Définition automatique du niveau de la difficulté textuelle de documents

Molka Tounsi Dhouib, Ekaterina Kostrykina, Catherine Faron

Définition automatique du niveau de la difficulté textuelle de documents

Molka Tounsi Dhouib*, Ekaterina Kostrykina**
Catherine Faron*,**

**Université Côte d'Azur, Polytech Nice Sophia, France
nom.prenom@univ-cotedazur

*Université Côte d'Azur, CNRS, Inria, I3S, Sophia-Antipolis, France
nom@i3s.unice.fr

Résumé. La lisibilité ou la difficulté textuelle représentent une information importante pour de nombreuses applications éducatives telles que les systèmes de recommandation. Aujourd'hui, les meilleurs résultats pour cette tâche sont obtenus en utilisant des techniques de traitement automatique de la langue (TAL) et d'apprentissage profond. Dans ce travail, nous proposons une approche multilingue qui se base sur la combinaison des modèles de réseaux de neuronaux avec des caractéristiques linguistiques extraites du texte afin d'améliorer la qualité de l'évaluation. Nous avons testé notre approche sur deux benchmarks, et les résultats montrent que cette combinaison améliore les performances du système.

1 Introduction

De nos jours, de plus en plus de ressources électroniques pour l'auto-éducation sont disponibles. Cependant, la manière dont l'apprenant absorbe le matériel dépend en grande partie de son expérience dans le domaine d'étude et de la lisibilité d'une ressource électronique particulière. La lisibilité textuelle est une détermination de la facilité ou de la difficulté à comprendre le texte lu. La lisibilité est un aspect important de l'apprentissage, car il est nécessaire d'estimer si le texte transmet l'information d'une manière appropriée pour les apprenants. En outre, la rapidité avec laquelle une personne comprend un texte écrit détermine sa capacité à acquérir de nouvelles connaissances et à rester motivé pour cela. C'est pourquoi l'automatisation de la tâche d'évaluation de la lisibilité d'un texte est de plus en plus populaire aujourd'hui. Cette tâche peut être considérée comme un problème de classification de textes selon leur degré de complexité (Mirończuk et Protasiewicz, 2018). L'objectif du travail présenté dans cet article est d'établir et d'évaluer une approche multilingue permettant de définir automatiquement la difficulté textuelle des documents. Les questions de recherche que nous aborderons sont les suivantes : (i) Quel modèle pouvons-nous utiliser pour évaluer la difficulté textuelle ? (ii) Quelles caractéristiques linguistiques pouvons nous utiliser pour cette tâche ? (iii) Est ce que ces caractéristiques peuvent être facilement adaptées à d'autres langues ? Le reste de cet article est structuré comme suit : La section 2 donne un aperçu de l'état de l'art sur l'évaluation

de la lisibilité de textes. La section 3 détaille notre approche avec la description de l'extraction de caractéristiques personnalisées. La section 4 rapporte et discute les résultats de nos expériences menées sur deux corpus de l'état de l'art. La section 5 conclut et donne quelques orientations pour des travaux futurs.

2 État de l'art

Plusieurs approches de l'évaluation automatique de la lisibilité des textes ont été présentées dans la littérature (Vajjala, 2022) dont le point commun est l'étude des différentes caractéristiques du texte (statistiques, linguistiques, syntaxiques, morphologiques et sémantiques). Une première direction traditionnelle consiste à utiliser des mesures de lisibilité statistiques telles que la formule de Flesch (Kincaid et al., 1975), de Coleman (Coleman et Liau, 1975), les indices de Gunning FOG (Gunning, 1968), ou de Dale-Chal (Chall et Dale, 1995). Ces mesures ne prennent en compte que peu de caractéristiques brutes comme le nombre moyen de syllabes par mot, la longueur moyenne des phrases, les fréquences de mots ou les listes de vocabulaire. Ces mesures présentent deux inconvénients : (i) leur dépendance à la langue, d'autant plus que la plupart d'entre elles ne s'intéressent qu'à des textes en anglais, (ii) leur manque de fiabilité car elles peuvent qualifier un texte comme facile à lire même si son contenu est complètement absurde (Petersen et Ostendorf, 2009; Feng et al., 2010). Une récente direction de recherche consiste à utiliser des techniques de traitement automatique des langues (TAL) et d'apprentissage automatique et profond afin de traiter plusieurs caractéristiques et proposer des stratégies multilingues (Madrazo Azpiazu et Pera, 2020; Bengoetxea et Gonzalez-Dios, 2021; Heilman et al., 2008; Mohammadi et Khasteh, 2019). (Filighera et al., 2019) ont proposé une approche basée sur différents modèles pour produire des représentations vectorielles des textes (tels que word2vec, ELMO, GloVe et BERT) avec des modèles de réseaux de neurones pour la tâche de la prédiction, et plus précisément des architectures de type biLSTM et CNN. (Madrazo Azpiazu et Pera, 2020) ont proposé une étude visant à analyser la faisabilité d'un système multilingue d'évaluation de la lisibilité. Ce travail présente l'étude de plusieurs caractéristiques linguistiques, statistiques et sémantiques avec des techniques d'apprentissage automatique comme Random Forest, Decision Tree, Support Vector Machine (SVM) et Multinomial Naive Bayes. Vec2Read (Azpiazu et Pera, 2019) présente une stratégie d'évaluation automatique de la lisibilité basée sur l'apprentissage profond en utilisant des caractéristiques telle que la catégorie grammaticale des mots ou leurs caractéristiques morphologiques. MultiAzterTest (Bengoetxea et Gonzalez-Dios, 2021) permet d'évaluer la lisibilité textuelle pour l'anglais, l'espagnol et le basque. Cet outil se base sur plusieurs caractéristique linguistique (i.e la fréquence des mots, la longueur des phrases, le niveau de vocabulaire, l'utilisation de connecteurs grammaticaux). Sur la base de ces caractéristiques et d'une représentation du texte construite avec Fastext, un classifieur SVM est utilisé pour classer le texte.

Nous proposons une approche multilingue qui se base sur l'apprentissage profond et combine plusieurs caractéristiques linguistiques. La particularité de notre approche par rapport aux travaux existants réside dans (i) l'utilisation d'un modèle de classification BERT et (ii) l'intégration de caractéristiques simples pour améliorer les performances de l'approche.

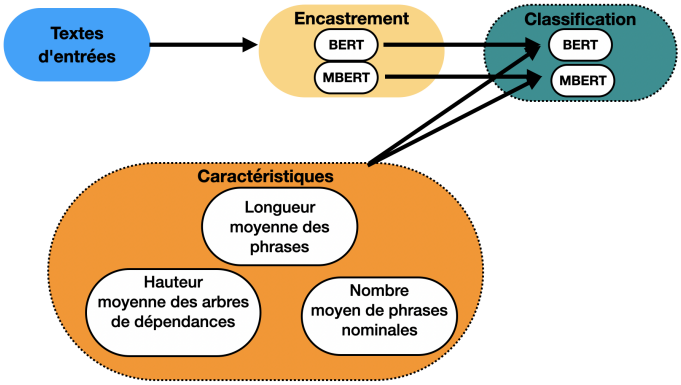

FIG. 1 – *Aperçu général de l'approche proposée pour l'évaluation automatique de la lisibilité de textes*

3 Approche proposée

Nous présentons une approche multilingue basée sur la classification des textes en utilisant des réseaux de neurones et plus précisément les modèles de BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019). Nous avons choisi les modèles BERT car ils sont pré-entraînés sur de vastes corpus, ont démontré une excellente performance et peuvent être facilement adaptés pour la tâche. De plus, ce type de modèle permet de généraliser notre approche à différentes langues et différents domaines. Cependant l'approche peut s'appliquer à d'autres modèles. Nous avons étendu notre modèle de base en utilisant deux caractéristiques linguistiques issues de l'état de l'art, à savoir : (i) la longueur moyenne des phrases et (ii) la longueur moyenne de l'arbre de dépendance syntaxique des phrases. Même si ces deux premières caractéristiques sont corrélées, néanmoins, il y a des exceptions où des phrases courtes peuvent avoir une structure grammaticale complexe (clauses subordonnées). Nous avons encore enrichi le modèle avec une nouvelle caractéristique qui est le nombre moyen des expressions nominales par phrase.

Dans la suite, nous expliquons les avantages, et décrivons la manière d'extraire ces caractéristiques à partir des textes et le processus d'intégration de ces dernières aux modèles de BERT. La figure 1 montre une vue globale de notre approche.

3.1 Caractéristiques linguistiques des textes prises en compte

3.1.1 Longueur moyenne des phrases (S)

La première caractéristique qui devrait améliorer la qualité de la classification des textes selon leur lisibilité est la longueur moyenne des phrases du texte. Elle a été utilisée dans plusieurs travaux de l'état de l'art comme par exemple (Azpiazu et Pera, 2019; Filighera et al., 2019). En effet des phrases excessivement longues rendent le texte difficile à comprendre et perturbent la fluidité de la lecture. Pour extraire cette caractéristique, nous avons commencé

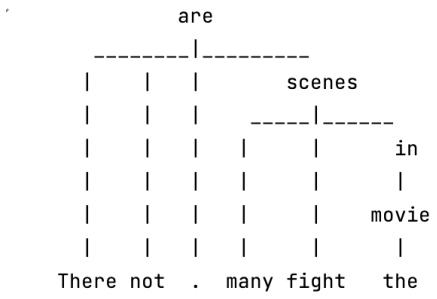

FIG. 2 – *Exemple d'un arbre de dépendance*

par diviser le texte en phrases à l'aide de la fonction sent_tokenize de la bibliothèque NLTK [1]. Ensuite, afin de diviser une phrase en mots, nous utilisons la fonction word_tokenize qui divise une phrase en une liste de mots, en tenant compte de la ponctuation et d'autres caractères spéciaux. Enfin, nous calculons la longueur moyenne des phrases pour chaque texte.

3.1.2 Hauteur moyenne des arbres de dépendance (D)

Cette caractéristique représente une mesure linguistique couramment utilisée en traitement automatique de la langue (Azpiazu et Pera, 2019; Heilman et al., 2008) pour évaluer la complexité syntaxique d'un texte. L'hypothèse est qu'un texte qui contient des phrases avec une structure grammaticale complexe implique des arbres de dépendance plus profonds, ce qui rend la compréhension difficile car le lecteur doit suivre des relations grammaticales complexes entre les mots. Avec cette caractéristique, nous avons mesuré la complexité d'un texte en s'appuyant sur l'étude des relations grammaticales entre les mots d'une phrase en termes de dépendance entre eux. Dans la grammaire de dépendance, une phrase est représentée comme un ensemble de relations binaires entre les mots, appelées dépendances. Chaque mot d'une phrase est considéré comme un nœud dans un arbre de dépendance, et les arêtes entre les nœuds représentent les relations grammaticales entre les mots. Afin d'extraire ce type de caractéristiques, nous avons divisé le texte en phrases et, nous avons utilisé la bibliothèque Spacy [2] pour effectuer une analyse linguistique, qui comprend l'analyse morphologique et syntaxique des phrases. Nous avons extrait l'arbre de dépendance de chaque phrase et nous avons parcouru ces arbres pour calculer la hauteur moyenne des phrases du texte. La figure 2 montre un exemple d'arbre de dépendance pour la phrase *"There are not many fight scenes in the movie"*. Nous pouvons voir dans cet exemple que la hauteur de cet arbre de dépendance est de 5 niveaux.

3.1.3 Nombre moyen de phrases nominales (N)

Une autre caractéristique qui devrait améliorer la qualité de la classification de textes selon leur lisibilité est le nombre moyen d'expressions nominales par phrase. Une phrase ordinaire contient généralement une ou deux expressions nominales, tandis que les phrases plus longues

1. https ://www.nltk.org
2. https ://spacy.io

et plus complexes peuvent en contenir trois ou plus. Le nombre moyen d'expressions nominales par phrase dans les phrases complexes tend à être plus élevé, car il est courant d'utiliser un langage plus descriptif pour transmettre des idées complexes. Ces phrases peuvent contenir jusqu'à quatre ou cinq expressions nominales. Cette caractéristique n'est pas couramment utilisée pour évaluer la difficulté d'un texte ce qui a motivé notre décision d'évaluer son apport dans cette tâche. Afin d'extraire ce type de caractéristiques, nous avons utilisé la bibliothèque lingfeat [3] pour effectuer une analyse linguistique, qui comprend l'analyse morphologique et syntaxique des phrases. Ensuite, nous avons extrait les informations syntaxiques du texte à l'aide de la même bibliothèque. Enfin, nous avons sélectionné le nombre moyen d'expressions nominales par phrase dans le texte. La bibliothèque lingfeat ne prenant en charge que la langue anglaise, nous avons testé l'intérêt de cette caractéristique sur les textes anglais seulement.

3.2 Ajout des caractéristiques au modèle

Pour intégrer chacune des caractéristiques et les combiner, nous avons ajouté une couche supplémentaire dans les modèles BERT et MBERT. Chacune de ces caractéristiques est représentée par un vecteur. Par exemple, la longueur moyenne des phrases est représentée par [5], la hauteur moyenne des arbre de dépendance est représentée par [4] et le nombre moyen de phrases nominales est représenté par [3]. La combinaison de ses caractéristiques revient alors à concaténer les trois vecteurs pour construire un seul vecteur plus long (c.à.d. [5, 4, 3]). Ce vecteur de caractéristiques est concaténé avec les vecteurs de BERT de taille 768 pour construire un vecteur unique utilisé comme entrée pour le modèle.

4 Expérimentations

4.1 Données et protocole d'évaluation

OneStopEnglishCorpus [4] est une collection de textes écrits en anglais de différents types et genres, tels que des textes de fiction, des textes académiques, des articles de journaux, etc. Les textes du jeu de données sont divisés en trois niveaux de difficulté : (i) débutant, (ii) intermédiaire et (iii) avancé. Le jeu de données contient 567 textes, avec 189 textes par niveau de difficulté. VikiWiki [5] est un jeu de données multilingues qui contient des textes divisés en deux niveaux de complexité : facile et difficile. Les textes sont issus de Wikipedia et Vikidia et sont rédigés en 6 langues : (i) anglais, (ii) espagnol, (iii) français, (iv) italien, (v) basque et (vi) catalan. Dans ce travail, nous avons considéré les cinq premières langues. Afin d'entraîner et d'évaluer le modèle, nous avons décidé de diviser les ensembles de données comme suit : (i) 70% comme ensemble d'entraînement, (ii) 10% comme ensemble de validation et (iii) 20% comme ensemble de test. Nous avons réalisé 34 expérimentations sur les deux jeux de données en faisant varier la version de BERT utilisée et les paramètres considérés, et nous avons utilisé l'*accuracy* [6] comme métrique d'évaluation.

3. https ://github.com/brucewlee/lingfeat
4. https ://github.com/nishkalavallabhi/OneStopEnglishCorpus
5. https ://github.com/ionmadrazo/VikiWiki
6. https ://huggingface.co/spaces/evaluate-metric/accuracy

4.2 Résultats

Les tableaux 1 et 2 présentent les résultats de notre approche respectivement sur les corpus OneStopEnglishCorpus et VikiWiki en utilisant le modèle BERT, et sur le corpus VikiWiki en utilisant le modèle mulitilingue BERT. Comme le montre le tableau 1, sur le corpus OneStopEnglishCorpus, nous obtenons les meilleurs résultats en intégrant les trois caractéristiques extraites à partir du texte (c.à.d. la longueur moyenne des phrases, la hauteur moyenne de l'arbre de dépendance et le nombre moyen d'expressions nominales par phrase). En comparant les résultats des textes en anglais du corpus VikiWiki, nous trouvons que le modèle BERT_base qui est entraîné sur un corpus de textes en anglais fournit de meilleurs résultats que le modèle multilingue MBERT entraîné sur des textes de différentes langues. Cependant, l'intégration des trois caractéristiques statistiques des textes à ces modèles permettent d'obtenir la même performance (0.98).

Modèle	Accuracy	
	OneStopEnglish	VikiWiki
BERT_base	0.78	0.97
BERT_base_S	0.87	0.97
BERT_base_D	0.91	**0.98**
BERT_base_S_D	0.92	0.97
BERT_base_N	0.93	**0.98**
BERT_base_SL_D_N	**0.94**	**0.98**

TAB. 1 – *Accuracy des modèles basés sur BERT sur les corpus OneStopEnglish et VikiWiki.*

Modèle	Accuracy				
	Anglais	Français	Espagnol	Italien	Basque
BERT_M	0.95	0.90	0.90	0.95	0.87
BERT_M_S	0.96	0.91	0.91	0.96	0.89
BERT_M_D	0.97	**0.92**	0.91	0.96	0.91
BERT_M_S_D	0.97	**0.92**	**0.92**	**0.97**	**0.92**
BERT_M_N	**0.98**	-	-	-	-
BERT_M_S_D_N	**0.98**	-	-	-	-

TAB. 2 – *Accuracy des modèles basés sur MBERT sur le corpus VikiWiki.*

Le tableau 2 montre que la prise en compte de plusieurs combinaisons de caractéristiques extraites à partir du texte donne toujours les meilleurs résultats. Les différents résultats obtenus montrent qu'il est approprié d'étendre les modèles de réseaux neuronaux avec des caractéristiques linguistiques. Cette approche peut aider à sauvegarder les informations importantes de la structure du texte, qui peuvent être perdues au cours du processus de représentation des textes, et les prendre en considération lors de la prédiction du niveau de lisibilité des documents. En effet, l'utilisation de la caractéristique supplémentaire simple de la longueur moyenne des phrases d'un texte, a permis d'obtenir de meilleurs résultats que le modèle BERT simple. Ce-

pendant, l'utilisation de la hauteur moyenne d'un arbre de dépendance et du nombre moyen d'expressions nominales par phrase dans le texte a permis d'améliorer encore les résultats. Nos expérimentations montrent que le nombre d'expressions nominales dans une phrase est également une mesure représentative de la complexité et donc de la lisibilité d'une phrase. D'après les différentes expérimentations réalisées sur cinq langues différentes, nous pouvons constater que pour le basque par exemple, chaque caractéristique utilisée séparément améliore les performances du système et le meilleur résultat a été obtenu en combinant toutes les caractéristiques. Au contraire, pour le français l'utilisation de la hauteur moyenne de l'arbre de dépendance est suffisante pour déduire la difficulté du texte. Finalement, d'après nos résultats sur les deux corpus de la langue anglaise nous pouvons déduire que ce n'est pas que la langue qui peut définir les caractéristiques à utiliser pour cette tâche, mais la qualité des textes et les domaines influencent le choix des caractéristiques à mettre en place.

5 Conclusion

Dans cet article, nous avons présenté une approche pour l'évaluation de la lisibilité des textes. Nous avons commencé par considérer des textes en anglais, et nous avons étendu notre approche pour traiter d'autres langues. Notre but était de définir une approche multilingue. Notre approche combine des modèles de réseaux de neurones avec des caractéristiques statistiques de textes : (i) longueur moyenne d'une phrase dans un texte, (ii) hauteur moyenne de l'arbre de dépendance d'une phrase et (iii) nombre moyen de phrases nominales. Nous avons obtenu les meilleurs résultats en combinant les trois types de caractéristiques linguistiques.

En perspectives, à court terme nous allons étendre nos expérimentations en utilisant la méthode de la validation croisée pour évaluer notre approche et la comparer avec l'état de l'art. Par la suite, nous projetons de prendre en compte la difficulté des textes par rapport à un domaine spécifique et la corrélation de la difficulté au vocabulaire. Une première piste est de considérer la fréquence des mots pour intégrer la notion de vocabulaire. Nous souhaitons aller plus loin en intégrant des caractéristiques sémantiques différenciant la complexité selon le domaine du document.

Références

Azpiazu, I. M. et M. S. Pera (2019). Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics 7*, 421–436.

Bengoetxea, K. et I. Gonzalez-Dios (2021). Multiaztertest : A multilingual analyzer on multiple levels of language for readability assessment. *arXiv preprint arXiv :2109.04870*.

Chall, J. S. et E. Dale (1995). Readability revisited : The new dale-chall readability formula. *(No Title)*.

Coleman, M. et T. L. Liau (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology 60*(2), 283.

Devlin, J., M. Chang, K. Lee, et K. Toutanova (2019). BERT : pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, et T. Solorio

(Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics.

Feng, L., M. Jansche, M. Huenerfauth, et N. Elhadad (2010). A comparison of features for automatic readability assessment.

Filighera, A., T. Steuer, et C. Rensing (2019). Automatic text difficulty estimation using embeddings and neural networks. In *Transforming Learning with Meaningful Technologies : 14th European Conference on Technology Enhanced Learning, EC-TEL 2019, Delft, The Netherlands, September 16–19, 2019, Proceedings 14*, pp. 335–348. Springer.

Gunning, R. (1968). Readability yardsticks. *The Technique of Clear Writing. New York : McGraw-Hill*.

Heilman, M., K. Collins-Thompson, et M. Eskenazi (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pp. 71–79.

Kincaid, J. P., R. P. Fishburne Jr, R. L. Rogers, et B. S. Chissom (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Madrazo Azpiazu, I. et M. S. Pera (2020). Is cross-lingual readability assessment possible ? *Journal of the Association for Information Science and Technology 71*(6), 644–656.

Mirończuk, M. M. et J. Protasiewicz (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications 106*, 36–54.

Mohammadi, H. et S. H. Khasteh (2019). Text as environment : A deep reinforcement learning text readability assessment model. *arXiv preprint arXiv :1912.05957*.

Petersen, S. E. et M. Ostendorf (2009). A machine learning approach to reading level assessment. *Computer speech & language 23*(1), 89–106.

Vajjala, S. (2022). Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, pp. 5366–5377. European Language Resources Association.

Summary

For many educational applications, such as learning resource recommendation systems, the textual difficulty of a text is a key information. Today, the best results for this task are obtained by using NLP and deep learning techniques. However, the use of these methods can result in the loss of statistical linguistic information that is important for determining text readability more accurately. In our work, we propose an approach for assessing text readability by combining neural network models with linguistic features extracted from the text and integrated into the model to improve the quality of the neural network models. Experimental results show that this combination improves system performance.