Impact of textual data augmentation on linguistic pattern extraction to improve the idiomaticity of extractive summaries

Abdelghani Laifa[1][0000−0002−2327−4675], Laurent Gautier[2,3][0000−0002−6210−410X], and Christophe Cruz[3][0000−0002−1698−3027]

[1] Laboratoire d'Informatique de Bourgogne, Maison des Sciences de l'Homme, Université de Bourgogne, France
Abdelghani_Laifa@etu.u-bourgogne.fr
[2] Maison des Sciences de l'Homme, Université de Bourgogne, France
Laurent.Gautier@u-bourgogne.fr
[3] Laboratoire d'Informatique de Bourgogne LIB, EA 7534, Univ. Bourgogne Franche-Comté, Dijon, France
Christophe.cruz@u-bourgogne.fr

Abstract. The present work aims to develop a text summarisation system for financial texts with a focus on the fluidity of the target language. Linguistic analysis shows that the process of writing summaries should take into account not only terminological and collocational extraction, but also a range of linguistic material referred to here as the "support lexicon", that plays an important role in the cognitive organisation of the field. On this basis, this paper highlights the relevance of pre-training the CamemBERT model on a French financial dataset to extend its domain-specific vocabulary and fine-tuning it on extractive summarisation. We then evaluate the impact of textual data augmentation, improving the performance of our extractive text summarisation model by up to 6%-11%.

Keywords: Text summarisation · Terminology · Linguistic Patterns · Natural Language Processing · Corpus Linguistics · Deep learning.

1 Context and objectives

The work presented here is part of a larger project conducted at the University of Burgundy, at the crossroads between the language sciences and data science. The main research question is how to extract patterns from their environment in order to improve the readability of automatic summaries within the field of finance. The project also questions, as in the following case study around what we propose to call the "support lexicon", the limits of strictly terminological and/or collocational inputs when dealing with specialised discourses. We explore a set of economic reports from the *Banque de France*, which are "serial texts" that combine various statistics with more explicative sequences. The extraction based on the usual approaches in terminology quickly showed that an essential

part of the lexicon was neglected. We therefore considered this to be a limit of such approaches and explored the possibilities of focusing on the support lexicon and the information it conveys, especially for deep learning. Thus, we present an extractive summarisation model for French based on self-attention. In this paper, we highlight our method of extracting lexico-grammatical patterns using the attention mechanism, which is a part of a neural architecture capable of highlighting relevant features in the input data. Because of the limited number of monthly reports available, we used the data augmentation principle to generate artificial data from the original dataset. The augmented corpus, composed of 226 initial reports and 226 artificial reports, allowed us to improve the performance of the fine-tuned model.

The corpus belongs to a discourse type that can be qualified as "conjuncture discourse" [1, 6, 20] and is produced by national central banks. It generally presents a very high degree of recurrence: both at the level of the contents themselves and of the form, with a very rigid macrostructure. The corpus is in French, and each report has its own summary. From a quantitative perspective, the corpus contains 323 monthly reports published between 1994 and 2020, and is made up of 6,554,396 words, 4,070,955 lemmas and 317,076 sentences. Each report has an average length of 1500 words, and each summary has an average length of 200 words. The remainder of this paper is structured as follows. Section 2 provides elements about the limits of terminology-based models and introduces automatic text summarisation. Section 3 presents the approach adopted here to produce summaries, starting with an explanation of the pre-training and the fine-tuning processes, as well as the extraction of syntactic and semantic patterns. It ends with the data augmentation method. Section 4 describes the results and evaluations.

2 Background

Pattern extraction based solely on the terminological approach has been revealed to to be limited. The extraction of lexico-grammatical patterns based on attention mechanisms therefore seemed to be a solution to improve the idiomaticity of the summaries produced.

2.1 From Terminology to Patterns and constructions

Recent work on specialised discourses has very clearly emphasised the limits of approaches based on words as isolated units. The continuous extension of the field of "phraseology" [5, 8, 13, 14] aims to capture recurrent segments that are more significant than words. Three holistic approaches are currently implemented in this field, revealing the role played by support lexicon, as a lexicon which does not fall within traditional terminology but is "applied", through patterns, to other terminology-collocational structures [12].

Frame semantics already implemented in terminology, is a cognitive linguistics paradigm aimed at an organised representation of knowledge linked to a concept that results from the experience of the speaker. Frames lead to particular encodings in language through combinatorics and preferential syntactic productions. **Lexico-grammatical patterns** challenge the traditional view of lexical modules (dictionaries), on which grammatical rules operate. Not only does the type of constrained text studied here implement a restricted repertoire of the French grammatical language system, but it does so only in synergy with the lexicon concerned with building blocks of meaning [7]: "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". They are the markers of a double idiomaticity in the corpus: idiomaticity of the language, and idiomaticity of the field allowing the experts to convey field-related information without ambiguity. **Construction grammars** represent the highest degree of abstraction and generalisation of frames and allow the syntax-semantic interface to be modeled with a high degree of granularity. As they are usage-based, they also have their starting point in recurring patterns.

2.2 Text summarisation

Text summarisation consists of creating a short version of a text document by extracting the essential information. There are two main approaches: extractive and abstractive. The extractive approach extracts the document's most salient sentences and combines them into a summary. In contrast, the abstractive approach aims to generate a summary as humans do, by extracting and paraphrasing the original text. There are relatively few works on text summarisation in French compared to other languages, and they focus mainly on extractive approaches by scoring sentences and selecting the highly scored ones for the summary. The lack of a French benchmark corpus makes evaluation for French summarisation more difficult. CamemBERT [17] is the first BERT-type [2] language model. It was trained on a French dataset containing the texts of numerous web pages.

3 Methodology

Fundamentally, the BERT-type model is a stack of Transformer encoder layers [23] that consist of multiple self-attention "heads". For every input token in a sequence, each head computes key, value, and query vectors, used to create a weighted representation. The outputs of all heads in the same layer are combined and run through a fully connected layer. Each layer is wrapped with a skip connection and followed by layer normalisation. The input representations are computed as follows: Each word in the input is first tokenised with SentencePiece [11], and then three embedding layers (token, position, and segment) are combined to obtain a fixed-length vector. Special token [CLS] is used for

classification predictions because it stores the combined weights provided by the heads of each sentence, and [SEP] separates the input segments. We will break down our approach into two essential stages, pre-training and fine-tuning. We first continued the training of the original CamemBERT model on the financial dataset using the masked language modeling task. Then, we fine-tuned our pre-trained CamemBERT model on an extractive summarisation task.

3.1 Pre-training the original CamemBERT on a financial dataset

In this method, we inserted all the corpus texts except the summaries into the model. The model learned the domain-specific financial vocabulary as well as the reports' lexico-grammatical patterns by predicting the randomly masked input tokens . The idea behind this method is to expand the knowledge of the original CamemBERT and to create a new French CamemBERT model that can understand and evaluate textual data in the field of finance.

3.2 Fine-tuning of our model on extractive summarisation

The fine-tuning method consists of adding a sentence classifier on top of the final encoder layer to predict which sentences should be included in the summary. For this, we converted our summarisation dataset, which consists of report-summary peers, to an extractive summarisation dataset by assigning a label 1 to each sentence of the report included in the summary and a label 0 to the other sentences in the report. Then, we divided the corpus into a training (226 reports), a validation (48 reports) and a test dataset (49 reports), in order to train and test the sentence classifier. At the end of this process, we froze our model so that only the parameters of the sentence classifier were learned from scratch. The model was pre-trained for 2 epochs over 44,800 steps on 1 GPU (Tesla K80) with a learning rate of $5e^{-5}$, setting a batch size for training to 10 and 100 for fine-tuning. Model checkpoints are saved and evaluated on the validation set every 500 steps. The sentence classifier was fine-tuned on the same GPU for 3 epochs with a learning rate of $2e^{-5}$.

3.3 Heads with linguistic knowledge

CamemBERT uses 12 layers of attention, and also incorporates 12 attention "heads" in every layer. Since model weights are not shared between layers, the CamemBERT model has up to 12 x 12 = 144 different heads of attention mechanisms. The question is: what is the capability of the encoder's attention mechanism in capturing linguistic knowledge as lexico-grammatical patterns? We extracted the attention mapped from our pre-trained model with BertViz [24] to explore the attention patterns of various layers/heads and to determine the linguistic patterns that mapped to our understanding of lexico-grammatical parsing.

Syntactic dependency After extracting the attention maps, we evaluated the prediction direction for each attention head: the direction of the head's word referring to the dependent and the opposite one. Three syntactic dependencies

were explored to emphasise the interdependence of vocabulary (lexis) and syntax (grammar). **Results:** As Figure 1 shows, in (A), heads 4 and 12 of layer 6 allowed the verbs "redresser" and "inscrire" to correctly address their attention to their subject. In (B), heads 4 and 11 of layer 4 allowed the noun's determiner to attend their noun, and in (C) the head 2 of layer 1 allowed the prepositions to attend their complements. Our model's attention heads performed remarkably well, illustrating how syntactic dependencies can emerge from the pre-training phase. We also noticed that other attention heads were able to capture these syntactic patterns but with less significant attention. We also experimented with the attention heads of the original CamemBERT on the financial data. The results were not significant because the attention weights were dispersed over the whole sequence without taking into account the grammar of the specific domain, in particular for experiment (A).

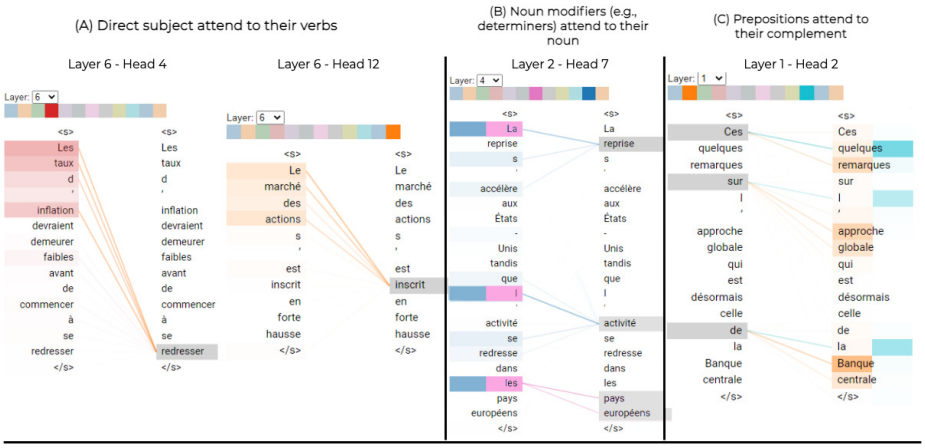

Fig. 1. CamemBERT attention heads that correspond to some syntactic linguistic phenomena. In this visualisation, the darkness of a line indicates the strength of the attention weight. All attention to/from words is coloured according to the head's colour, which emphasises the attention pattern. Selected words are highlighted with a grey box.

Semantic dependency Coreference resolution is a challenging task that requires context understanding, reasoning, and domain-specific perception. It consists of finding all the linguistic expressions in a given text that refer to the same real-world entity. This is why most of the linguistic phenomena studies [10,16,25] have been devoted to BERT's syntactic phenomena rather than its semantic phenomena [3, 22]. We used attention heads for the challenging semantic task of coreference resolution. We evaluated the attention heads of our model on coreference resolution using the financial dataset. **Results:** As shown in Figure 2, we found that some heads (9 - 12) of layer 6 achieved excellent coreference resolution performance, where "leurs" and "celle" correctly referred to their entities "les enterprises d'assurance" and "l'approche globale" respectively.

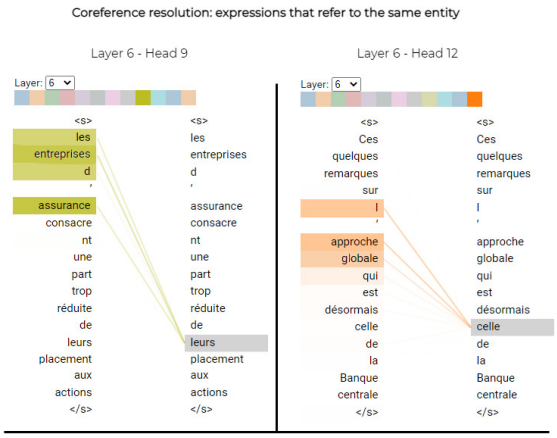

Fig. 2. Visualisation of coreference resolution attentions.

Compared to syntactic experimentation, we noticed that few attention heads are specialised in capturing semantic phenomena (only the heads 9 and 12 of the layer 6 in our case).

3.4 Data augmentation in CamemBERT

BERT-type models have been pre-trained on the task of "masked word prediction", so that the input sentences will have some masked words, and the model tries to predict them by suggesting adequate words according to the context of the input sentence. In our case, we used our pre-trained CamemBERT model for this task to obtain the artificial data. In doing so, we used a Stanford part-of-speech tagger trained in French to read the dataset reports and assign parts of speech to each word. Then, we selected all the modifiers (adjectives and adverbs) in the reports to mask them and predict with our pre-trained CamemBERT model, which adjectives and adverbs could appropriately substitute the masked modifiers (except the original ones). We thereby generated a synthetic training dataset containing new modifiers which retained the same meaning and patterns. We trained the model again on the new training dataset that included both the original and the synthetic data. This technique was chosen because our model considers the context of the sentence that includes the masked word before predicting the appropriate word.

4 Results and evaluation

We evaluated the quality of the summary using ROUGE [15]. Unigram and bi-gram overlap (R-1 and R-2) are reported as a means of evaluating informativeness and the longest common subsequence (R-L) as a means of evaluating fluency. We compared the performance of the different CamemBERT models, and

we further included the Lead-3 [21] multilingual baseline that extracts the first 3 sentences from any given article and the PyTextRank [19], considered as multilingual baseline, which is an implementation that includes the TextRank [18], PositionRank [4] and Biased TextRank [9] algorithms, used for extractive summarisation by getting the top-ranked phrases from a text document.

Table 1. Fine-tuned CamemBERT model scores before and after data augmentation

Models	R_1	R_2	R_L
Lead-3 baseline	0.3342	0.1220	0.1735
PyTextRank baseline	0.3979	0.1269	**0.2616**
CamemBERT (Original)	0.4354	0.1056	0.2277
Fine-tuned CamemBERT (Without data augmentation)	**0.5301**	0.3238	**0.4203**
Fine-tuned CamemBERT (With data augmentation)	**0.6136**	0.4201	**0.5102**

Table 1 shows the significant advantage of training the model on our custom financial dataset, where the model learned a new domain-specific vocabulary, allowing us to generate more specific summaries. In addition, data augmentation on the Bank of France's textual monthly reports data contributed to an improvement of 6% to 11% in terms of the model's performance and the automatically-produced summaries.

5 Conclusion

The extraction of lexico-grammatical patterns from the attention mechanisms allowed our model trained on the bank reports to produce more accurate and domain-specific summaries. We have highlighted the impact of textual data augmentation on the model's performance and on the extracted summaries. The lack of data does not allow the model to cover all of the semantically relevant aspects of the domain-specific data. More original financial data could improve the model's performance.

References

1. Desmedt, L., Gautier, L., Llorca, M.: Les discours de la conjoncture économique. L'Harmattan, Paris (2021)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
3. Ettinger, A.: What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics **8**, 34–48 (2020)
4. Florescu, C., Caragea, C.: PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada (2017)
5. Gautier, L.: Figement et discours spécialisés. Frank und Timme, Berlin (1998)

6. Gautier, L.: Les discours de la bourse et de la finance. Frank und Timme, Berlin (2012)
7. Gledhill, C., Kübler, N.: What can linguistic approaches bring to english for specific purposes? ASp. la revue du GERAS (69), 65–95 (2016)
8. Granger, S., Meunier, F.: Phraseology: An interdisciplinary perspective. John Benjamins Publishing (2008)
9. Kazemi, A., Pérez-Rosas, V., Mihalcea, R.: Biased TextRank: Unsupervised graph-based content extraction. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 1642–1652. Barcelona, Spain (2020)
10. Kim, T., Choi, J., Edmiston, D., goo Lee, S.: Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction (2020)
11. Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing (2018)
12. Laifa, A., Gautier, L., Cruz, C.: Extraire des patterns pour améliorer l'idiomaticité de résumés semiautomatiques en finances : le cas du lexique support. In: ToTh 2020 - Terminologie et Ontologie. Université Savoie Mont-Blanc, Presses Universitaires Savoie Mont-Blanc, Chambéry, France (2020)
13. Legallois, D., Charnois, T., Larjavaara, M.: The Grammar of Genres and Styles: From Discrete to Non-discrete Units. Walter de Gruyter GmbH & Co KG (2018)
14. Legallois, D., Tutin, A.: Présentation: Vers une extension du domaine de la phraséologie. Langages (1), 3–25 (2013)
15. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
16. Marecek, D., Rosa, R.: From balustrades to pierre vinken: Looking for syntax in transformer self-attentions (2019)
17. Martin, L., Müller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: Camembert: a tasty french language model (2019)
18. Mihalcea, R., Tarau, P.: TextRank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. pp. 404–411. Association for Computational Linguistics, Barcelona, Spain (2004)
19. Nathan, P.: PyTextRank, a Python implementation of TextRank for phrase extraction and summarization of text documents (2016)
20. Rocci, A., Palmieri, R., Gautier, L.: Introduction to thematic section on text and discourse analysis in financial communication. Studies in Communication Sciences **15**(1), 2–4 (2015)
21. See, A., Liu, P.J., Manning, C.D.: Get to the point: Summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1073–1083. Association for Computational Linguistics, Vancouver, Canada (2017)
22. Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R.T., Kim, N., Durme, B.V., Bowman, S.R., Das, D., Pavlick, E.: What do you learn from context? probing for sentence structure in contextualized word representations (2019)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30**, 5998–6008 (2017)
24. Vig, J.: A multiscale visualization of attention in the transformer model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 37–42 (2019)
25. Vilares, D., Strzyz, M., Søgaard, A., Gómez-Rodríguez, C.: Parsing as pretraining. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)