour certaines des langues cibles. Tels quels, les modèles que nous avons expérimentés sont entraînés en utilisant des données génériques et testés sur des données du domaine médical, impliquant ainsi deux niveaux de transfert

Tâche 1	M-BERT + STS	MUSE	Sentence M-BERT + STS	maximum
Sp-Cor	0.7499	0.7421	0.7679	-
p-value	3.1295e-75	7.0960e-73	6.1899e-81	-
EDRM	0.6533	0.6663	0.6838	0.8220
Tâche 2	SGD	MUSE	XGB	maximum
MAP	0.9850	0.9906	0.9396	0.9906

TABLE 4 – Évaluation finale des trois meilleurs modèles par tâche, évalués dans les résultats des tableaux 1 et 2. La colonne *maximum* montre les meilleurs performances des différents modèles participants.

d'apprentissage, la langue et le domaine, pour une tâche assez complexe (prédiction des votes sur six classes (0-5)).

Concernant la tâche 2 où l'objectif était de trouver une phrase parallèle à une phrase source dans un ensemble de phrases cibles, le plus performant des modèles que nous avons expérimentés, le modèle MUSE, obtient les meilleures performances parmi les différents systèmes participants. L'intérêt de ce modèle est qu'il n'a pas du tout été entraîné avec les données d'entraînement fourni par les organisateurs, ce qui en fait un modèle intéressant pour le développement d'applications multilingues sans donnée d'entraînement.

4 Conclusion

Nous avons présenté dans cet article les résultats de notre participation à l'atelier DEFT 2020 avec différents modèles multilingues d'appariement de texte, permettant de résoudre des tâches 1 et 2, auxquelles nous avons participé. Les expérimentations que nous avons menées pour résoudre la tâche 1 montrent que dans un contexte où le volume de données d'entraînement dans une langue est très limité, il est possible d'opérer un transfert d'apprentissage d'une langue à l'autre en recourant à des modèles d'apprentissage profond multilingues, comme MUSE et M-BERT. Cette approche permet d'exploiter conjointement des données d'apprentissage dans une langue pour laquelle on dispose d'un plus grand nombre de données et des données moins nombreuses dans la langue cible. Cette approche obtient cependant des résultats assez loin des performances obtenues par les meilleurs systèmes participants. A l'inverse, nous avons obtenu les meilleurs résultats dans la tâche 2 avec une approche pourtant encore plus restrictive où les données d'entraînement ne sont pas du tout exploitées (modèle MUSE).

Les modèles entraînés durant nos expérimentations sont des modèles de langue généralistes, pré-entraînés sur des textes qui ne relèvent pas du domaine médical. Une piste intéressante pour poursuivre ces travaux serait d'entraîner un modèle de langue multilingue spécialisé sur ce domaine, qui permettrait de disposer de représentations de phrases mieux adaptées aux corpus. A ce jour en effet, à notre connaissance seuls des modèles monolingues de langue anglaise spécialisés dans le domaine médical sont disponibles (Lee *et al.*, 2019). En outre, une extension de ces modèles pour prendre en compte une tâche de multiclassification peut être intéressante afin de déterminer le vote correspondant à un couple de phrases en entrées.

Références

BAZIZ M., BOUGHANEM M., AUSSENAC-GILLES N. & CHRISMENT C. (2005). Semantic cores for representing documents in ir. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, SAC '05, p. 1011–1017, New York, NY, USA : ACM. DOI : 10.1145/1066677.1066911.

BOTTOU L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, p. 177–186. Springer. DOI : 10.1007/978-3-7908-2604-3_16.

BOWMAN S. R., ANGELI G., POTTS C. & MANNING C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)* : Association for Computational Linguistics. arXiv preprint : L-1508.05326.

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.

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

CHEN T. & GUESTRIN C. (2016). Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, p. 785–794. DOI : 10.1145/2939672.2939785.

CHIDAMBARAM M., YANG Y., CER D., YUAN S., SUNG Y., STROPE B. & KURZWEIL R. (2018). Learning cross-lingual sentence representations via a multi-task dual-encoder model. *CoRR*. arXiv preprint : abs/1810.12836.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint : abs/1911.02116.

CONNEAU A., LAMPLE G., RINOTT R., WILLIAMS A., BOWMAN S. R., SCHWENK H. & STOYANOV V. (2018). XNLI : evaluating cross-lingual sentence representations. *CoRR*. arXiv preprint : abs/1809.05053.

DEHGHANI M., GOUWS S., VINYALS O., USZKOREIT J. & KAISER Ł. (2018). Universal transformers. arXiv preprint : L-1807.03819.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. arXiv preprint : L-1810.04805.

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.

GUO J., FAN Y., AI Q. & CROFT W. B. (2016). A deep relevance matching model for ad-hoc retrieval. New York, NY, USA : Association for Computing Machinery.

KUSNER M. J., SUN Y., KOLKIN N. I. & WEINBERGER K. Q. (2015). From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, p. 957–966 : JMLR.org. DOI : 10.5555/3045118.3045221.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2019). Biobert : a pre-trained biomedical language representation model for biomedical text mining. *CoRR*. arXiv preprint : abs/1901.08746.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized BERT pretraining approach. *CoRR*. arXiv preprint : abs/1907.11692.

MANNING C., RAGHAVAN P. & SCHÜTZE H. (2010). Introduction to information retrieval. *Natural Language Engineering*, **16**(1), 100–103. DOI : 10.5555/1394399.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. arXiv preprint : L-1911.03894.

PIRES T., SCHLINGER E. & GARRETTE D. (2019). How multilingual is multilingual bert ? *CoRR*. arXiv preprint : abs/1906.01502.

RADFORD A., NARASIMHAN K., SALIMANS T. & SUTSKEVER I. (2018). Improving language understanding with unsupervised learning. *Technical report, OpenAI*. Technical Report.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence embeddings using Siamese BERT-networks. 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. 3982–3992, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

WILLIAMS A., NANGIA N. & BOWMAN S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1112–1122 : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1101.

YANG Y., CER D., AHMAD A., GUO M., LAW J., CONSTANT N., ABREGO G. H., YUAN S., TAR C., SUNG Y.-H., STROPE B. & KURZWEIL R. (2019). Multilingual universal sentence encoder for semantic retrieval. arXiv preprint : L-1907.04307.