Simplification de textes via l'identification de passages faisant référence à des informations implicites et l'estimation d'une similarité stylistique

Sujet de stage de M2 Informatique
Patrice Bellot — Liana Ermakova
Aix-Marseille Université — CNRS, LIS, équipe R2I
Université de Bretagne Occidentale, HCTI EA-4249
patrice.bellot@lis-lab.fr
liana.ermakova@univ-brest.fr

15 octobre 2021

Présentation générale. *Ce stage s'inscrit dans le cadre du projet collaboratif SimpleText[1] que nous menons avec les Universités de Bretagne Occidentale, de Bretagne Sud, de Lyon II, d'Avignon en France mais aussi avec l'Université d'Amsterdam, l'Université Sechenov à Moscou et l'Université de Minho (Portugal). Le financement du stage n'est pas encore assuré mais une poursuite en doctorat dans l'une des universités partenaires est d'ores et déjà envisagée. La ou le stagiaire présentera ses travaux aux partenaires du projet puis durant le track SimpleText de la conférence CLEF 2022[2].*

La simplification de textes peut faciliter la lecture de documents : elle consiste à adapter automatiquement l'écriture d'un texte à un lectorat cible. Il peut s'agir de personnes ne maîtrisant pas suffisamment la langue des documents, de personnes non expertes des domaines abordés ou encore présentant des difficultés cognitives liées au processus de lecture lui-même. Automatiser la simplification de textes présente de nombreux challenges liés au traitement automatique des langues et à l'extraction et la recherche d'information. Une première difficulté, au cœur de ce stage, consiste à mesurer la complexité d'un document (François 2009) et, par extension, à estimer à quel point un texte a été simplifié sans pour autant avoir perdu (trop) d'informations ou trop changé de style ou de tonalité générale. De façon pratique, cette estimation nécessite d'avoir au préalable identifié les facteurs reflétant la complexité sachant qu'il a été montré que les mesures de lisibilité telles que celles de Flesch et de Dale-Chall ne sont pas convaincantes

[1] https://simpletext-madics.github.io

[2] https://clef2022.clef-initiative.eu

pour un lectorat général (Leroy et al., 2013)[3]. De façon générale, la qualité de la simplification peut s'estimer selon plusieurs dimensions : la compréhension du texte (Eme & Rouet, 2001), la perte, l'ajout ou la modification erronée d'informations factuelles ou « distorsion informationnelle » (Nurbakova et al., 2020), le changement ou la perte du style.

Les autres difficultés, non traitées dans ce stage mais dont il faut néanmoins avoir conscience, concernent le processus de simplification — réécriture — lui-même, aussi bien dans la forme (lexique et syntaxe mais aussi style et formes discursives (Del Ré et al., 2019)) que dans le fond (informations, opinions et points de vue…). La simplification correspond à un processus où sont combinées des étapes s'apparentant au résumé automatique avec la suppression des passages de texte inopportuns ou trop redondants et qui brouillent la compréhension, à l'amélioration de la lisibilité par l'identification des passages les plus complexes et la réécriture selon des formes lexicales et syntaxiques simplifiées (Hijazi, 2020), et empruntant à la recherche d'information par l'identification puis l'explication des termes les plus complexes (sémantique distributionnelle ou analyse morphologique) mais aussi par celle d'informations implicites ou autrement dit de connaissances étrangères aux lecteurs mais perçues comme communes par l'auteur, du moins pour le lectorat cible supposé[4]. La simplification se distingue notamment du résumé automatique (voir Ermakova et al., 2019 pour un aperçu général) pour lequel seules les idées ou informations jugées essentielles sont conservées et la taille des documents réduite d'un facteur important. En effet, simplifier un texte ne consiste pas toujours à réduire sa taille puisqu'il peut s'avérer nécessaire de développer un concept pour le rendre explicite ou d'ajouter des informations ou des définitions — qui sont à chercher sur le Web à partir d'une requête automatiquement générée — pour expliquer les expressions ou termes les plus complexes (Bellot et al., 2016).

Objectifs du stage et étapes. *On se propose de conduire deux études. Le temps consacré à chacune d'elle,* a priori *équivalent, dépendra de l'intérêt de la ou du stagiaire et des résultats intermédiaires obtenus. La première étude défriche une problématique peu explorée et comprend une expérimentation exploitant des classifieurs neuronaux et des modèles de langue (par exemple, Jurassic-1[5], Mutlilingual T5[6], etc.). La*

[3] alors qu'elles le sont dans le cas de certaines dyslexies (Sitbon & Bellot, 2008) où la facilitation de la lecture permet celle de la compréhension. Cela dit, la prise en compte de la lisibilité permet de tenir compte de critères complémentaires à ceux de la seule pertinence informationnelle et n'est donc pas à écarter complètement pour la simplification (Tavernier & Bellot, 2011).

[4] Notons aussi que l'implicite peut concerner des intentions, surtout dans les textes narratifs, moins dans les textes journalistiques ou scientifiques (Marin et al., 2007) ou encore des émotions (Balahur et al., 2012).

[5] https://studio.ai21.com/
[6] https://github.com/google-research/multilingual-t5

deuxième étude consiste surtout à une revue de la littérature et à l'expérimentation et l'intégration de modules logiciels existants afin de contribuer à une mesure d'estimation de la qualité multidimensionnelle d'une simplification textuelle.

La première étude concerne l'identification de l'implicite, autrement dit des passages textuels nécessitant la recherche de connaissances non explicites dans le texte mais importantes à sa compréhension. Cette question est relativement nouvelle. Il s'agira tout d'abord d'identifier les jeux de données déjà disponibles (par exemple Becker et al., 2019) et de confronter les schémas d'annotation proposés aux textes que l'on souhaite simplifier et qui seront issus de la tâche SimpleText de CLEF[7]. L'ensemble pourra alors être exploité pour apprendre un modèle de classification et d'annotation de « *passages faisant référence à des informations implicites* » dont il faudra déterminer les descripteurs les plus utiles ou « *termes à expliquer* ». Des modèles de langue pourront être employés au sein d'architectures neuronales pour la classification et la recherche du contexte explicite (cas d'usage, définition, analogie, exemples, etc.) dans un source externe (e.g. Wikipédia) puis comparés à des approches plus transparentes (forêts aléatoires). Le tout fournira, via une approche de classification simple, une première base de performance pour des travaux ultérieurs puisqu'il ne semble pas que l'on puisse en trouver dans la littérature ou alors restreintes à quelques formes d'implicite (Atkinson et al., 2009). Notons toutefois que Becker et al. (2021) proposent une approche pour générer, à partir de modèles de langue pré-appris, la connaissance (au sens de suites lexicales) nécessaire à la compréhension de paires de phrases. Cela peut alors entraîner la question suivante en ce qui nous concerne : est-il nécessaire, pour une simplification textuelle, d'identifier au préalable les phrases faisant appel à de la connaissance implicite ou devons-nous systématiquement identifier cette dernière et la rajouter dans le texte « simplifié » ? Quelle est alors la limite acceptable dans l'explicitation d'un texte (surcharges et redondances, dérives informationnelles inévitables) ?

La seconde étude concerne l'identification des caractéristiques stylistiques, des figures de style présentes et de la tonalité (émotion, polarité), autant d'aspects que le processus de simplification doit être en mesure de conserver. Cette seconde étude, réalisée à partir de logiciels existants et dont certains restent à identifier, permettra de proposer une mesure de similarité entre deux textes selon des indices surfaciques qui définissent le style de l'auteur (anaphores, répétitions, formes syntaxiques atypiques ou fréquentes…), selon des figures de style plus ou moins complexes (métaphores, oxymores, analogies…) mais aussi selon sa tonalité ou sa polarité, des émotions communiquées, de l'ironie ou des sarcasmes (Joshi et al., 2017). Il s'agira tout d'abord d'effectuer une revue de la littérature concernant les éléments qui viennent d'être cités (quelles approches, quels jeux de données, quels outils logiciels, quelles performances) puis d'appliquer

[7] https://simpletext-madics.github.io/

et d'évaluer un certain nombre de solutions sur un ensemble de textes. Il s'agira d'estimer à quel point les similarités entre les textes, selon les dimensions évoquées, sont corrélées à celles que l'on peut estimer avec beaucoup moins d'effort via des représentations continues distribuées (plongements lexicaux) ou des modèles de langues tels que SentenceBERT (Reimers & Gurevych, 2019) adaptés au calcul de nombreuses similarités deux à deux. Cette seconde étude contribuera à l'estimation automatique de la qualité de la simplification textuelle.

Références :

- Atkinson, J., Ferreira, A., & Aravena, E. (2009). Discovering implicit intention-level knowledge from natural-language texts. Knowledge-Based Systems, 22(7), 502-508.
- Balahur, A., Hermida, J. M., & Montoyo, A. (2012). Detecting implicit expressions of emotion in text: A comparative analysis. Decision support systems, 53(4), 742-753.
- Becker, M., Korfhage, K., & Frank, A. (2019). Implicit knowledge in argumentative texts: an annotated corpus. arXiv preprint arXiv:1912.10161.
- Becker, M., Liang, S., & Frank, A. (2021, June). Reconstructing Implicit Knowledge with Language Models. In Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures (pp. 11-24).
- Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., & Tannier, X. (2016). INEX Tweet Contextualization task: Evaluation, results and lesson learned. Information Processing & Management, 52(5), 801-819.
- Alessandra Del Ré, Fabrice Hirsch et Christelle Dodane, « L'ironie dans le discours : des premières productions enfantines aux productions des adultes », Cahiers de praxématique [En ligne], 70 | 2018, mis en ligne le 22 janvier 2019. URL : http://journals.openedition.org/praxematique/4796 ; DOI : https://doi.org/10.4000/praxematique.4796
- Eme Elsa, Rouet Jean-François, « Les connaissances métacognitives en lecture-compréhension chez l'enfant et l'adulte », Enfance, 2001/4 (Vol. 53), p. 309-328. DOI : 10.3917/enf.534.0309. URL : https://www.cairn.info/revue-enfance1-2001-4-page-309.htm
- Liana Ermakova, Patrice Bellot, Pavel Braslavski, Jaap Kamps, Josiane Mothe, Diana Nurbakova, Irina Ovchinnikova, Eric Sanjuan. Text Simplification for Scientific Information Access: CLEF 2021 SimpleText Workshop. 43rd edition of the annual BCS-IRSG European Conference on Information Retrieval : Advances in Information Retrieval (ECIR 2021), Mar 2021, Lucca (virtual), Italy. [PDF]
- Liana Ermakova, Patrice Bellot, Pavel Braslavski, Jaap Kamps, Josiane Mothe, et al.. Overview of SimpleText CLEF 2021 Workshop and Pilot Tasks. 12th Conference and Labs of the Evaluation Forum (CLEF 2021), Sep 2021, Bucharest (on line), Romania. pp.2212 - 2227. [PDF]
- Ermakova, L., Cossu, J. V., & Mothe, J. (2019). A survey on evaluation of summarization methods. Information processing & management, 56(5), 1794-1814.

- François, T. (2009, June). Modèles statistiques pour l'estimation automatique de la difficulté de textes de FLE. In Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles. REncontres jeunes Chercheurs en Informatique pour le Traitement Automatique des Langues (pp. 61-70).

- Hijazi, R. (2020, June). Transformations syntaxiques entre niveaux de simplification dans le corpus Newsela. In 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). Volume 3: Rencontre des Étudiants Chercheurs en Informatique pour le TAL (pp. 137-150). ATALA; AFCP.

- Joshi, A., Bhattacharyya, P. & Carman, M. J. Automatic Sarcasm Detection. ACM Computing Surveys 50, 1–22 (2017).

- Leroy, G., Endicott, J.E., Kauchak, D., Mouradi, O., Just, M.: User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. Journal of medical Internet research 15(7), e144 (2013)

- Brigitte Marin, Jacques Crinon, Denis Legros et Patrick Avel, « Lire un texte documentaire scientifique : quels obstacles, quelles aides à la compréhension ? », Revue française de pédagogie [En ligne], 160 | juillet-septembre 2007. URL : http://journals.openedition.org/rfp/786 ; DOI : https://doi.org/10.4000/rfp.786

- Nurbakova, D., Ermakova, L., & Ovchinnikova, I. (2020, November). Understanding the Personality of Contributors to Information Cascades in Social Media in response to the COVID-19 Pandemic. In 2020 International Conference on Data Mining Workshops (ICDMW) (pp. 45-52). IEEE.

- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

- Sitbon, L., & Bellot, P. (2008). A readability measure for an information retrieval process adapted to dyslexics. In Second international workshop on Adaptive Information Retrieval (AIR 2008 in conjunction with IIiX 2008) (pp. 52-57).

- Tavernier, J., & Bellot, P. (2011, December). Flesch and dale-chall readability measures for INEX 2011 question-answering track. In International Workshop of the Initiative for the Evaluation of XML Retrieval (pp. 235-246). Springer, Berlin, Heidelberg.