CoFiF: A Corpus of Financial Reports in French Language

Sina Ahmadi - Tobias Dauder
Insight Centre for Data Analytics
Overview

1. Introduction

2. Corpus and its importance

3. Objective

4. Our corpus: CoFiF (creation, analysis and experiments)

5. Conclusion
Corpus and its importance

“A corpus is a collection of natural language (text, and/or transcriptions of speech or signs) constructed with a specific purpose.”*

- Large corpora play a progressively important role in driving NLP research
- More and more used to feed notorious neural network models
- Many corpora are available for different fields in various languages

What about financial domain in French?

* O’Keeffe, Anne, and Michael McCarthy, eds. *The Routledge handbook of corpus linguistics*. Routledge, 2010.
Objective

- Many resources available for English, such as: the Wall Street Journal (WSJ) Corpus [Paul and Baker, 1992]
  - the 10-k Corpus [Ko-ganet al., 2009]
  - the 8-k Corpus [Leeet al., 2014]
- No corpus to date deals with French texts in the field of economics and finance

Our goal → to create a French corpus in the domain of business and economics
Our corpus: CoFiF

- The first corpus comprising company reports in the French language
- It contains over 188 million tokens in 2655 reports
- The corpus spans over 20 years, ranging from 1995 to 2018
Corpus creation

1. Selection of enterprises according to:
   - **Cotation Assistée en Continu (CAC) 40**: France’s main stock index containing 40 of the 100 largest companies by market capitalization of the stock exchange
   - **CAC Next 20**: the 20 largest ones which are listed following the ones in the CAC40

2. Collecting 4 types of documents by consulting company’s website:
   - reference documents (*documents de références*)
   - annual results (*résultats annuels*)
   - semestrial results (*résultats semestriels*)
   - trimestrial results (*résultats trimestriels*)
Corpus analysis: reports distribution per year

![Graph showing the distribution of reports per year for CAC40, CAC Next 20, and Total.]
Experiments: language models

Evaluation of the corpus using two character-level language models:

• a forward recurrent neural network (RNN)
• a backward one

| Sentence                                                                 | Perplexity |
|-------------------------------------------------------------------------|------------|
| Perspectives d’avenir et principaux risques.                            | 1.7892     |
| Perspectives avenir et principaux risques.                              | 2.9605     |
| Le chiffre d’affaires de l’activité autocars augmente principalement suite à une amélioration du prix moyen, et ce malgré un recul des volumes de 3 %. | 1.9745     |
| Le chiffre d’affaires l’activité autocars augmente principalement suite de une amélioration du prix moyen, et ce malgré un recul dès volumes 3 %. | 2.9471     |
| Cette stratégie permettrait ainsi d’accroître les péages ferroviaires perçus par Groupe Eurotunnel pour l’utilisation de son infrastructure. | 2.4411     |
| Cette stratégie permettrait ainsi d’accroître les péages ferroviaires perçus par Groupe Eurotunnel que l’utilisation son infrastructure. | 2.8991     |
Experiments: word embeddings

We trained a Word2Vec model* on the cleaned textual data of CoFiF and evaluated the quality of the retrieved word embeddings.

| bénéfice  | perte       | croissance | impôt  | économie  |
|----------|-------------|------------|--------|-----------|
| profit   | 0.715       | dépréciation 0.666 | progression 0.857 | impôts 0.783 | agriculture 0.672 |
| versement | 0.544       | moins-value 0.599 | décroissance 0.782 | fonctionnelle 0.606 | installation 0.609 |
| solde    | 0.534       | variation 0.571 | hausse 0.731 | imputation 0.592 | énergie 0.603 |
| résultat | 0.512       | insuffisance 0.515 | amélioration 0.719 | amortissement 0.535 | problématique 0.593 |
| dividende | 0.512       | diminution 0.505 | dynamique 0.716 | déduction 0.533 | innovation 0.581 |

* Mikolov, Tomas, et al. "word2vec." URL https://code.google.com/p/word2vec (2013).
Getting CoFiF

CoFiF is openly available for non-commercial use under the Attribution-NonCommercial-ShareAlike 4.0 International license. More information at https://cofif.github.io.
Conclusion

• CoFiF is the first French large corpus in financial domain

• Enables progress in applying NLP approaches on the textual data related to the financial sector in francophone countries, particularly France, Canada, Belgium and Switzerland.

• It provides a comprehensive and factual overview of a company’s shape and their language can also be consulted in linguistic terms.