Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Évaluation d'un générateur automatique de reformulations médicales

Ioana Buhnila[1] Amalia Todirascu[1]
(1) LiLPa UR 1339, Université de Strasbourg, 67000 Strasbourg, France
ioana.buhnila@etu.unistra.fr, todiras@unistra.fr

RESUME

Les textes médicaux sont difficiles à comprendre pour le grand public à cause des termes de spécialité. Ces notions médicales ont besoin d'être *reformulées* en utilisant des mots de la langue commune. La **reformulation** représente le processus de réécriture qui a le rôle d'expliquer ou simplifier une phrase ou syntagme. Nous présentons la méthodologie de construction d'un jeu de données original (termes et reformulations) permettant la détection et génération des nouvelles reformulations médicales. Pour compléter ce corpus, nous menons des expériences de génération automatique de reformulations médicales sous-phrastiques avec l'outil **APT** (Nighojkar & Licato, 2021), qui s'appuie sur des techniques d'apprentissage profond. Nous adaptons le modèle de langue de type Transformer **T5** (Raffel et al., 2020) avec des termes médicaux et leurs reformulations annotés manuellement en français et en roumain, langue romane peu dotée en ressources pour le TAL. Nous présentons une analyse détaillée des résultats de la génération automatique des paraphrases.

ABSTRACT

Evaluation of an automatic medical paraphrase generator
Medical texts are difficult to understand for laypeople due to technical terms. Medical concepts need to be *paraphrased* using words from the common language. **Paraphrasing** is the process of rewriting with the aim of explaining or simplifying a sentence or phrase. We present the method for manually building a dataset for paraphrase detection and generation. We are conducting experiments on automatic generation of sub-phrastic medical paraphrases with the machine learning tool **APT** (Nighojkar & Licato, 2021), using deep learning techniques. We train the Transformer **T5** language model (Raffel et al., 2020) with manually annotated medical terms and paraphrases in French and Romanian, a Romance language lacking resources for NLP applications. We present a detailed analysis of the results of the paraphrase generation.

MOTS-CLES : génération automatique de la reformulation, apprentissage automatique, texte médical, Transformer T5
KEYWORDS: paraphrase generation, machine learning, medical text, T5 Transformer

1 Introduction

Les connaissances médicales de spécialité sont facilement consultables en ligne. Pourtant, elles ne sont pas toujours facilement *compréhensibles* pour le grand public. Cette difficulté est due au grand nombre de *termes médicaux* employés dans ces textes. *Le terme* est une unité lexicale de spécialité

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 80–93.

qui représente des connaissances spécifiques à un domaine du savoir (Costa, 2005). La terminologie médicale est difficile à comprendre pour un non-spécialiste à cause de l'origine grecque ou latine de dénominations, par exemple le terme « myocardiaque », formé avec une base latine « myo » (muscle) et une base grecque « cardia » (cœur) (Grabar & Hamon, 2016). Ces connaissances spécialisées sont faciles à comprendre pour un spécialiste, mais leur sens reste souvent obscur pour le grand public, d'où la nécessité d'une vulgarisation constante dans ce domaine (Grabar & Hamon, 2016). *Vulgariser* la médecine représente le processus de simplification lexicale et syntaxique qui a comme but de rendre les notions médicales très techniques compréhensibles en fonction du public cible. Informer correctement le grand public sur les questions médicales est une tâche qui demande un effort soutenu, vu l'innovation continue et les défis sanitaires constants (Pecout et al., 2019).

Nous considérons que les reformulations sont nécessaires dans la vulgarisation scientifique (Vargas, 2008 ; Cardon, 2018). *La reformulation* est un processus linguistique de transformation du discours qui a le rôle d'expliquer, simplifier ou pointer une phrase ou un syntagme. Nous travaillons sur des *reformulations sous-phrastiques*, ne dépassant pas la portée de la phrase, qui se présentent sous forme d'explications, de paraphrases (syntagmes synonymiques) ou de définitions de termes scientifiques monolexicaux et polylexicaux, de type « <terme>placebo</terme> ou <ref>absence d'intervention</ref> ». Notre objectif est double : de *construire* manuellement *un corpus valide* de *termes médicaux et leurs reformulations* (au sens large du terme) et de *générer automatiquement d'autres reformulations médicales* à partir de nos données annotées précédemment sur deux corpus, en français et en roumain, et ainsi agrandir automatiquement nos corpus de reformulations. Ce qui nous motive davantage dans notre recherche est le nombre restreint de ressources pour la reformulation dans le domaine médical pour la simplification de textes (Cardon, 2018 ; Cardon & Grabar, 2019), particulièrement pour des langues de moindre diffusion comme le roumain (Buhnila, 2021).

Nous présentons dans la section 2 l'état de l'art sur la génération des reformulations, suivi par notre méthodologie et les données exploitées (dans la section 3). Dans la section 4 nous présentons nos expériences de génération automatique, l'analyse et l'évaluation des résultats de génération de reformulations médicales. La conclusion, la comparaison avec l'état de l'art et des perspectives de recherches sont présentées dans la section 5.

2 Contexte

Certaines études ont été réalisées sur la reformulation dans le domaine de la médecine (Elhadad & Sutaria, 2007 ; Deléger & Zweigenbaum, 2009 ; Grabar & Hamon, 2015). Des ressources lexicales comme WordNet (Miller, 1998) pour la langue générale et UMLS (Bodenreider, 2004) ou Snomed (Spackman et al., 1997 ; Donnelly, 2006) pour le domaine médical sont utiles pour l'identification automatique des paraphrases selon leur degré de synonymie (Cardon & Grabar, 2019 ; Koptient et al., 2019). Il existe plusieurs travaux sur la simplification lexicale automatique des textes écrits (Specia et al., 2012 ; Shardlow, 2014 ; Grabar & Hamon, 2015 ; Saggion, 2017 ; Cardon, 2021). Ces études sont réalisées sur des termes simples et majoritairement sur la paraphrase phrastique. Nous travaillons sur les reformulations sous-phrastiques de termes médicaux simples et polylexicaux. Nous prenons en compte des reformulations variées (définitions, exemples, explications) qui sont très différentes au niveau lexical et syntaxique par rapport à l'original. Notre but est de construire un corpus de reformulations qui illustre ces phénomènes et de générer des nouvelles reformulations médicales.

Dans ce sens, nous nous intéressons à l'apprentissage automatique par réseaux de neurones, car plusieurs méthodes peuvent être appliquées pour travailler sur la paraphrase :

- *Similarité sémantique* textuelle (STS), qui mesure le degré d'équivalence des textes ou des phrases similaires qui contiennent des mots en commun (Agirre et al., 2016) ;
- *Identification de paraphrase (PI)*, qui identifie si deux phrases ou segments ont le même sens (Brockett & Dolan, 2005 ; Xu et al., 2015) ;
- *Génération des paraphrases (PG)*, qui crée de nouveaux textes à partir des données d'entrée et des modèles de langues (Gupta et al., 2018 ; Bowman et al., 2016).

Les méthodes basées sur des mesures de similarité textuelle (STS ou PI) comptabilisent les mots qui ont une forme similaire. Ces méthodes ont des limites quant à l'identification des paraphrases qui utilisent des mots ou des phrases avec des formes très différentes. Ainsi, nous adoptons une architecture neuronale de type paradigme contradictoire, qui vise à identifier des *différences lexicales et syntaxiques*, qui permet aussi de *générer des paraphrases* (PG) pour créer des reformulations les plus diverses possibles, en gardant le plus possible le même contenu sémantique (moins pour les reformulations de type explication). Nous considérons cette méthode la plus adaptée à notre tâche d'identification de la reformulation, vu la grande diversité des reformulations médicales identifiées dans nos corpus : paraphrases, exemplifications, synonymes, de type définition, explication ou abréviation. Nous présentons en détail cette architecture dans le point suivant.

2.1 L'architecture neuronale APT

Pour générer des reformulations à partir du terme médical, nous faisons appel à l'architecture neuronale **APT** (*Adversarial Paraphrasing Task*) (Nighojkar & Licato, 2021). Cette architecture utilise une méthode pour générer des reformulations composées avec des significations équivalentes mais présentant des différences lexicales et syntaxiques. Ce modèle vise à identifier le sens général d'une phrase, non pas uniquement le sens des mots séparés. L'architecture **APT** est construite sur deux principes :

- *La similarité de sens* : cette similarité se vérifie par le fait que deux phrases qui sont mutuellement implicites sont sémantiquement équivalentes, et sont donc, des paraphrases ;
- *La dissimilarité de la structure* : mesurée avec BLEURT (Sellam et al., 2020), un score qui évalue les textes générés automatiquement en se basant sur les plongements des mots du modèle de langue BERT (Devlin et al., 2019). Ce score évalue la similarité lexicale et syntaxique de chaque paire de phrases. BERT est adapté pour la tâche d'évaluation de la qualité de la prédiction automatique par rapport à l'évaluation humaine. Le score BLEURT utilise un ensemble réduit de données annotées par les humains pour adapter BERT à cette tâche d'évaluation. Pourtant, les données de référence utilisées pour créer BLEURT ont été générées automatiquement par plusieurs types de transformations : mot ou segment de phrase masqué et remplacé à l'aide de BERT, traduction et rétrotraduction de la phrase et suppression aléatoire des mots.

Dans leur étude, Nighojkar et Licato (2021) mènent les expériences sur des paraphrases phrastiques en anglais, appartenant au langage général. Ils utilisent un score MI (implication mutuelle) entre 2 segments, qui mesure les inférences nécessaires pour déduire le sens du premier à partir du deuxième et vice-versa, à l'aide du modèle de langue **T5-base**. Ainsi, ils sélectionnent des paraphrases qui ont un score d'implication mutuelle MI grand et un score BLEURT (score de similarité de phrases) réduit. Les expériences avec les corpus anglais annotés manuellement ont donné des meilleurs résultats par rapport aux données extraits automatiquement de Twitter. Nous testons ce modèle contradictoire d'identification de la reformulation (**APT**) sur nos données sous-phrastiques (terme médical versus reformulation), en français et en roumain, appartenant au domaine médical. Par rapport à Nighojkar et Licato (2021), nous travaillons avec des reformulations sous-phrastiques présentant peu de similarités lexicales ou syntaxiques avec l'original.

Pour lancer nos expériences de génération, nous adoptons des modèles de langues de type Transformer, en faisant appel aux données annotées manuellement, plus fiables que des jeux de données de grande taille acquises automatiquement. Nous présentons ce concept et notre méthode par la suite.

2.2 Modèle de langues de type Transformer

Un *modèle de langue* est une représentation des informations linguistiques variées qui participent à la construction du sens dans une langue donnée par des règles linguistiques et des normes d'usage spécifiques à cette langue. Ces normes sont converties en vecteurs (des valeurs numériques attribuées aux mots et séquences de mots) pour les rendre exploitables par une machine. Le modèle de langue estime la probabilité qu'un mot ou un syntagme soit présent dans une langue, en fonction du contexte. Les *Transformers* (Vaswani et al., 2017) sont des modèles de langues pour l'apprentissage automatique profond par des réseaux de neurones. Les *Transformers* traitent les données de manière séquentielle (position des mots) à l'aide de mécanismes d'auto-attention et de systèmes de type encodeur-décodeur (chaque mot reçoit un vecteur numérique).

Nous utilisons le modèle de langue **T5** (*Text-to-Text Transformer*) de Google (Raffel et al., 2020), car il contient quatre langues, dont nos langues d'étude : l'allemand, l'anglais, le français et le roumain. Ce Transformer est pré-entrainé sur C4 (*Colossal Clean Crawled Corpus)* un corpus nettoyé avec une taille colossale de 7 téraoctets, extrait du corpus Web de Common Crawl. **T5** a été pré-entrainé pour plusieurs tâches spécifiques du TAL, dont l'identification de la paraphrase et la similarité de phrases. Nous l'adaptons pour notre propre jeu de données (nous travaillons sur des reformulations sous-phrastiques au lieu des reformulations phrastiques) dans un domaine spécifique. Nous présentons notre méthode et les données exploitées pour générer des reformulations médicales dans le point suivant.

3 Méthodologie

Afin de générer des reformulations spécifiques au domaine médical, nous devons adapter le modèle de langue **T5** à l'aide de données médicales. Nos données sont des paires (terme médical, reformulation médicale), en français et en roumain, données que nous avons extrait semi-automatiquement à l'aide des ontologies et des reformulations. Ces données sont annotées manuellement, suivant la méthodologie proposée par (Buhnila, 2021 ; 2022b). Nous adaptons le modèle **T5** pour le français, ainsi que pour le roumain et nous utilisons l'architecture **APT** (Nighojkar & Licato, 2021) pour la génération de prédictions de reformulations médicales. Ces prédictions sont analysées et évaluées en suivant notre échelle de lisibilité pour élargir les corpus de reformulations. Nous détaillons ces étapes et les données utilisées par la suite.

3.1 Données médicales annotées

Pour le français nous travaillons sur deux corpus : **CLEAR Cochrane** (Grabar & Cardon, 2018) et **ClassYN** (Todirascu et al., 2012). Ces corpus écrits et comparables sont constitués des textes scientifiques du domaine médical destinés aux experts et des textes simplifiés pour le grand public. Le corpus **CLEAR Cochrane** a une taille de **4 355 054** tokens et le corpus **ClassYN** de **1 779 423** tokens. Pour la langue roumaine, nous exploitons le corpus **GrandMed-Ro** (Buhnila, 2018) constitué à partir des textes de vulgarisation extraites de la toile avec Sketch Engine (Kilgarriff et al., 2014) (6 440 951 tokens).

Nous identifions automatiquement les termes médicaux dans notre corpus français avec l'annotateur **SIFR-BioPortal** (Tchechmedjiev et al., 2018), utilisant l'ontologie médicale **SNOMED-3.5VF** (Côté, 1998) (qui contient 150 906 concepts médicaux) et des scripts en Perl pour l'extraction des phrases contenant les termes. À notre connaissance, il n'existe pas de liste de termes médicaux en roumain en libre accès. Pour cela, nous avons extrait les entités nommées médicales du corpus médical annoté **MoNERo** (Mitrofan et al., 2019) et nous avons créé une **terminologie médicale en roumain** de 14 133 termes. Notre méthode consiste dans l'extraction automatique de phrases qui contiennent simultanément des termes médicaux et des *marqueurs lexicaux ou grammaticaux de reformulation*, de type « c'est-à-dire », « autrement dit », « encore appelé », « est un », « également appelé », « désigne », signifie », etc., et leurs équivalents roumains (Grabar & Hamon, 2015 ; Antoine & Grabar, 2016 ; Buhnila, 2022a). Nous prenons en compte également les *marqueurs orthographiques* comme les doubles points ou les parenthèses.

Les phrases qui contiennent des termes médicaux et des marqueurs de reformulations sont extraites automatiquement. Ces phrases sont par la suite annotées semi-automatiquement et analysées manuellement du point de vue lexical et sémantico-pragmatique afin d'identifier des reformulations de termes médicaux (selon les critères de Buhnila, 2021 ; 2022a ; 2022b). Nous définissons les *relations lexicales* comme le lien lexical qui existe entre les deux segments, le terme médical et la reformulation. Nous analysons et annotons les relations lexicales de type *synonymie, hypéronymie, hyponymie et méronymie*[1], dans le contexte du texte médical (Condamines, 2018 ; Ramadier, 2016 ; Săpoiu, 2013). *Les fonctions sémantico-pragmatiques* représentent les raisons qui poussent le locuteur à utiliser la reformulation dans les textes écrits du domaine médical (de type *définition, paraphrase, synonymie, dénomination*[2]*, exemplification, explication*) (Malaise et al., 2004 ; Eshkol-Taravella & Grabar, 2017 ; Buhnila, 2022a). Les phrases ont été annotées par au moins deux annotateurs non-spécialistes du domaine de la médecine. L'accord inter-annotateur Kappa (Cohen, 1960) pour l'annotation des phrases contenant des reformulations valides ou sans reformulation est, pour le français, de **0,78** (**0,62** pour toutes les phrases du corpus CLEAR et **0,95** pour ClassYN), et pour le roumain, de **0,82** (pour 1 010 phrases annotées du corpus GrandMed-Ro).

Afin de construire des listes de reformulations correctes pour chaque langue, nous avons réalisé une adjudication entre les deux annotations. Lorsqu'il y avait des différences dans les deux annotations (statut *oui* versus statut *non*), nous avons choisi le statut de plus adapté de la reformulation en suivant le guide d'annotation. À l'issue de nos analyses, nous avons constitué une liste de **8 626** paires de *terme – reformulation* extraites des corpus français CLEAR Cochrane et ClassYN et **3 027** paires du corpus roumain GrandMed-Ro.

3.2 Données d'entraînement

Nous avons lancé nos expériences avec la plateforme **APT** et la version **T5-base** du Transformer qui a 220 millions de paramètres, permettant des prédictions de bonne qualité en anglais. Nous l'adaptons pour le français et le roumain en s'appuyant sur les données annotées et validées manuellement. Pour chaque expérience monolingue, nous avons utilisé trois types de données :

1. *Un corpus d'entraînement* : **8 146** paires *terme – reformulation* pour le français et **2 727** paires pour le roumain ;

[1] La relation d'holonymie n'est pas présente dans les reformulations identifiées. Nous n'avons pas traité la relation d'antonymie, car nous cherchons des reformulations qui gardent une équivalence de sens et non pas des contraires.

[2] Notre définition de la *dénomination* : le terme est reformulé à l'aide d'un autre nom (ou terme), en gardant une relation lexicale d'équivalence sémantique (la synonymie), mais sans l'intention d'expliciter ou simplifier le terme reformulé.

2. *Un corpus de validation* : l'évaluation est réalisée pendant le processus d'apprentissage sur des blocs de vingt exemples corrects de paires *termes – reformulation* ;
3. *Un corpus de test* : **480** paires *terme – reformulation* extraites aléatoirement de la liste de reformulations du corpus français et **300** paires *terme – reformulation* du corpus roumain (car nombre réduit de phrases annotées). Ces exemples n'ont pas été utilisés pour l'entraînement pour éviter les biais.

Dans la section suivante nous présentons en détail les résultats de ces expériences, la précision des prédictions correctes pour les deux langues, le français et le roumain.

4 Expériences et résultats

Nous avons modifié les paramètres de l'architecture **APT** pour avoir entre 1 et 5 prédictions de reformulations médicales pour chacun de 480, respectivement 300 termes médicaux de la liste de test. L'adaptation du modèle **T5-base** avec les reformulations a duré 24 heures pour chacune de langues de notre étude. Nous avons utilisé une taille maximale de 256 mots (pour les reformulations), le taux d'apprentissage (3e-4), 4 epochs et des batchs (entrainement et validation) de taille de 20 paires, le paramètre concernant la réduction des poids (0,01), un optimiseur AdamW (l'epsilon 1e-8). Nous évaluons manuellement chaque prédiction et nous présentons les résultats de notre analyse.

4.1 Génération des reformulations médicales

Nous avons obtenu **2 268 prédictions** de reformulations médicales générées automatiquement pour les **480** termes médicaux du corpus de test en français, et **1 490 prédictions** en roumain pour les **300** termes de test. Par exemple, le terme médical « drépanocytose » est associé à deux types d'informations : la reformulation médicale issue de nos annotations manuelles « est une maladie héréditaire du sang causée par des anomalies de la production d'hémoglobine » (sous l'intitulé *Truth,* reformulation de référence) et des prédictions automatiques (*Prediction*) : « maladie du sang génétique » ; « , maladie héréditaire de l'hémoglobine », « d'autres maladies génétiques » ; « tel que maladie génétique » ; « : maladie d'Alzheimer ». Nous observons que pour cet exemple, les quatre premières prédictions sont correctes. Parmi les 480 termes médicaux en français, **180 (37,50%)** termes ont été attribués des prédictions de reformulations correctes. Pour le roumain, **80 (26,66%)** termes ont été bien reformulés par le modèle de langue **T5-base**. La faible précision pour le roumain s'explique par la taille réduite de données disponibles pour l'entraînement (2 727 reformulations annotées par rapport à 8 146 pour le français) et par la morphosyntaxe plus complexe de la langue (déclinaisons, cas marqués morphologiquement, signes diacritiques absents dans le corpus qui a été utilisé pour construire le modèle T5).

Prédiction avec T5-base	**CLEAR et ClassYN**		**GrandMed-Ro**	
	N° termes	**Précision**	**N° termes**	**Précision**
au moins une prédiction correcte	180	37,50%	80	26,66%
aucune prédiction correcte	300	62,50%	220	73,33%
Total	**480**	**100%**	**300**	**100%**

TABLE 1 : Résultats de génération automatique de prédictions avec T5-base

Afin d'évaluer la qualité des prédictions automatiques de reformulations, nous avons créé un **guide d'annotation** sur une **échelle d'annotation** permettant de les évaluer, que nous présentons en détail ci-dessous.

4.2 Evaluation des prédictions automatiques

L'échelle d'annotation conçue pour évaluer les prédictions est construite en trois parties :

1. *La première* évalue chaque prédiction de reformulation générée :
   - La valeur 2 : pour les prédictions identiques à la reformulation médicale initiale (Truth, qui est la valeur de référence) ;
   - La valeur 1 : pour les prédictions correctes, mais différentes de la reformulation médicale initiale ;
   - La valeur 0 : pour les prédictions incorrectes. Même s'il y a des parties de reformulations correctement construites, nous avons attribué la valeur 0 en cas de mots inventés ou des mots inadaptés dans le contexte (par exemple on parle de *l'inflammation de la peau* et on utilise des symptômes des *infections urinaires*) ;
   - La valeur -1 : pour les répétitions du terme médical.
2. *La deuxième* calcule la moyenne de toutes les prédictions d'un terme ;
3. *La troisième* évalue si au moins une des prédictions générées est correcte parmi toutes les reformulations générées pour un terme médical à la fois :
   - La valeur 1 : le terme médical a au moins une prédiction de reformulation médicale correcte ;
   - La valeur 0 : le terme médical n'a reçu que des prédictions de reformulations médicales qui ne sont pas correctes.

Nous présentons notre analyse et évaluation des résultats de prédictions et nous calculons la précision de génération automatique de reformulations du modèle T5-base.

4.3 Analyse de prédictions de reformulations

Nous analysons en détail les scores donnés à chacune de 2 268 prédictions générées automatiquement par le modèle T5-base (voir Table 2). Très peu de prédictions sont identiques à la valeur de référence (Truth) : pour 73 prédictions (3,21%) le modèle a généré correctement les reformulations correspondantes pour 27 termes médicaux. Ces prédictions sont de plusieurs types :

- **Abréviations** : Terme : *troubles associés à l'entorse cervicale* ; Truth: *(TAEC)* ; Prediction: *(TAEC)* ;
- **Hypéronymes** : Terme : *maladies cardiovasculaires, l'ostéoporose et la démence* ; Truth: *telles que maladies chroniques* ; Prediction: *maladies chroniques* ;
- **Paraphrases** : Terme : *syndrome confusionnel* ; Truth: *(délirium)* ; Prediction: *(délirium)* ;
- **Dénominations** : Terme : *sclérose latérale amyotrophique* ; Truth: */ maladie du motoneurone* ; Prediction: */ maladie du motoneurone*.

Dans les cas des *dénominations*, l'apparition fréquente de la paire respective de terme – reformulation dans le corpus d'apprentissage joue un rôle important dans la précision de la prédiction générée. Par exemple, nous comptons 16 prédictions générées automatiquement de type « maladie du motoneurone » pour le terme « sclérose latérale amyotrophique », qui apparait 11 fois dans la liste de termes du corpus de test et 41 fois ensemble avec le terme dans le corpus d'entraînement. Nous remarquons également que, même si APT a été paramétré pour générer maximum 5 prédictions pour chaque terme médical, il peut trouver la reformulation correcte exacte

(selon Truth) avec un seul essai, comme par exemple pour les termes médicaux polylexicaux « essais contrôlés randomisés » ; Truth: (ECR) ; Prediction: (ECR) ou « Organisation mondiale de la Santé » ; Truth: (OMS) ; Prediction: (OMS).

Échelle d'annotation des prédictions	**CLEAR et ClassYN**		**GrandMed-Ro**	
	N° prédictions	**Précision %**	**N° prédictions**	**Précision %**
Score 2	73	3,21%	3	0,2%
Score 1	244	10,75%	117	7,85%
Scores positifs	***317***	***13,97%***	***120***	***8,05%***
Score 0	1848	81,48%	1 320	88,59%
Score -1	94	4,14%	50	3,35%
Scores négatifs	***1942***	***85,62%***	***1 370***	***91,94%***
Total	**2 268**	**100%**	**1 490**	**100%**

TABLE 2 : Statistiques sur les scores de l'échelle d'évaluation de prédictions avec T5-base

Parmi toutes les prédictions générées pour le français, **317 (13,97%)** sont des reformulations correctes pour les **180 termes** ayant au moins une prédiction correcte (**37,50%**). **T5-base** a généré également d'autres reformulations correctes que celle de la colonne *Truth* parmi ses prédictions, **244 (10,75%)** plus précisément. Ainsi, nous avons obtenu des *nouvelles reformulations*, en dehors de celles qui ont été données comme exemples lors de l'entraînement. Concernant les prédictions correctes en roumain, nous observons que la majorité (**98,25%**) de prédictions générées sont des reformulations nouvelles (*score 1*) par rapport aux données d'apprentissage, les reformulations annotées (*score 2*). Par exemple, pour le terme « Infectiile herpetice » (les infections par l'herpès), dont la reformulation initiale était « sunt afectiuni eruptive de natura inflamatorie » (*sont des affections éruptives de nature inflammatoire*), **T5-base** a généré la reformulation « cum ar fi o infectie virala contagioasa » (*telle qu'une infection virale contagieuse*). Nous avons identifié 112 (7,51%) prédictions qui contiennent des répétitions de type « boala boala » *(maladie maladie)*, « un tip tip » *(un type type)*, « cum ar fi un virus viral viral » *(comme un virus viral viral)*, « fiind afectiune cronica cronica » *(étant une affection chronique chronique)*, « afectiune afectiune » *(affection, affection)*.

Pour éliminer les répétitions du même mot dans la prédiction générée, nous avons mené une deuxième expérience (**Exp 2**) avec **T5-base** et l'architecture **APT** (**sans répétition du même mot**). Nous avons évalué les prédictions générées lors de cette deuxième expérience pour le corpus français et le corpus roumain et nous comparons les deux versions (**Exp 1** et **Exp 2**, sans répétitions) ci-dessous. Pour le français, la précision des termes avec au moins une prédiction correcte a augmenté à **44,79%** (**7,29%** de plus que dans l'**Exp 1**). Pour le roumain, nous avons identifié **19% de termes supplémentaires avec au moins une prédiction correcte** pour obtenir une nouvelle précision de **45,66%** (Table 3).

Les prédictions correctes (*score 1* et *score 2*) restent quand même limitées : **381** (**16,80%**) pour le français et **222** (**14,94%**) pour le roumain. La contrainte sans répétitions imposée à la deuxième expérience augmente le nombre de prédictions correctes (*scores positifs*) de **2,83%** pour le français et de **6,89%** pour le roumain (voir Table 4).

	CLEAR et ClassYN				GrandMed-Ro			
Modèle	**T5-base Exp 1**		**T5-base Exp 2**		**T5-base Exp 1**		**T5-base Exp 2**	
Données statistiques	**N° trm**	**%**	**N° trm**	**%**	**N° trm**	**%**	**N° trm**	**%**
au moins une prédiction correcte	180	37,50%	**215**	**44,79%**	80	26,66%	**137**	**45,66%**
aucune prédiction correcte	300	62,50%	265	55,20%	220	73,33%	163	54,33%
Total	480	100%	480	100%	300	100%	300	100%

TABLE 3 : Statistiques sur les résultats de prédictions de l'expérience 1 (répétitions possibles) et 2 (sans répétitions)

Nous avons calculé le **score inter-annotateur Kappa** (Cohen, 1960) pour **1 196 prédictions** en français et **1 234 prédictions** en roumain (générées pour **250** termes de chaque langue), annotées par deux annotateurs francophones, non-spécialistes du domaine de la médecine. Nous avons obtenu un score inter-annotateur Kappa de **0,44** pour le français et de **0,48** pour le roumain. Ces accords sont *modérés* car ils concernent quatre valeurs d'annotation différentes selon le guide d'annotation de prédictions automatiques (*2, 1, 0* et *-1*). Par conséquent, ces scores sont très précis pour chaque valeur. Nous avons calculé également le score Kappa pour les **250 termes annotés par langue**. Pour les 250 termes annotées en français le score est de **0,42** et en roumain de **0,55**, également des scores inter-annotateur Kappa *modérés*. Les scores modérés s'expliquent par la difficulté de la tâche : il est difficile d'identifier la reformulation correcte (surtout quand les mots inventés utilisent des préfixes ou suffixes utilisés couramment pour créer des termes).

	CLEAR et ClassYN				GrandMed-Ro			
Échelle d'évaluation	**T5-base Exp 1**		**T5-base Exp 2**		**T5-base Exp 1**		**T5-base Exp 2**	
	N°	**%**	**N°**	**%**	**N°**	**%**	**N°**	**%**
Score 2	73	3,21%	50	2,20%	3	0,2%	7	0,47%
Score 1	244	10,75%	330	14,55%	117	7,85%	214	14,41%
Scores positifs	***317***	***13,97%***	***382***	***16,85%***	***120***	***8,05%***	***223***	***15,01%***
Score 0	1 848	81,48%	1 796	79,22%	1 320	88,59%	1 216	81,88%
Score -1	94	4,14%	89	3,92%	50	3,35%	45	3,03%
Scores négatifs	***1 942***	***85,62%***	***1 885***	***83,14%***	***1 370***	***91,94%***	***1 261***	***84,91%***
Total	**2 268**	**100%**	**2 267**	**100%**	**1 490**	**100%**	**1 485**	**100%**

TABLE 4 : Statistiques sur les scores de l'échelle d'évaluation de prédictions : expériences 1 et 2

Nous avons analysé les prédictions annotées avec le *score 1* afin d'identifier celles qui sont des nouvelles reformulations par rapport au jeu de données d'entraînement (les 8 146 paires *terme-reformulation* pour le français et les 2 727 paires pour le roumain). Nous analysons les prédictions uniques, sans doublons, et nous observons que lors de l'**Exp 2**, **T5-base** a généré **81,55%** de nouvelles reformulations en français et **86%** en roumain (Table 5). Les nouvelles reformulations obtenues sont en général des reformulations assez simples, des variantes du nom de la maladie sous forme d'adjectif (*schizophrénie : une maladie schizophrénique*).

	CLEAR et ClassYN				GrandMed-Ro			
Modèle	**T5-base Exp 1**		**T5-base Exp 2**		**T5-base Exp 1**		**T5-base Exp 2**	
	N° ref	**%**	**N° ref**	**%**	**N° ref**	**%**	**N° ref**	**%**
nouvelle prédiction	135	**68,52 %**	84	**81,55 %**	94	**84,68 %**	172	**86 %**
prédiction entraînement	62	31,47 %	19	18,44 %	17	15,31 %	28	14 %
Total sans doublons	**197**	**100 %**	**103**	**100 %**	**111**	**100 %**	**200**	**100 %**

TABLE 5 : Statistiques sur les prédictions automatiques de score 1

Les résultats des scores négatifs sont dus également aux mots inventés ou mal orthographiés, générés par le Transformer. En français nous avons identifié seulement 8 mots inventés (0,35%) de type « maladie *nichéolaire* », « une maladie caractérisée par une *ote* de cœur ». Cependant, pour le roumain le nombre de mots inventés est beaucoup plus grand (**Exp 1** : 71 mots (4,76%) ; **Exp 2** : 129 mots (8,68%)), comme nous l'observons dans les exemples suivants : « reprezinta o afectiune in care se dezvolta un *tumefl rea* in *cortija* » (*représente une affection dont se developpe [mot inventé] mauvaise [mot inventé]*), « numiti *articulatari* » (*nommés [mot inventé]*), « este o boala a *virusesului* » (*est une maladie de [mot inventé]*).

Nous remarquons que ces mots inventés ont un lemme correct, comme dans les deux derniers exemples, « articulație » (*articulation*) et « virus » (*virus*). La difficulté pour le roumain vient de la déclinaison avec article enclitique (attaché à la fin du mot) et par la présence des cas en roumain. Dans l'exemple « este o boala a *virusesului* », la forme correcte du dernier mot serait la forme articulée, cas génitif du mot *virus*, c'est-à-dire « a virusului » (*du virus*). La forme incorrecte « a virusesului » montre que le Transformer n'a pas trouvé la bonne particule à ajouter à la fin du mot pour illustrer le cas génitif (qui exprime la possession en roumain) ou qu'il a ajouté la particule au mot anglais *viruses* (forme au pluriel de *virus*). Nous observons ce problème également avec le mot « articulatari », dont la forme correcte est « articulații » (*articulations*). L'absence de signes diacritiques du roumain (ă, î, â, ț, ș), problème récurrent dans les textes de vulgarisation sur la toile, complexifie encore la tâche du Transformer.

D'autres erreurs de génération concernent l'insertion des mots ou séquences de mots qui n'ont pas de lien avec le contexte dans la reformulation. Le terme « cholangite sclérosante primitive » (« maladie intestinale inflammatoire » (reformulation de référence)) est reformulé comme « une maladie chronique qui peut être caractérisée par des troubles cholestatiques, des antécédents et des taches causés par une inflammation des voies et des structures de la peau ». « Les structures de la peau » n'ont pas de lien direct avec les maladies intestinales. Ce type d'erreur est difficilement identifiable mais apparaît fréquemment dans les paraphrases annotées avec le *score 0*.

4.4 Analyse de la lisibilité des prédictions

Nous avons analysé le niveau de lisibilité (Laframboise, 1978 ; François, 2011) des reformulations générées automatiquement. Nous avons constitué un guide d'annotation de la lisibilité pour les non-spécialistes sur trois niveaux :

- *Niveau 1, facile à comprendre* : la reformulation médicale est plus facile à comprendre que le terme médical (il y a que des mots plus simples dans la reformulation) ;

- *Niveau 2, même complexité* : même niveau de complexité ou de technicité entre le terme médical et sa reformulation, c'est-à-dire que le sens de deux parties est difficile à comprendre par l'annotateur ;
- *Niveau 3, difficile à comprendre* : la reformulation médicale est plus complexe ou plus technique que le terme et par conséquent plus difficile à comprendre.

L'évaluation sur les deux langues est réalisée par deux annotateurs non-spécialistes de la médecine, francophones ayant comme langue maternelle le roumain. Nous limitons cette analyse seulement aux nouvelles prédictions de reformulations (*score 1*). Nous n'avons pas évalué les reformulations incorrectes (*score 0*) à cause de la présence de termes inventés et de sens perturbés (mots inadaptés au contexte).

Niveau de lisibilité	**CLEAR et ClassYN**		**GrandMed-Ro**	
	Annot 1	**Annot 2**	**Annot 1**	**Annot 2**
	N° (%)	**N° (%)**	**N° (%)**	**N° (%)**
Niveau 1	221 (66,96%)	247 (74,84%)	179 (83,64%)	193 (90,18%)
Niveau 2	72 (21,81%)	42 (12,72%)	27 (12,61%)	2 (0,93%)
Niveau 3	37 (11,21%)	41 (12,42%)	8 (3,73%)	19 (8,87%)
Total	**330 (100%)**		**214 (100%)**	

TABLE 6 : Évaluation du niveau de lisibilité des prédictions

Les résultats de l'évaluation montrent que, en moyenne, **70,90%** des prédictions en français et **86,91%** en roumain sont plus faciles à comprendre que le terme médical. Ces prédictions peuvent servir à la simplification automatique de termes médicaux.

5 Conclusion

Nos expériences préliminaires prouvent que l'architecture **APT** peut être utilisée également pour générer des **reformulations sous-phrastiques médicales** avec une précision de **44,79%** pour le français et **45,66%** pour le roumain (si l'on considère les termes qui ont au moins une bonne prédiction). Si Nighojkar et Licato (2021) ont généré un grand nombre de paraphrases de la langue générale en anglais, nos expériences sont menées sur des données du domaine médical en français et en roumain, ce qui rend la tâche plus difficile. Les résultats sont comparables entre les deux langues.

Nous avons créé des **ressources** annotées et vérifiées manuellement partageables[3] et un **guide d'annotation et d'évaluation des reformulations** issues des textes naturels et des générations automatiques. Le guide peut s'appliquer pour l'évaluation d'autres jeux de données (termes et de leur reformulation) pour d'autres domaines. Nous nous analysé les nouvelles reformulations générées automatiquement et nous avons interprété les résultats. Notre adaptation d'un seul paramètre pour éviter les répétitions montre que l'architecture **APT** permet **d'augmenter** la précision (concernant le nombre de termes ayant au moins une reformulation correcte générée automatiquement) à **45%**. Nous exploiterons les erreurs observées afin de les éviter dans des expériences futures pour améliorer nos résultats de prédictions sur les deux langues d'étude, en particulier pour éviter les mots inventés. L'outil DERIF (Namer, 2002) pourrait être utilisé pour générer des paraphrases à partir des mêmes familles lexicales ou évaluer certaines reformulations générées automatiquement. L'évaluation manuelle des prédictions générées prouve qu'elles peuvent constituer des reformulations médicales faciles à comprendre par le grand public.

[3] Les ressources, le code et le guide seront disponibles à partir de 15 juin 2023 sur la plateforme github : https://github.com/ibuhnila/refomed

Références

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. 10th International Workshop on Semantic Evaluation*; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511. ACL (Association for Computational Linguistics).

ANTOINE E. & GRABAR N. (2016). Exploitation de reformulations pour l'acquisition d'un vocabulaire expert/non expert. In *TALN 2016 : Traitement Automatique des Langues Naturelles*. Paris, France. HAL : https://hal.archives-ouvertes.fr/hal-01426816.

BODENREIDER O. (2004). The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research 32* (90001): 267D - 270. DOI : https://doi.org/10.1093/nar/gkh061.

BOWMAN S., GAUTHIER J., RASTOGI A., GUPTA R., MANNING C. D. & POTTS C. (2016). A Fast Unified Model for Parsing and Sentence Understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), p. 1466-1477.

BROCKETT C. & DOLAN W. B. (2005). Support Vector Machines for Paraphrase Identification and Corpus Construction. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, p. 1-8. http://aclweb.org/anthology/I/I05/I05-5001.

BUHNILA I. (2018). *Simplification lexicale entre les textes scientifiques et les textes de vulgarisation du domaine de la médecine.* Mémoire de Master. Université de Strasbourg, France.

BUHNILA I. (2021). Building a Corpus of Medical Paraphrases in Romanian. In *Proceedings of the The 16th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR-2021*, Iasi, p. 139-152.

BUHNILA I. (2022a). Le Rôle Des Marqueurs et Indicateurs Dans l'analyse Lexicale et Sémantico-Pragmatique de Reformulations Médicales. *8e Congrès Mondial de Linguistique Française (CMLF)*, 4-8 juillet 2022, Orléans, SHS Web of Conferences 138: 10005. DOI : https://doi.org/10.1051/shsconf/202213810005.

BUHNILA I. (2022b). Identifying Medical Paraphrases in Scientific versus Popularization Texts in French for Laypeople Understanding. In *Proceedings of the Third Workshop on Scholarly Document Processing*. Gyeongju, Republic of Korea, p. 69-79. Association for Computational Linguistics.

CARDON R. (2018). Approche lexicale de la simplification automatique de textes médicaux. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, p. 159-73. Rennes, France.

CARDON R. (2021). *Simplification automatique de textes techniques et spécialisés*. Informatique et langage [cs.CL]. Thèse de doctorat. Université de Lille. Français. ⟨NNT : 2021LILUH007⟩. ⟨tel-03343769v2⟩.

CARDON R. & GRABAR N. (2019). Automatic detection of parallel sentences in comparable biomedical corpora. In *TALN 2019*. Toulouse, France. HAL : https://hal.archives-ouvertes.fr/hal-02430446.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.,* 20, p. 27-46.

CÔTÉ R. A. (1998). Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International. Version 3.5. *Northfield, IL: College of American Pathologists*.

CONDAMINES A. (2018). Nouvelles perspectives pour la terminologie textuelle. J. Altmanova; M. Centrella; K.E. Russo. *Terminology and Discourse, Peter Lang*, p. 1-13.

COSTA R. (2005). Texte, terme et contexte. In *Actes des septièmes Journées scientifiques du réseau de chercheurs Lexicologie Terminologie Traduction*, p. 79-88. Bruxelles, Belgique.

DELÉGER L. & ZWEIGENBAUM P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, p. 2–10. BUCC '09. Suntec, Singapore: Association for Computational Linguistics.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*. http://arxiv.org/abs/1810.04805.

DONNELLY K. (2006). SNOMED-CT: The Advanced Terminology and Coding System for EHealth. *Studies in Health Technology and Informatics*, 121, p. 279-90.

ELHADAD N. & SUTARIA K. (2007). Mining a Lexicon of Technical Terms and Lay Equivalents. In *Biological, translational, and clinical language processing*, p. 49–56. Prague, Czech Republic: Association for Computational Linguistics. https://www.aclweb.org/anthology/W07-1007.

ESHKOL-TARAVELLA I. & GRABAR N. (2017). Taxinomie dans les reformulations du point de vue de la linguistique de corpus. *Syntaxe et Sémantique, vol. 18, no. 1*, p. 149-184.

FRANÇOIS T. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. Thèse de Doctorat. Université Catholique de Louvain. Louvain, France.

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 : https://doi.org/10.18653/v1/W18-7002.

GRABAR N. & HAMON T. (2015). Extraction automatique de paraphrases grand public pour les termes médicaux. In *22ème Traitement Automatique des Langues Naturelles*, 14. Caen, France.

GRABAR N. & HAMON T. (2016). Exploitation de la morphologie pour l'extraction automatique de paraphrases grand public des termes médicaux. *Traitement Automatique des Langues*, Varia, 57 (1), p. 85-109.

GUPTA A., AGARWAL A., SINGH P. & RAI P. (2018). A deep generative framework for paraphrase generation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

KILGARRIFF A., BAISA V., BUŠTA J., JAKUBÍČEK M., KOVÁŘ V., MICHELFEIT J., RYCHLÝ P. & SUCHOMEL V. (2014). The Sketch Engine: ten years on. *Lexicography 1*, p. 7-36.

KOPTIENT A., CARDON R. & GRABAR N. (2019). Simplification-induced transformations: typology and some characteristics. In *BioNLP 2019*. Florence, Italy. DOI : https://doi.org/10.18653/v1/W19-5033.

LAFRAMBOISE Y. (1978). La lisibilité : Qu'est-ce que la lisibilité ? Quels éléments rendent un texte lisible et un autre pas ? *Québec français 32*, p. 27-29.

MALAISE V., ZWEIGENBAUM P. & BACHIMONT, B. (2004). Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. In *Actes de la 11ème conférence sur le Traitement Automatique des Langues Naturelles*. Articles longs, p. 149–158, Fès, Maroc. ATALA.

MILLER G. A. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

MITROFAN M., BARBU MITITELU V. & MITROFAN G. (2019). MoNERo: A Biomedical Gold Standard Corpus for the Romanian Language. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, p. 71-79. Florence, Italy: Association for Computational Linguistics. DOI : https://doi.org/10.18653/v1/W19-5008.

NAMER, F. (2002). Acquisition automatique de sens à partir d'opérations morphologiques en français : études de cas. In *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles*. Articles longs, p. 237–246, Nancy, France. ATALA.

NIGHOJKAR, A. & LICATO, J. (2021). Improving Paraphrase Detection with the Adversarial Paraphrasing Task. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), p. 7106–7116, Online. Association for Computational Linguistics.

PECOUT A., TRAN T. M. & GRABAR N. (2019). Améliorer la diffusion de l'information sur la maladie d'Alzheimer : étude pilote sur la simplification de textes médicaux. *Ela. Etudes de linguistique appliquée*, no 195 (3), p. 325-41.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research (JMLR)*, 21(140).

RAMADIER L. (2016). Indexation et apprentissage de termes et de relations à partir de comptes rendus de radiologie. Thèse en Informatique. Université Montpellier, Français. ⟨NNT : 2016MONTT298⟩. ⟨tel-01479769v2⟩.

SAGGION H. (2017). Automatic Text Simplification. *Synthesis Lectures on Human Language Technologies* 10 (1), p. 1-137. DOI : https://doi.org/10.2200/S00700ED1V01Y201602HLT032.

SAPOIU C. (2013). *Hiponimia în terminologia medicală. Modalităţi de abordare în semantică şi lexicografie.* Piteşti, Editura Trend, 199 pages.

SELLAM T., DAS D. & PARIKH A. (2020). BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7881–7892, Online. Association for Computational Linguistics.

SPACKMAN, K. A., CAMPBELL K. E. & CÔTÉ, R. A. (1997). SNOMED RT: a reference terminology for health care. In *Proceedings of the AMIA Annual Fall Symposium*, p. 640-44.

SPECIA L., KUMAR J. S. & MIHALCEA R. (2012). SemEval-2012 task 1: English Lexical Simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task*, and Volume 2: *Proceedings of the Sixth International Workshop on Semantic Evaluation*, p. 347–355. SemEval '12. Montréal, Canada: Association for Computational Linguistics.

SHARDLOW M. (2014). A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications* 4 (1). DOI : https://doi.org/10.14569/SpecialIssue.2014.040109.

TODIRASCU A., PADO S., KRISCH J., KISSELEW M. & HEID U. (2012). French and German Corpora for Audience-Based Text Type Classification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation* (LREC'12), p. 1591–1597. Istanbul, Turkey: European Language Resources Association (ELRA).

TCHECHMEDJIEV A., ABDAOUI A., EMONET V., ZEVIO S. & JONQUET C. (2018). SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes. *BMC bioinformatics*, 19(1), 405.

VARGAS E. (2008). Un comportement de type céramique, c'est-à-dire cassant : les reformulations intratextuelles dans les émissions de vulgarisation télévisées allemandes. In *Pragmatique de la reformulation. Types de discours - Interactions didactiques*. Sous la direction de M. SCHUWER. M.-C. LE BOT, E. RICHARD, p. 21-38. Rennes: Presses Universitaires de Rennes.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

XU W., CALLISON-BURCH C. & DOLAN W. B. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, p. 1-11.