Extraire des patterns pour améliorer l'idiomaticité de résumés semi-automatiques en finances : le cas du lexique support

Abdelghani Laifa, Laurent Gautier, Christophe Cruz

Extraire des *patterns* pour améliorer l'idiomaticité de résumés semi-automatiques en finances : le cas du lexique support

Abdelghani LAIFA*, Laurent GAUTIER**,
Christophe CRUZ***

* TIL EA4182, LIB EA7534, Université Bourgogne Franche-Comté,
F-21000, Dijon, France
Abdelghani_laifa@etu.u-bourgogne.fr
**MSH EA4182, TIL EA4182, Université Bourgogne,
F-21000, Dijon, France
laurent.gautier@u-bourgogne.fr
*** LIB EA7534, Université Bourgogne Franche-Comté,
F-21000, Dijon, France
christophe.cruz@ubfc.fr

Abstract. This paper presents a work aiming at developing a semi-automatic drafting system for summaries of economic and financial texts, paying particular attention to the idiomaticity and fluency of the target language. To do so, the study starts from the analysis of a corpus of periodical reports of the Banque de France. Linguistic work shows that the writing of summaries that focuses solely on terminological and collocational extraction ignores a whole range of vocabulary, which is captured here as a "support lexicon", playing an important role in the cognitive organization of the field. On this basis, this work in deep learning discusses the relevance of our lexico-grammatical pattern extracting method using the self-attention mechanism, and its impact on guiding CamemBERT abstractive summarization model through data augmentation. A first experimentation using the corpus under consideration and focusing the extraction method is presented.

1. Contexte et objectifs

Les travaux présentés dans cet article s'inscrivent dans le cadre d'un projet de thèse à l'Université de Bourgogne mené à l'interface entre sciences du langage (linguistique de corpus, traitement automatique des langues naturelles) et science des données (apprentissage automatique). D'un point de vue linguistique, les travaux récents sur les discours spécialisés mettent très clairement l'accent sur les limites des approches partant strictement des mots en tant qu'unités isolées. De même, les progrès récents dans le domaine de la rédaction automatique de résumés grâce aux méthodes d'apprentissage automatique basées sur les réseaux de neurones profonds et sur le principe d'auto-attention démontrent l'intérêt de considérer les mots dans leurs environnements proches et éloignés.

Ainsi, la question que nous posons ici est celle de savoir comment extraire les *patterns* de mots dans leur environnement proche et comment ces *patterns* améliorent l'idiomaticité des résumés auto-

matiques. Le cadre d'application choisi pour cette étude est celle du domaine des finances avec une idiomaticité propre. Ce faisant, elle questionne du même coup, comme dans l'étude de cas qui va suivre autour de ce que nous proposons de dénommer 'lexique support,' les limites d'entrées strictement terminologiques et/ou collocationnelles des discours spécialisés.

La démonstration qui va suivre part de l'exploration d'un corpus de rapports de conjoncture de la Banque de France, des textes sériels associant des statistiques diverses présentées sous forme de tableaux à du texte scientifique plus ou moins vulgarisé (section 2). L'extraction sur la base des approches les plus usuelles en terminologie, comme l'approche par concordance de termes, de collocations ou autres combinatoires qualifiées ici de termino-centrées montre vite qu'une partie importante du lexique est délaissé alors même qu'il s'agit de vecteurs langagiers fondamentaux de l'information spécialisée à transmettre : la part d'incertitude relevant des modalités épistémiques et la dimension aspectuelle liée au scénario de comparaison qui est à la base de ces textes. Nous considérons donc ceci comme une limite de ces approches (section 3) et explorons les possibilités de prise en charge et de modélisation de ce lexique support et des informations qu'il transmet, en particulier en vue de la rédaction automatique de résumés par approche abstractive. Ainsi, nous présenterons un modèle de rédaction automatique basé sur l'auto-attention « *self-attention* » pour la rédaction extractive de résumés (section 4). Sur la base de ces éléments, le chapitre se clôt par notre méthode d'extraction du lexique support à l'aide du mécanisme d'attention, ainsi que l'extraction des schémas lexico-grammaticaux. Ces éléments extraits seront les paramètres de notre méthode d'augmentation des données permettant l'ajustement fin du modèle de rédaction de résumé par approche abstractive impactant l'idiomaticité des résumés générés et latente dans les patterns et dans le lexique. Pour terminer, nous avons procédé à une première étude d'extraction du lexique et des patterns (section 5).

2. Corpus

Le corpus interrogé ici relève d'une catégorie de discours pouvant être qualifiés de « discours de conjoncture » (Gautier 2012, Desmedt *et al* 2020) et produits, entre autres, par les Banques Centrales des États. Il s'agit globalement d'une publication dite sérielle, c'est-à-dire périodique présentant un très fort degré de récurrence : au niveau des contenus eux-mêmes qui visent les grands « thèmes » constitutifs du domaine et au niveau de la forme à travers des articles relativement brefs et organisés en courts paragraphes avec une macrostructure très apparente en facilitant la lecture rapide.

Le corpus à la base de ce chapitre est en français et reprend la publication de la Banque de France intitulée *Bulletin de la Banque de France*. Il a été compilé à la Maison des Sciences de l'Homme de Dijon par Hédi Maazaoui et couvre les années 1994 à 2020. Sur le temps de collecte, la forme même du *Bulletin* a évolué : d'un document unique d'une centaine de pages (parfois plus) publié le mois suivant la période-objet, il est passé à partir de mars 2018 à une publication d'articles individuels au fil de l'eau. En termes de contenu, une évolution notable doit être précisée : le *Bulletin* est passé d'une publication très figée correspondant aux phases de synthèse-bilan et prospective de l'activité économique et financière du mois écoulé suivie d'études ponctuelles et de tableaux statistiques à une publication « thématique » ne comptant précisément plus que des articles de fonds. Ce type de rapports a donné lieu à diverses études linguistiques, en particulier d'inspiration discursive (Resche 2003, Gautier 2012, Palmieri *et al.* 2015) pour lesquelles ce changement de nature est méthodologiquement important. L'unité et l'homogénéité des objets spécialisés traités tout comme celle des terminologies afférentes nous permettent néanmoins, pour la problématique esquissée en [1], de travailler sur l'intégralité de la collecte.

D'un point de vue quantitatif, ce corpus se caractérise comme suit :

Empan chronologique	1994-2020
Nombre de rapports	323
Nombre de mots/tokens	6,554,396
Nombre de lemmes	4,070,955
Nombre de phrases	317,076

TAB. 1 – *Caractérisation quantitative du corpus interrogé*

3. Limites de l'extraction de *patterns* termino-centrés

Les travaux récents sur les discours spécialisés mettent très clairement l'accent sur les limites des approches partant strictement des mots en tant qu'unités isolées. On a là une illustration évidente de l'extension quasi continue du champ de la « phraséologie » (Granger 2008, Legallois *et al* 2013, Legallois *et al*. 2018) visant à saisir les récurrences de segments supérieurs au mot. Pour la modélisation de connaissances en vue d'applications relevant des industries de la langue, trois approches holistiques ont, ces dernières années, marqué l'évolution de la recherche dans ce domaine :

- La théorie des scénarios (*frame semantics*, *cf*. Ziem 2014), qui a connu une implémentation spécifique en terminologie (Faber 2012), est un paradigme de linguistique cognitive visant une représentation organisée des connaissances liées à un concept, résultat de l'expérience du locuteur et trouvant un encodage particulier dans la langue à travers combinatoires et réalisations syntaxiques préférentielles. Ici, le *frame* de COMPARAISON entre états successifs de différents indicateurs économiques en est un exemple saillant.
- Les modèles basés sur les schémas/*patterns* lexico-grammaticaux qui bouleversent la vision traditionnelle de modules lexicaux, saisis sous forme de dictionnaires, sur lesquels opéreraient *toutes* les règles de grammaire de la langue considérée (Kopaczyk *et al* 2018). Non seulement le type de texte contraint étudié ici ne met en œuvre qu'un répertoire restreint des règles du français, mais il ne le fait qu'en synergie avec le lexique concerné. Nous suivons ici la définition des *patterns* comme blocs de sens (Gledhill *et al* 2016, 75) :

  "The typical linguistic features of ESP cannot be characterised as a list of discreet items (technical terminology, the passive, hedging, impersonal expressions, etc.), rather the most typical features of ESP texts are chains of meaningful interlocking lexical and grammatical structures, which we have called lexico-grammatical patterns".

  Ces schémas sont les garants d'une double idiomaticité dans un tel corpus : idiomaticité de la langue, idiomaticité du domaine. Ce sont eux qui permettent au(x) spécialiste(s) de transmettre l'information du domaine sans ambiguïté.
- Les grammaires de construction qui représentent en quelque sorte le degré d'abstraction et de généralisation (Goldberg 2005) ultime des *frames* et permettent de modéliser avec un haut degré de granularité l'interface syntaxe-sémantique. Ces grammaires étant basées sur l'usage, elles trouvent aussi leur point de départ dans les récurrences de type *patterns* décrites ci-dessus.

Il s'agit de montrer ici qu'une mise en œuvre de ce programme ne peut s'en tenir à une entrée strictement terminologique, mais doit, justement pour l'idiomaticité/ fluidité des résumés, intégrer aussi du lexique support défini, en première approximation, comme du lexique ne relevant pas de la terminologie du domaine mais « s'appliquant », à travers des *patterns,* propres aux autres structures terminologico-collocationnelles. Les étapes d'extraction discutées ici permettent de le mettre en évidence.

3.1. De l'extraction terminologique aux collocations

Le traitement sous SketchEngine du corpus présenté en [2] en utilisant le module d'extraction des mots clefs simples et complexes (Keywords) montre, sans surprise, une prédominance de N et de A ainsi que le résument les tableaux 2 et 3 reprenant les 30 premiers candidats termes identifiés par l'outil. La présence d'un certain nombre d'A en haut de classement (TAB. 2) (*financier, monétaire, annuel*) rend clairement nécessaire la prise en compte des termes complexes où l'on retrouve des termes composés (au sens strict) et des collocations (TAB. 3).

banque	Poste	billet
France	million	BCE
euro	net	direction
compte	crédit	marché
financier	cadre	devise
monétaire	eurosystème	milliard
titre	résultat	directeur
opération	taux	bancaire
annuel	réserve	exercice
zone	risque	central

TAB. 2 – *Liste des 30 candidats termes simples les plus fréquents*

zone euro	produit net	autre titre
politique monétaire	direction générale	opération principale
rapport annuel	code monétaire	crise financière
banque centrale	contrôle prudentiel	réserve obligatoire
compte annuel	fond propre	position nette
conseil général	banque centrale européenne	clientèle institutionnelle
stabilité financière	autre engagement	audit interne
autre produit	revenu fixe	banque centrale nationale
solde net	fond propre	moyenne annuelle
revenu monétaire	principe comptable	système financier

TAB. 3 – *Liste des 30 candidats termes complexes les plus fréquents (sont soulignés les A repérés dès l'extraction des candidats termes simples*)

Un premier traitement de ces résultats sur la base des techniques terminologiques classiques consiste à extraire pour ces termes les combinatoires récurrentes pour en mettre en évidence l'environnement collocationnel. Au vu de l'objectif final du travail en matière de rédaction semi-automatique de résumés, ces collocations peuvent être envisagées comme véhiculant des informations clefs du texte-source. Pour s'en tenir au seul domaine verbo-nominal au cœur de cette contribution, on identifie plusieurs schémas récurrents correspondant à un *frame*/scénario prototypique de ces textes, celui de la COMPARAISON puisque l'apport informationnel majeur de ces derniers réside les mouvements enregistrés pour différents indicateurs. C'est ce qu'indiquent les exemples suivants :

(1) Baisse : La capacité des entreprises à honorer leurs engagements financiers, évaluée par la cotation Banque de France, semble s'améliorer, après **avoir diminué** tendanciellement depuis la crise.

(2) Hausse : En revanche, les bons du Trésor ont continué d'**augmenter** au même rythme que précédemment (+ 16,5 % à fin février).

(3) Stabilité : Le solde des services apparaît stable d'un mois à l'autre (+ 6,9 milliards de francs, au lieu de + 7,1 milliards).

L'extraction systématique de ces structures permet de relever l'ensemble des constructions, au sens technique précisé ci-dessus, recourant à des collocations spécialisées plus ou moins saturées lexicalement et devant apparaître dans le résumé automatique compte tenu du rôle cognitif du scénario dans le domaine de spécialité. On distinguera ainsi, pour ne donner qu'un exemple, les structures à verbe plein (4) de celles à verbe support en (5) ou encore des structures attributives en (6) :

(4) X (X = indicateur) *diminue, baisse, augmente, croît, se redresse…*

(5) X (X = indicateur) *enregistre / connaît / affiche une baisse, une hausse…*

(6) X (X = indicateur) *est / paraît/ apparaît stable…*

3.2. Le poids du lexique support

Ce que montre toutefois l'examen systématique de ces combinatoires, c'est la présence simultanée et quasi systématique de deux autres types de constructions, que nous considérons être de niveau supérieur, c'est-à-dire liées non pas au domaine en tant que tel, mais au type de texte lui-même et qui « opèrent » pour ainsi dire sur les constructions précédentes : elles relèvent d'une part du marquage de l'aspectualité et d'autre part des modalités épistémiques. N'étant pas à proprement parler terminologiques, elles échappent à une approche strictement collocationnelle.

Le scénario de COMPARAISON évoqué dans la sous-section précédente suppose la mise en relation des valeurs de l'indicateur économique concerné à deux moments t et t-1. Par-delà les mouvements saisis dans les exemples (4) à (6), les textes regorgent d'éléments de mise en relation de ces mouvements avec les tendances relevées lors de la période précédente : ce marquage, qui vise le découpage chronologique interne des mouvements, peut ainsi être assimilé à la catégorie linguistique de l'aspect. C'est ce qui apparaît des exemples (7) avec un verbe support explicite ou (8) dans le choix même du verbe support (*rester* vs. *être*) :

(7) En revanche, les bons du Trésor **ont continué** d'**augmenter** au même rythme que précédemment (+ 16,5 % à fin février).

(8) L'encours des livrets A et bleus est **resté stable** (après une hausse de 0,2 % le mois précédent)

A un autre niveau, on relève un ensemble de marqueurs linguistiques de l'incertitude et / ou du pronostic liés à la nature même du type de texte, comme en (9):

(9) Depuis 2012, la limite de LTV a été réduite de 1% par an, de 106% initialement à 101% à partir de janvier 2017, et **devrait** diminuer à 100% en 2018.

Il s'agit, cognitivement, d'un scénario de nature EPISTEMIQUE directement lié au scénario de COMPARAISON discuté ci-dessus. En effet, ces rapports présentant, après le bilan par rapport à la période t-1, un pronostic sur l'évolution de l'indicateur, il y a une part intrinsèque d'incertitude liée à la prévision qui vient s'emboîter sur la construction liée au mouvement, ici de baisse (*diminuer*). Linguistiquement, ce scénario se traduit par une série de marqueurs spécifiques, essentiellement les verbes *pouvoir* et *devoir* au conditionnel qui seront autant de points d'attention à intégrer au système de résumé automatique.

Notre proposition ici est de considérer ces deux niveaux comme relevant d'un lexique-support qui ne saurait être ignoré parce que non strictement terminologique. C'est précisément l'apport des trois méthodologies listées en début de section que d'en avoir montré l'importance. Importance dont il convient de tenir compte dans le développement de l'outil de résumé. Ainsi, la qualité de l'idiomaticité d'un résumé requière la formalisation d'un lexique support couplé à des schémas/*patterns* lexico-grammaticaux. Nous proposons une méthode en deux phases, dont seule la première sera présentée ici, consistant dans un premier temps à extraire le lexique support et les *patterns* à l'aide d'un modèle de langage pour la génération de résumé formé par le corpus présenté précédemment. Dans un deuxième temps, le lexique et les *patterns* sont exploités pour produire un corpus textuel artificiel où l'accent est mis sur l'usage du lexique en accord avec les *patterns*. Le principe de l'augmentation des données (van Dyk *et al* 2001) repose sur le principe d'augmenter de façon artificielle les données lorsque celles-ci ne sont pas suffisamment abondantes pour former le modèle de réseau de neurones.

4. Rédaction automatique des résumés

La rédaction automatique de résumés est une technique de traitement automatique des langues visant à condenser un document ou un ensemble de documents à l'aide d'un sous-ensemble de phrases représentatives extraites ou générées. Les méthodes les plus récentes et efficaces sont basées sur des modèles d'apprentissage profond. Ainsi, les méthodes de génération de résumés se caractérisent par deux approches différentes. L'approche extractive de texte : ici, le modèle résume de longs documents et les représente dans des phrases plus courtes et plus simples dont le nombre est un paramètre de l'algorithme. L'approche abstractive : le modèle produit un résumé du document à partir de l'information latente du document. Les architectures de base des réseaux de neurones qui permettent d'apprendre ce type de tâches sont les architectures Seq2Seq (encodeur-décodeur avec avec attention) (Sutskever *et al.* 2014), les réseaux de neurones récurrents LSTM (RNN) (Hochreiter *et al* 1997), les modèles BERT (Devlin *et al.* 2019), et Transformer (Vaswani *et al.* 2017) ainsi que le mécanisme d'attention (Vaswani *et al.* 2017).

Dans les travaux préliminaires que nous présentons ici pour la première phase de notre méthode, nous avons concentré nos efforts sur l'approche extractive et le modèle BERTSUM ainsi que sur son mécanisme d'auto-attention où chaque mot d'une phrase se caractérise par son contexte. Ce mécanisme est crucial pour identifier à la fois le lexique support et les *patterns* lexico-grammaticaux.

4.1. Le modèle BERT et sa variante BERTSUM pour les résumés

L'acronyme BERT (Devlin *et al.* 2019) signifie "Bidirectional Encoder Representations from Transformers". Ce modèle de réseau neuronal est préformé à l'aide de larges corpus textuels pour produire des services de traitement du langage naturel. L'entraînement d'un modèle BERT est coûteux, car les modèles requièrent plusieurs gigaoctets de texte pour les former, et d'autre part, la formation des modèles lors de l'entraînement nécessite des machines puissantes et des temps de calcul longs. Le modèle BERT a été est pré-entraîné sur plus de 16 Go de données contenant 3,3 milliards de mots. La version de BERT la plus volumineuse a nécessité 4 jours complets et a mobilisé 64 unités TPU Google pour finaliser son entraînement. Toutefois, une caractéristique avantageuse de ces modèles pré-entraînés est d'offrir la possibilité d'ajuster ces modèles en les entrainant sur un petit jeu de données spécifique pour une tâche spécifique (ici la rédaction automatique des résumés) afin d'obtenir des résultats plus intéressants sans qu'il soit nécessaire de recourir à des machines puissantes. Ainsi, la communauté scientifique partage ses modèles que les chercheurs et industriels peuvent tester et adapter à leur besoin. L'étape qui consiste à entraîner un modèle pré-entraîné sur un petit jeu de données se nomme la phase de réglage fin, formation fine ou « fine-tuning ».

Le modèle BERT a été adapté pour l'extraction de résumés en intégrant notamment une information positionnelle des phrases. Ce nouveau modèle se nomme BERTSUM. L'image ci-dessus montre de légères différences entre le modèle d'origine et le modèle utilisé pour le résumé.

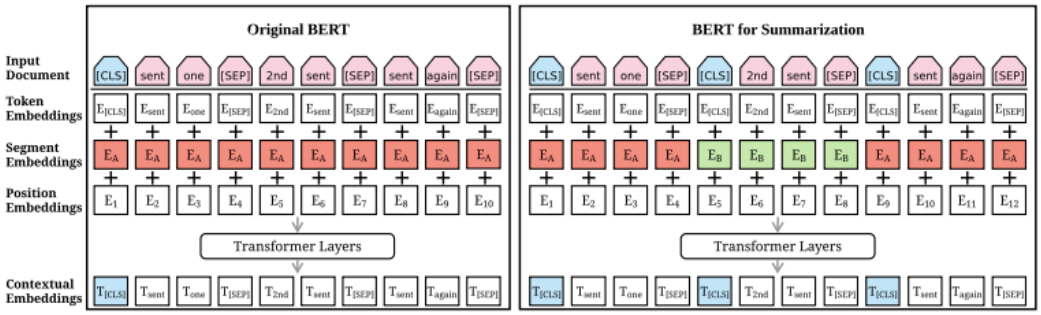


FIG. 1 – *Les différences entre le modèle BERT original et BERTSUM. (Liu 2019)*

BERTSUM (Liu 2019) utilise la méthode d'enchâssement des mots (Levy *et al* 2014) (*word embeddings*) produisant un vecteur pour chaque mot du document afin d'alimenter le modèle. Ici, le jeton [CLS] est ajouté au début de chaque phrase notifiant l'algorithme du début de phrase. Il existe également une différence dans l'enchâssement des segments ou phrases. Chaque phrase se voit attribuer une incorporation de Ea ou Eb selon que la phrase est paire ou impaire. Si la séquence de segments est [s1, s2, s3] alors les enchâssements de segment sont [Ea, Eb, Ea]. Ainsi, l'enchâssement contextuel d'un mot contient l'enchâssement du mot auquel est concaténé l'enchâssement de sa position et l'enchâssement du segment. BERTSUM attribue des scores à chaque phrase qui représente la valeur que cette phrase ajoute à l'ensemble du document. Ainsi [s1, s2, s3] se voit attribuer [score1, score2, score3]. Les phrases avec les scores les plus élevés sont ensuite collectées et réorganisées pour donner le résumé global de l'article.

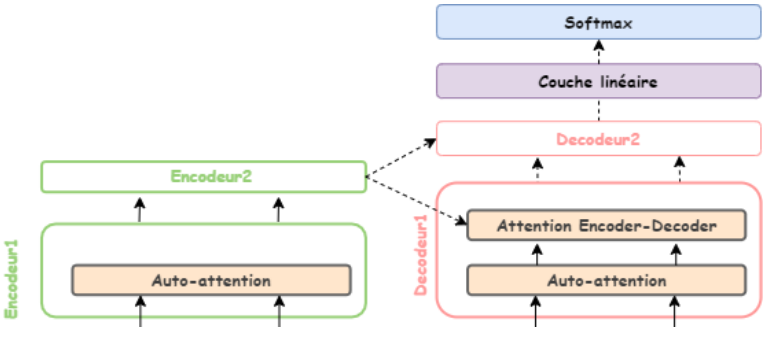


FIG. 2 – *Architecture de l'encodeur-décodeur du Transformer*

Le modèle BERT s'appuie sur le composant « Transformer » (Vaswani *et al.* 2017) exploitant le mécanisme d'auto-attention – dont l'objet est d'analyser une séquence d'entrée – et décide à chaque étape quelles autres parties de la séquence sont importantes. Le modèle BERT est constitué d'un encodeur et d'un décodeur (Figure 2).

Un encodeur se compose d'une couche d'auto-attention reliée à l'encodeur précédent ou l'entrée pour le premier (enchâssement contextuel des mots), et des réseaux de neurones à couches denses (cadre bleu nommé FF sur la figure 3). Pour chaque entrée de l'encodeur, l'auto-attention prend en compte plusieurs autres entrées en même temps et décide lesquelles sont importantes en leur attribuant des poids différents.

Un décodeur se compose, de manière similaire, de couches d'attentions reliées au décodeur précédent et de couches denses, la particularité du décodeur étant d'être composé aussi d'une couche d'attention supplémentaire intercalée entre ses deux couches et reliée à la sortie du dernier encodeur. En d'autres termes, le décodeur prendra alors en entrée la phrase codée et les poids fournis par le mécanisme d'auto-attention de la couche du décodeur.

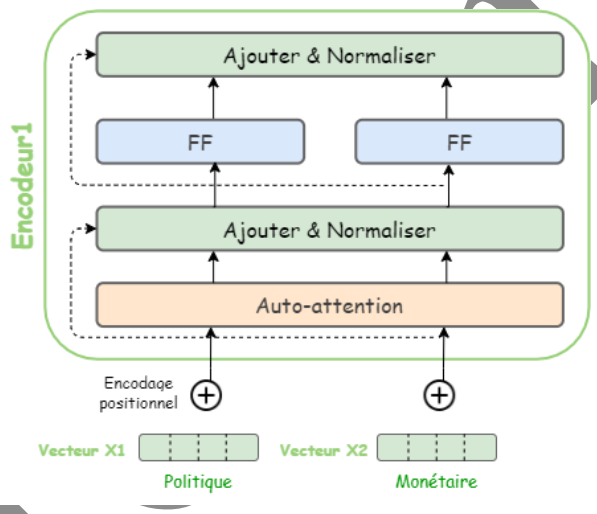


FIG. 3 – *Architecture de l'encodeur*

Le réseau de neurones FF est un réseau de neurones à propagation avant où les vecteurs ne se déplacent que dans une seule direction : à partir des nœuds d'entrée vers les nœuds de sortie.

4.2. Le mécanisme de « auto-attention »

Le mécanisme d'auto-attention a été introduit dans l'article « Attention is all you need » (Vaswani *et al* 2017). Ce mécanisme d'attention est composé de plusieurs couches d'attentions ou *layers* ; chaque couche comportant plusieurs têtes d'attentions *heads,* chaque tête se charge d'attribuer un poids d'attention en fonction du contexte de la phrase.

Considérant la phrase S « Les taux d'inflation devraient demeurer faibles avant de commencer à ***se redresser*** », la question se pose de savoir à quoi réfère le verbe *se redresser* ? Si la question est simple pour un humain disposant du *frame* spécialisé correspondant, elle n'est pas aussi simple pour un algorithme. Lorsque le modèle traite le verbe *se redresser*, l'auto-attention lui permet d'associer *se redresser* à *taux d'inflations*. Par ailleurs, la phrase S contient deux autres éléments identifiés en [3] comme relevant du lexique-support : le verbe *demeurer* qui transmet une information de durée contextuelle (les dits taux étaient faibles dans la période précédente, n'ont pas remonté et devraient le rester) et le verbe *devoir* au conditionnel relevant des marqueurs épistémiques et encodant l'état de fait visé comme relevant de la conjecture.

Au fur et à mesure que le modèle traite chaque mot et chaque position dans la séquence d'entrée, l'auto-attention « regarde » de son côté les autres positions dans la séquence d'entrée pour trouver des indices pouvant aider à améliorer la représentation des mots.

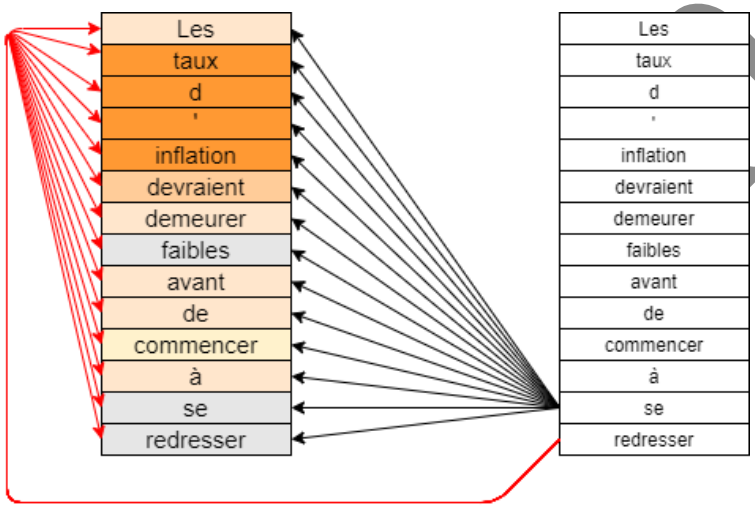


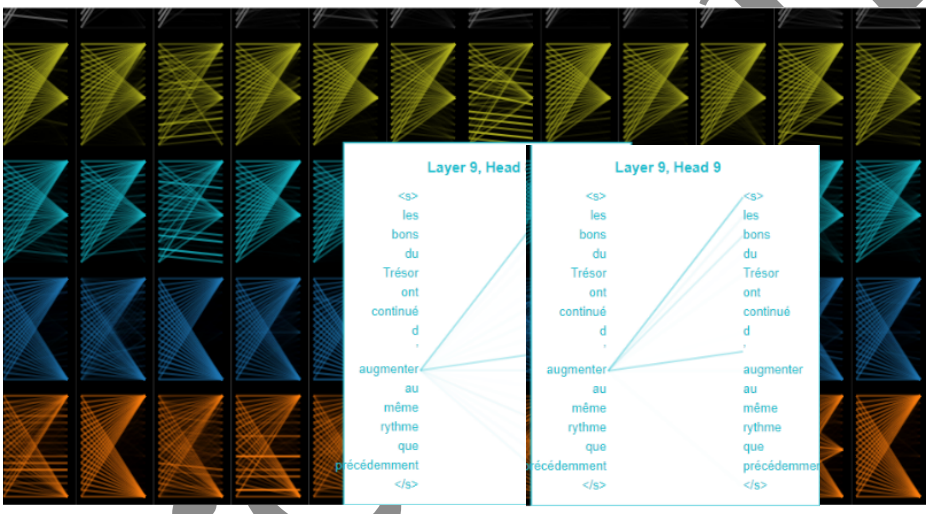
FIG. 4 – *Le mécanisme de "Self-Attention"*

Durant la phase d'entrainement, l'auto-attention permet au modèle d'associer à chaque mot de la phrase un autre mot de la même phrase. Cette association s'exprime par un poids, qui est à son tour présenté par l'intensité de la couleur (FIG 4) : cette dernière illustre le mécanisme d'attention appliqué sur chaque mot de la séquence, les flèches noires et rouges présentent l'attention portée par #se et #redresser respectivement, vers tous les autres mots de la phrase. On constate que les jetons « #se #redresser » portent plus d'attention sur les jetons « #taux #d #' #inflation », cela s'explique par le fait que durant le traitement des mots, l'auto-attention parvient à déterminer le contexte de la séquence et réussit à référer le verbe « se redresser » à son sujet « taux d'inflation ». Ainsi, l'auto-attention est un mécanisme d'attention reliant différentes positions d'une séquence afin de calculer une représentation contextuelle de la même séquence.

4.3. Extraction du lexique support et visualisation de l'attention

Supposons que chaque rapport est composé de X phrases : en attribuant à chaque mot de chaque phrase un certain poids d'attention qui évoluera au cours de l'entrainement du modèle (FIG. 5), nous constatons à un temps t que l'attention attribuée par exemple à « les bons du trésor » dans la (couche 9, tête 8) est moins pondérée que l'attention attribuée dans la (couche 9, tête 9). Cette pondération est représentée par l'intensité de la couleur de la ligne connectant deux mots. Ainsi, nous exploitons cette pondération et le mécanisme d'auto-attention pour extraire les schémas lexico-grammaticaux. Dans cette figure, nous pouvons constater qu'au fil de l'entrainement, le modèle accorde à #bon #du #trésor plus d'attention au verbe #augmenter montrant ainsi la capacité de ce mécanisme à comprendre le contexte de la phrase. Ce mécanisme d'auto-attention est appliqué sur tous les mots de toutes les phrases et il est représenté en image de fond de la figure 5. Nous pouvons extraire manuellement les schémas lexico-grammaticaux et le lexique support dans chaque rapport mensuel de la banque de France avec un seuil de pondération afin de réduire le nombre de cas à étudier.

FIG. 5 – *Identification du lexique support grâce au mécanisme d'auto-attention*



5. Expérimentation

Le premier modèle de langage de type BERT en français se nomme CamemBERT. Il a été entraîné sur le corpus OSCAR (Javier *et al.* 2019), une section française du jeu de données CommonCrawl (Nils *et al.* 2019). Ce corpus est composé de textes provenant d'une grande quantité de pages Web. C'est lui qui a été exploité pour notre expérimentation.

Cette section présente comment le modèle de langage CamemBERT est formé à l'aide du corpus. Le résultat de cette formation et les résultats de l'auto-apprentissage permettront d'extraire le lexique support et les *patterns*. Couplé à un seuil de pondération des liens d'attention, nous ajoutons la méthode PCA pour visualiser les phrases classées pour les résumés afin de procéder à une analyse linguistique des phases avec les sémantiques latentes les plus proches et dont les pondérations sont les plus élevées.

5.1 Rapports de la banque de France

Avant de commencer l'entrainement du modèle, la première étape consiste à extraire le texte essentiel des rapports au format PDF et les enregistrer au format texte avec l'encodage Unicode grâce à PdfMiner (Yusuke 2007).

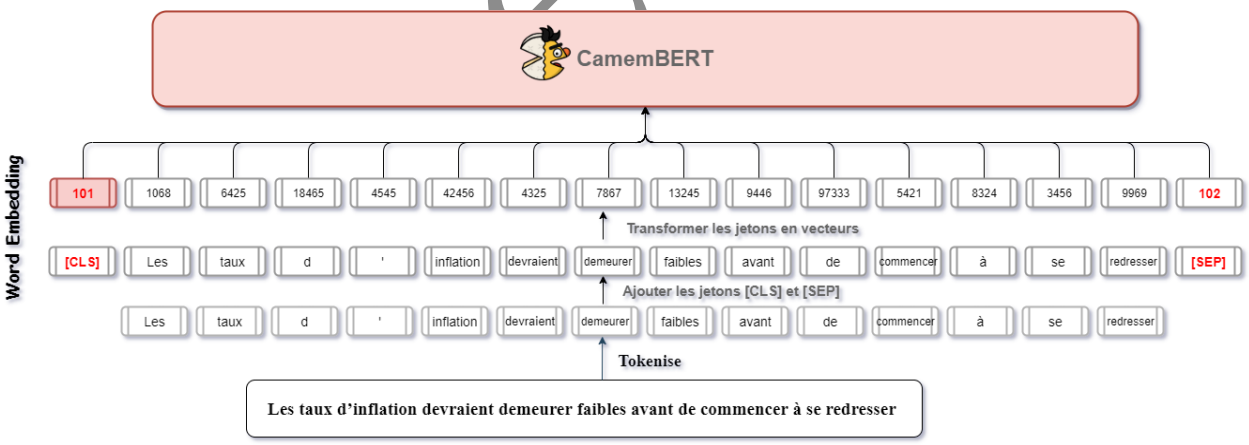
Rapport
Les canaux de transmission de la politique monétaire — à savoir les mécanismes par lesquels une décision de politique monétaire, en affectant le comportement des agents économiques, agit sur la croissance et les prix — sont des processus complexes. Les effets des variations de taux d'intérêt sur les variables macroéconomiques ne sont pas systématiques, mais dépendent de l'état de l'économie et des anticipations des agents. Par ailleurs, ils ne s'exercent qu'avec des délais assez longs, généralement évalués entre quatre et six trimestres. Les différentes études disponibles, dont l'important travail réalisé par les banques centrales de l'Eurosystème en 2001, ont montré que, dans la zone euro, les réponses aux impulsions de politique monétaire, avec pourtant des profils relativement proches de ceux observés aux États-Unis, avaient un impact sur l'activité et les prix sensiblement moins élevés 1.

TAB. 4 – *Exemple de texte extrait*

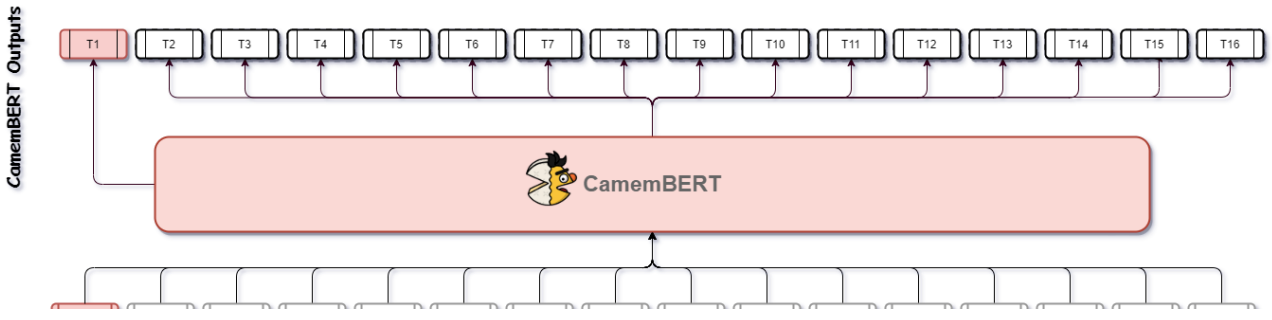
5.2 Traitement des données

La deuxième étape consiste à utiliser le CamemBERT Tokenizer pour d'abord diviser le texte en jetons uniques définissant ainsi un dictionnaire. Ensuite, nous ajoutons les jetons spéciaux nécessaires pour les classifications de phrases dont le jeton [CLS] à la première position et [SEP] à la fin de la phrase. La troisième étape remplace chaque jeton par son identifiant dans la table d'enchâssement, composant obtenu avec le modèle pré-entrainé CamemBERT. Ce traitement est appliqué à toutes les phrases du rapport. La figure suivante présente les différentes étapes présentées.

FIG. 6 – *Le processus de traitement de données et de création de la séquence de vecteurs*



L'état suivant consiste à ajuster le modèle avec le corpus de rapports de la Banque de France en appliquant l'auto-attention. Les phrases d'entrée ont maintenant la forme appropriée pour être insérée



dans le modèle CamemBERT. Le passage du vecteur d'entrée via CamemBERT fonctionne exactement comme BERT. La sortie du modèle CamemBERT est composée des vecteurs de probabilité contenant les caractéristiques de chaque jeton d'entrée. Le premier vecteur représente le jeton 101 qui correspond à [CLS].

FIG. 7 – *Le passage des entrées par CamemBERT*

Les vecteurs [T2-T16] sont des scores calculés par le modèle CamemBERT, ces scores expriment la représentativité de chaque jeton au sein de la séquence, tandis que T1 représente le contexte de toute la séquence.

5.3 Classification des données

Comme il s'agit à la fin d'une tâche de classification de phrases, nous ignorons tout sauf le premier vecteur associé au jeton [CLS]. Ce jeton capture le contexte de la phrase transmis par l'auto-attention. Nous transmettons tous les premiers vecteurs de chaque phrase du texte comme entrée du modèle de classification. Ce dernier classifie chaque phrase selon sa similarité sémantique. Ainsi le résultat de ce classement offre un sous-ensemble de phrases utilisant le lexique support et les *patterns* lexico-grammaticaux objet de l'analyse linguistique.

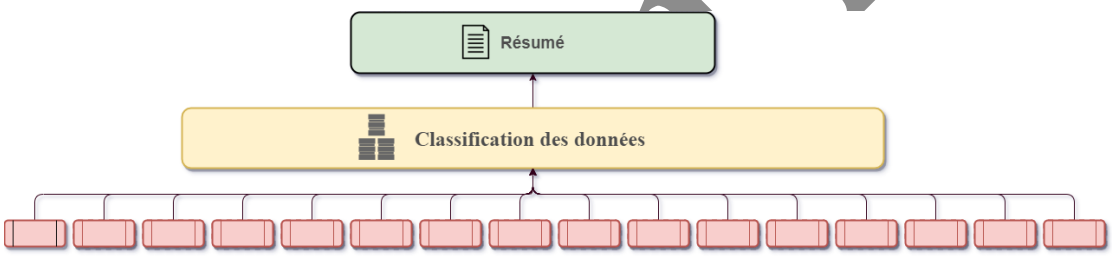


FIG. 8 – *Le passage des jetons [CLS] de toutes les phrases à travers le classificateur des données*

5.4 Visualisation des données

Une fois que nous avons les vecteurs d'enchâssement de nos phrases, nous utilisons la méthode PCA (Principal Component Analysis) pour réduire la dimensionnalité du vecteur de la phrase de dimension 768 à 2. Ensuite, les vecteurs à deux dimensions sont tracés et forment un nuage de points illustré sur la figure suivante. Ainsi, les phrases qui partagent le même contexte sont proches dans l'espace à deux dimensions. Nous pouvons remarquer deux groupes denses de points.

Grace à cette visualisation, nous considérons que les *patterns* qui sont proches partagent un lexique support plus ou moins proche, Nous pouvons donc extraire les schémas lexico-grammaticaux qui partagent un contexte commun.

La première partie d'identification du lexique support avec le mécanisme d'auto-attention, et la deuxième partie de visualisation PCA qui projette les schémas lexico-grammaticaux qui ont un sens commun, sont les deux parties de notre première solution d'extraction du lexique support, sachant que le lexique support et les patterns sont des composants essentiels pour améliorer l'idiomaticité.

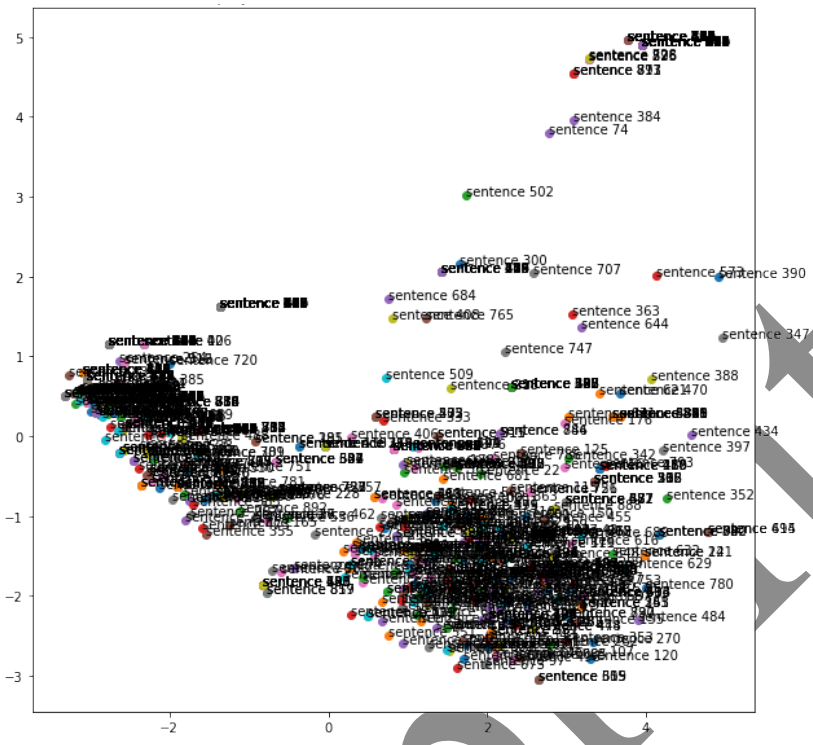


FIG. 9 – *Projection des multiples phrases du rapport selon le contexte*

La méthode PCA nous permet de sélectionner visuellement les phrases dans un périmètre proche d'un centroïde visuel. Nous pourrions appliquer une méthode des plus proches voisins afin de diviser l'espace en différentes classes représentant ainsi un thème par classe et les centroïdes associés nous permettraient d'extraire automatiquement les phrases les plus proches d'un même thème pour en extraire des sous-ensembles de phrases les plus représentatives du thème.

6. Evaluation

Cette section présente les scores de notre modèle de langage basé sur CamemBERT dont les résultats sont comparés avec le modèle originel. La formation fine comme nous pouvons nous en douter améliore les résultats mesurés avec la métrique ROUGE renforçant l'idée que les phrases exploitées pour extraire le lexique support et les *patterns* sont d'autant plus pertinentes.

ROUGE (Recall-Oriented Understanding for Gisting Evaluation) (Lin 2004) est une métrique utilisée en traitement automatique du langage pour évaluer le résumé automatique des textes. Les métriques comparent un résumé produit automatiquement à une référence ou à un ensemble de références qualifié.

ROUGE-N : ROUGE-N compare des n-grams entre un résumé candidat et la référence. ROUGE-N est calculé comme suit :

$$\left(\frac{\sum_{S \in (ReferemceSummaries)} \sum_{gram_n \in (S)} Count_{match}(gram_n)}{\sum_{S \in (ReferemceSummaries)} \sum_{gram_n \in (S)} Count(gram_n)} \right)$$

Où *n* représente la longueur des n-gram, *gram_n* et *Count_{match} (gram_n)* est le nombre maximal de n-grams occurrents dans un résumé candidat et la référence.

ROUGE-L : fait référence à la séquence commune la plus longue (LCS). L'avantage de l'utilisation de LCS est qu'il ne nécessite pas de correspondances consécutives mais des correspondances en séquence qui reflètent l'ordre des mots au niveau de la phrase. Puisqu'il inclut automatiquement les n-grams communs les plus longs dans la séquence, nous n'avons pas besoin d'une longueur de n-grams prédéfinie :

$$\text{ROUGE-L} = \left(\frac{LCS(X,Y)}{m}\right)$$

Où *LCS (X, Y)* fait référence à la longueur de la sous séquence commune la plus longue entre X et Y. Dans cet article, nous utiliserons ces deux métriques *ROUGE* pour l'évaluation automatique de nos résumés.

CamemBERT (Original)	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L
	0.435484	0.105691	0.032787	0.016529	0.227700
CamemBERT **(entrainé sur les rapports de la banque de France)**	**0.530120**	0.323887	0.269388	0.255144	**0.420348**

TAB. 5 – *Les scores ROUGE des deux modèles: CamemBERT et notre modèle CamemBERT ajusté.*

Comparé au modèle original, notre modèle offre de meilleurs résultats par rapport aux scores ROUGE. Les meilleurs scores de notre modèle présentés sur la table (TAB 5) expriment l'avantage significatif de l'entrainement du modèle sur nos données de la Banque de France, d'où le modèle a appris un nouveau dictionnaire propre au domaine de finance, ce qui nous a permis de générer des résumés plus spécifiques.

Conclusion

Les travaux présentés à l'interface entre sciences du langage et science des données posent la question de savoir comment extraire les *patterns* de mots dans leur environnement proche et comment ces *patterns* améliorent l'idiomaticité de résumés automatiques.

Le modèle de langage de rédaction automatique étudié dans ces travaux est basé sur le mécanisme l'auto-attention et permet la rédaction extractive de résumés. Ce mécanisme est crucial pour identifier à la fois le lexique support et les *patterns* lexico-grammaticaux.

Pour les travaux préliminaires que nous présentons ici pour la première phase de notre méthode, nous avons concentré nos efforts sur l'approche extractive et le modèle CamemBERT. Nous avons opté pour entraîner le modèle CamemBERT à l'aide des rapports de la Banque de France dans le domaine de la finance. Une fois le modèle formé finement à l'aide des rapports, nous avons évalué les scores avant et après cette formation du modèle d'origine CamemBERT. Les résultats sur le score ROUGE de cette formation fine montrent l'amélioration de la qualité des phrases extraites à des fins d'analyses linguistiques.

Pour conclure, le lexique support et les *patterns* lexico-grammaticaux seront les paramètres de la deuxième partie de notre méthode dont l'objet est l'augmentation des données permettant ainsi l'ajustement fin du modèle de rédaction de résumé par approche abstractive pour améliorer l'idiomaticité des résumés générés. La méthode d'augmentation n'est pas présentée ici et fera l'objet de travaux futurs.

Références

Abigail, See. Peter J, Liu. Christopher D, Manning. 2017. "Get To The Point: Summarization with Pointer-Generator Networks". CoRR, abs/1704.04368.

Ashish, Vaswani. Noam, Shazeer. Niki, Parmar. Jakob, Uszkoreit. Llion, Jones. Aidan N, Gomez. Lukasz, Kaiser. Illia, Polosukhin. 2017. "Attention Is All You Need". CoRR abs/1706.03762.

Brügger, Nils. Ian, Milligan. 2019. "The SAGE Handbook of Web History". SAGE Publications Limited.

Daniel, Kondratyuk. Milan, Staka. 2019. "75 Languages, 1 Model: Parsing Universal Dependencies Universally". CoRR abs/1904.02099.

David A, van Dyk. Xiao-Li Meng. 2001. "The Art of Data Augmentation, Journal of Computational and Graphical Statistics". 10:1, 1-50.

Derek, Miller.2019. "Leveraging BERT for Extractive Text Summarization on Lectures". CoRR, abs/1906.04165.

Desmedt, Ludovic. Gautier, Laurent. Llorca, Matthieu (Eds). 2020. "Les discours de conjoncture économique". Paris L'Harmattan.

Devlin, Jacob. Chang, Ming-Wei. Lee, Kenton. Toutanova, Kristina. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arxiv: 1810.04805.

Faber, Pamela (Eds). 2012. "A Cognitive Linguistics View of Terminology and Specialized Language". Berlin: de Gruyter.

Gautier, Laurent (Ed.). 2012. "Les discours de la bourse et de la finance". Berlin: Frank & Timme.

Gledhill, Chris, Kübler, Natalie. 2016. "What can linguistic approaches bring to English forSpecific Purposes?". In *ASp*, 69.

Goldberg, Adele. 2005. "Constructions at work. The nature of generalization in language". Oxford: Oxford University Press.

Granger, Sylvianne. Meunier, fanny (Eds). 2008. "Phraseology. An interdisciplinary perspective". Amsterdam: Benjamins.

Hochreiter, Sepp. Schmidhuber, Jürgen. 1997. "Long Short-term Memory Neural computation". 9, 1735-80.

Kopaczyk, Joanna. Tyrkkö, Jukka (Eds). 2018. "Applications of Pattern-driven Methods in Corpus Linguistics". Amsterdam: Benjamins.

Legallois, Dominique, Tutin, Agnès (Eds). 2013. "Vers une extension du domaine de la phraséologie". Paris: Larousse.

Legallois, Dominique. Charnois, Thierry. Larjavaara, Meri (Eds). 2018. "The grammar of genres and styles. From discrete to non-discrete units". Berlin: de Gruyter.

Levy, O. Goldberg, Y. 2014. "Dependency-Based Word Embeddings". In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 302–308.

Lin, Chin-Yew. 2004. “ROUGE: A Package for Automatic Evaluation of summaries”. in *ACL Workshop: Text Summarization*.

Lydia, Laxmi. Govindasamy, P. Lakshmanaprabu, S.K. Ramya, D. 2018. “Document Clustering Based On Text Mining K-Means Algorithm Using Euclidean Distance Similarity”. In *Journal of Advanced Research in Dynamical and Control Systems*. 10.

Martin, Louis. Muller, Benjamin. Ortiz Suárez, Pedro. Dupont, Yoann. Romary, Laurent. De la Clergerie, Eric. Seddah, Djamé. Sagot, Benoît. 2020. “CamemBERT: a Tasty French Language Model”. 10.18653/v1/2020.acl-main.645. 7203-7219.

Ortiz, Suarez. Pedro, Javier. Laurent, Romary. Benoît, Sagot. 2020. “A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages”. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics. 1703–1714.

Palmieri, Rudi. Rocci, Andrea. Gautier, Laurent (Eds). 2020. “Text and discourse analysis in financial communication”. Thematic Issue of Studies in *Communication Science*. Amsterdam: Elsevier.

Resche, Catherine. 2003. “Décryptage d’un genre particulier : les communiqués de presse de la Banque Centrale américaine”. In *ASp*, 39-40.

Shinyama, Yusuke. 2007. “PDFMiner - Python PDF Parser”.

Straka, Milan. 2018. “UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task”. In *Proceedings of the CoNLL 2018* Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics. 197–207.

Taku, Kudo. John, Richardson. 2018. “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing”. CoRR, abs/1808.06226.

Telmo, Pires. Eva, Schlinger. Dan, Garrette. 2019. “How multilingual is Multilingual BERT?”. CoRR abs/1906.01502.

Yang, Liu. 2019. “Fine-tune BERT for Extractive Summarization”. CoRR, abs/1903.10318.

Ziem, Alexander. 2014. “Frames of Understanding in Text and Discourse”. Theoretical foundations and descriptive applications. Amsterdam: Benjamins.

Résumé

Cet article présente des travaux visant à développer un système de rédaction automatique de résumés de textes économiques et financiers en attachant une attention particulière à l’idiomaticité et à la fluidité de la langue d’arrivée. Pour ce faire, l’étude part d’un corpus de rapports périodiques de la Banque de France relevant des discours de conjoncture. Le travail linguistique permet de montrer qu’une rédaction des résumés ne s’attachant qu’à l’extraction terminologique et collocationnelle stricte ignore tout un pan de vocabulaire, saisi ici comme « lexique support », jouant un rôle important dans l’organisation cognitive du domaine. Sur cette base, le travail présenté sur les modèles de langage en apprentissage profond met en avant la pertinence du mécanisme d’auto-attention pour identifier et extraire des schémas lexico-grammaticaux ainsi le lexique support, et l’impact sur le guidage du modèle de résumé abstractif de CamemBERT à travers l'augmentation des données. Une première expérimentation utilisant le corpus considéré ainsi que la méthode d'extraction sont présentées.