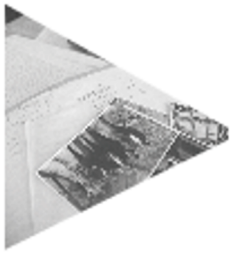
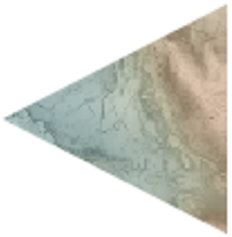


Extraire des *patterns* pour améliorer l'idiomaticité de résumés semi-automatiques en finances : le cas du lexique support

Abdelghani LAIFA TIL EA4182, LIB EA7534
Laurent Gautier TIL EA4182
Christophe CRUZ LIB EA7534



Structure de la présentation

1. Problématique et objectifs
2. Présentation du corpus
3. Limites de l'extraction de *patterns* termino-centrés
4. Rédaction automatique de résumés
5. Expérimentation
6. Evaluation

Conclusion

1. Problématique et objectifs

- Croisement entre
  - approches des textes spécialisés par delà l'unité-mot isolée
  - développements actuels de résumé automatique de texte par apprentissage profond
- Deux questions de recherche en synergie :
  - comment extraire les *patterns* de mots dans leur environnement proche ?
  - comment ces *patterns* améliorent-ils l'idiomaticité des résumés automatiques ?

2. Corpus

- Texte sériel : *Bulletin mensuel de la Banque de France*
- Domaines : Macroéconomie, finances, politiques publiques

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

- Tendance globale au dépassement des unités terminologiques isolées : extension du champ phraséologique
- 3 grands paradigmes dans la recherche récente :
  - Théorie des scénarios / *frame semantics* : représentation organisée des connaissances liées à un concept, résultat de l'expérience du locuteur => retombées sur combinatoires et figement
  - Modèles des *patterns*/schémas avec suspension de la dichotomie lexique-syntaxe

"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". (Gledhill/Kübler 2016, 75)

– Grammaire(s) de construction : degré ultime d'abstraction du modèle des *frames*/scénarios => permettent de modéliser avec un haut degré de granularité l'interface syn-taxe-sémantique.

3.1 De l'extraction terminologique aux collocations

- Extraction des termes mono- et polylexicaux
- Extraction systématique des combinatoires récurrentes
- Mise au jour d'un *frame* de comparaison, comme dans :

(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).

- Modélisation possible par extraction systématique des N et V présents dans le répertoire :

(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

- Extraction systématique révèle l'articulation des combinatoires figées avec du lexique-grammaire non saisi par une approche strictement termino
- « Lexique support » directement lié avec le scénario de comparaison précédent
- Dimension 1 : découpage chronologique des tendances/évolutions de la comparaison => aspect comme dans :

(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)

- Dimension 2 : mise en discours du caractère prospectif et incertain des prévisions => modalisation / évidentialité, comme dans :

(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.

- Idiomaticité »de spécialité » va dépendre *aussi* de la prise en compte de ces deux dimensions dans la rédaction des résumés

=> Introduction des « points d'attention » dans le système d'IA

4. Méthode informatique proposée pour la rédaction automatique des résumés

1ère partie

- Extraction des patterns avec CamemBERT

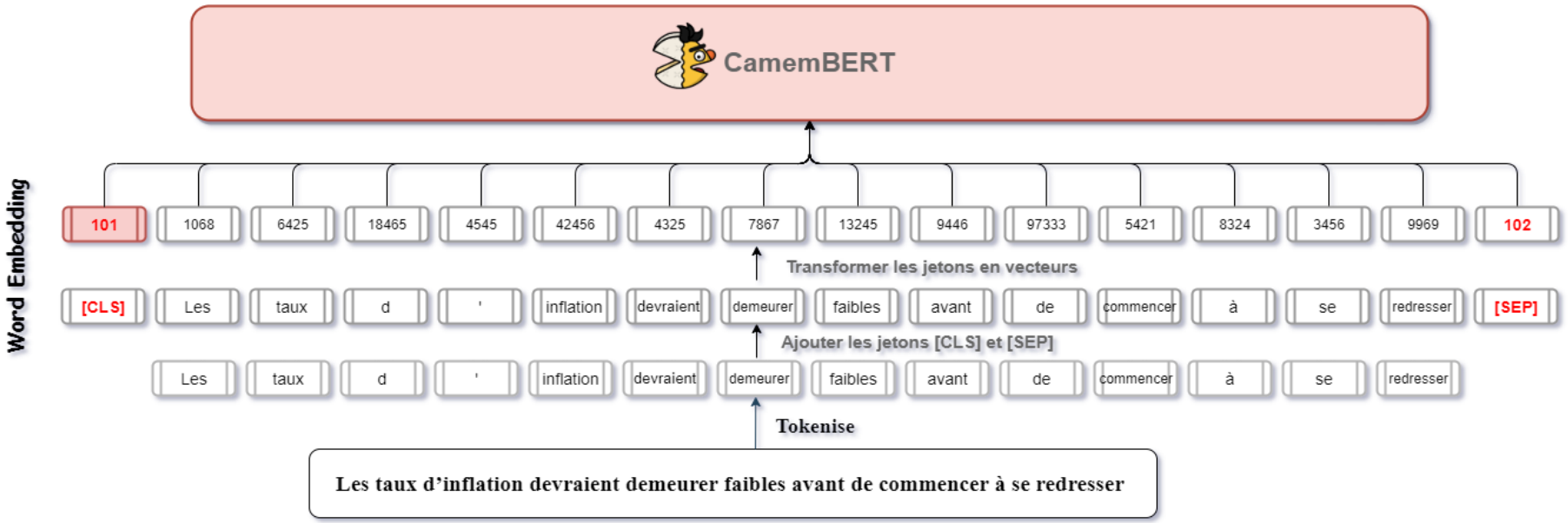
- Evaluation du modèle CamemBERT ajusté

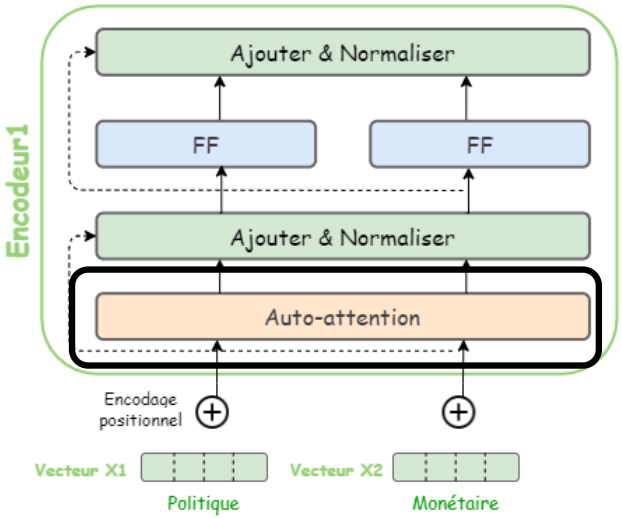
2ème partie

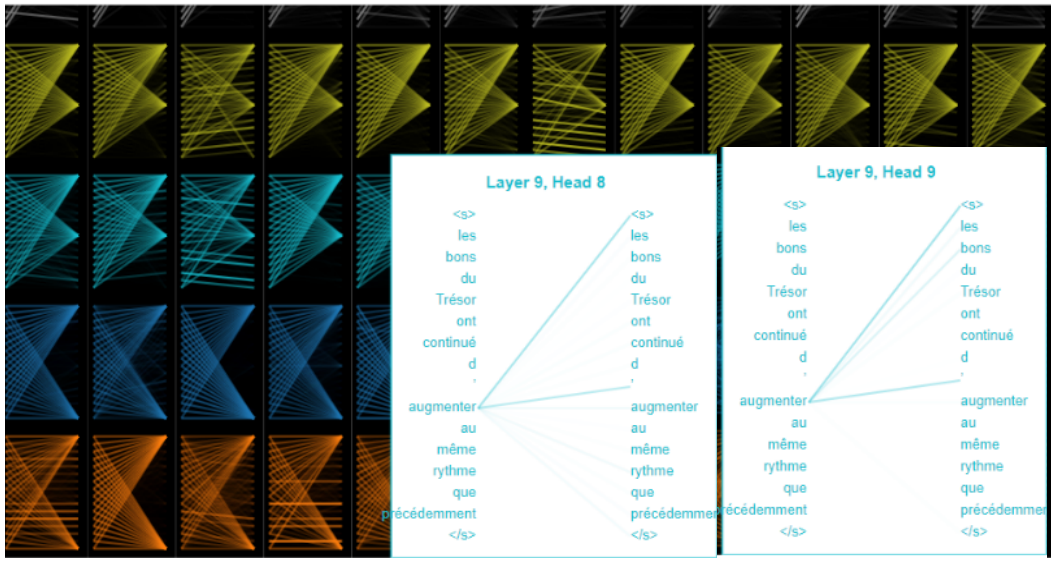
- Augmentation des données

5. Expérimentation

5.1 Extraction des patterns avec CamemBERT

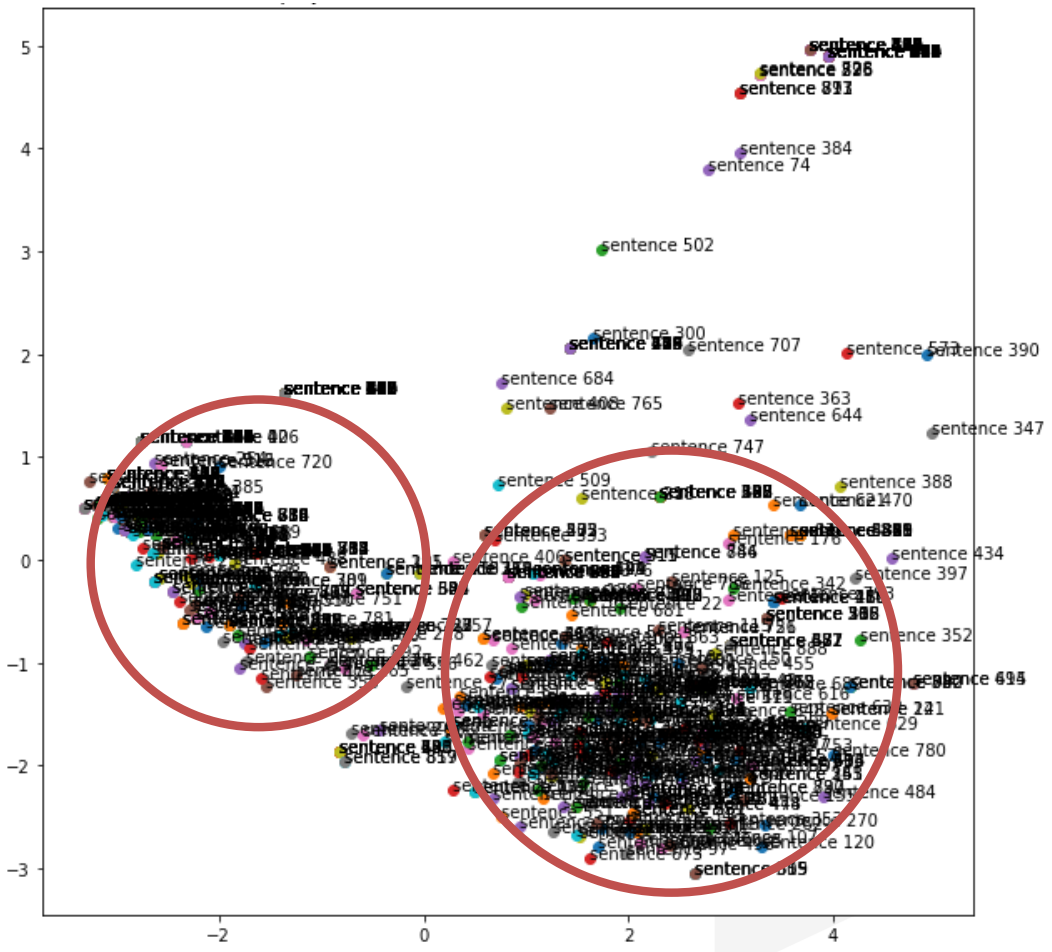






- Encodeur – décodeur
- Auto-attention
- Exemple

- PCA et visualisation des schémas lexico-grammaticaux



5.2 Augmentation des données

- Exemple

En revanche, les bons du Trésor ont continué d'augmenter au même rythme que <mask>

```
[{'score': 0.3298775851726532,
  'sequence': '<s> En revanche, les bons du Trésor ont continué
d'augmenter au même rythme que précédemment :)</s>',
  'token': 7488,
  'token_str': '_précédemment'},
 {'score': 0.24416683614253998,
  'sequence': '<s> En revanche, les bons du Trésor ont continué
d'augmenter au même rythme que prévu :)</s>',
  'token': 1936,
  'token_str': '_prévu'},
 {'score': 0.01743445359170437,
  'sequence': '<s> En revanche, les bons du Trésor ont continué
d'augmenter au même rythme que 2016. :)</s>',
  'token': 4890,
  'token_str': '_2016.'},
 {'score': 0.015762045979499817,
  'sequence': '<s> En revanche, les bons du Trésor ont continué
d'augmenter au même rythme que dernièrement :)</s>',
  'token': 17594,
  'token_str': '_dernièrement'}]
```

6. Evaluation du modèle CamemBERT ajusté

ROUGE (Recall-Oriented Understanding for Gisting Evaluation) est une métrique utilisée en traitement automatique du langage pour évaluer le résumé automatique des textes. Cette métrique compare un résumé produit automatiquement à une référence ou à un ensemble de références qualifié (Les rapports mensuels de la Banque de France pour notre cas).

	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L
CamemBERT (Original)	0.435484	0.105691	0.032787	0.016529	0.227700
CamemBERT (Ajusté)	**0.530120**	**0.323887**	**0.269388**	**0.255144**	**0.420348**

Conclusion

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.

Merci pour votre attention !

Abdelghani LAIFA

(Abdelghani_laifa@etu.u-bourgogne.fr)

Laurent GAUTIER

(Laurent.Gautier@u-bourgogne.fr)

Christophe CRUZ

(christophe.cruz@u-bourgogne.fr)