Reported speech detection in newspapers

Maxence Morin, Lisa Chabrier

Reported speech detection in newspapers

Morin, Maxence maxence.morin@orange.com [1, 2] and Chabrier, Lisa lisa.chabrier@inria.fr[3]

[11]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, FRANCE
[2]Orange Innovation, 14000 Caen
[3]Univ Lyon, Inria, INSA Lyon, CNRS, UCBL, LIRIS (UMR5205)

October 10, 2023

Abstract

This report describes the work produced at the occasion of the workshop SEEDS (*Semaines Études Entreprises en Data Sciences*) organized by GdR@MADICS. This workshop was hosted at the University of Troyes (UTT).

The industrial project to which we contributed during the workshop was proposed by the association SPOT, and consisted of the detection of reported speech in text extracted from the written press, and the association of such reported speech to a subject. We focused on the French newspaper "Le Parisien", according to the requirement of the association SPOT.

The workshop lasted five days, with an introduction session from the industrial product owners, three and a half days of work, and a presentation of the work from all groups on the last day. This report presents the methodology, the contribution, and the results produced during the workshop.

Contents

1 Introduction

SPOT is an association founded by Samy Monnier in 2020. SPOT allows for each social debate to gather and organize in a collaborative manner all the information already published in the media and on social networks. The website is available at https://www.spotdebats.org/. Thanks to Natural Language Processing (NLP), the website highlights tweets that correspond to a social debate written by people present in a short list. The short list is composed of elected officials and political figures from the RNE (*Répertoire National des Élus*).

The tool is composed of three parts. First part is the cleaning of the data. Second part is the process of getting features of articles and debates. Last part is about reported speech detection.

2 Data

SpotTheNews was used on a database of 8,000 news articles extracted from French newspaper "Le Parisien". Articles are described by a title, a content, an abstract, a URL, a date, the author and a Boolean for the accessibility of the article (i.e. reserved to subscribers of the newspaper). The content of the article is raw HTML, with a lot of extra information, as links, references to other article, links to twitter and so on.

2.1 Cleaning

The data cleaning pipeline is separated in three parts. The first is common for both parts of the project; the second is for the keyword extraction only and the third is used by the reported speech detection.

Common pipeline Whether it is for the keyword extraction, the keywords matching or the reported speech detection, a common pipeline has been implemented. Because the article content contains HTML tags, we used Beautiful Soup 4 to parse, remove useless tags and transform left tags in text. The content of the article is merged with the title and the abstract.

Keyword extraction To get features for an article, we had to remove stop words (e.g. "un", "de", "la") and put the whole to lower case. We also removed punctuation. Left words are essentially key-words.

Reported speech extraction Because the spaCy and Stanza separation of the article into sentence don't worked in our case, we use a custom function to split the article. This function relies on strong clues of the end of a sentence as the dot. Citation has to be fully extracted, even if the reported speech contains several sentences. To solve this problem, the function obfuscates the text between quotation before splitting into sentences. This allows us to keep reported speech whole. Reported speech are transformed at the end of the process in plain text.

3 Contribution

3.1 Keyword extraction

All keywords from the concatenation of the title and the content of the article are analysed using KeyBERT (Grootendorst 2020) python package. KeyBERT gives us the top fifteen keywords of the content, and the top five keywords for debate titles. Those keywords come with their relevance score ($score \in [0; 1]$).

3.2 News article matching

The main goal of the project is to gather who said what about which debate. To do this, we had to compare articles keywords and to debates keywords, and according to a matching threshold, select articles corresponding to a specific debate. The distance function used to compare keywords is cosines.

3.3 Named Entity Recognition (NER)

Naive approach At first, name recognition was only treated by SpaCy. The text is cut into lemmas, and each lemma has a type attached. When the type corresponded to "PERS", we would try to compare the name obtained to the people white list. For the comparison, several approaches were implemented. One method tolerant to small mistakes using a hamming distance was discarded due to the variation the journalist would use to cite a person (i.e. Emmanuel Macron, Mr Macron, Macron...) could not be safely detected. The current approach still has some limits but is compatible with a small database. The name found in the text is split into all its sub-elements (Names, surnames, "Mr" or "Mme"...). The simple algorithm will try to match two elements in the database. This can lead to errors, but it will usually find the right person. The rule is strict when there is an ambiguity:

no one is returned. This behavior could be changed with a small adjustment of the code.

Advanced approach In the advanced approach, we tried to detect co-reference to a person, which is very often used in press articles to avoid the repetition of the full name. It is often the case that the paraphrase used as a coreference is longer than the actual name. To achieve this, we used the independent package *cross-lingual-coreference* (definition by Lee et al. 2017 and python implementation by Berenstein 2022), which can be used in the general SpaCy framework. It is possible to immediately replace every coreference in the text, but the common alias used will be the first mention of the entity, which will not necessarily be the name of the person (See 1). Our context requires us to be able to identify some with a name, therefore we tweaked the behavior of the package into a homemade version that will identify a known person in the list of possibles alias, and only if someone was identified replace every mention in the text by the full name. This resolved text is then sent to the reported speech detection part of the code.

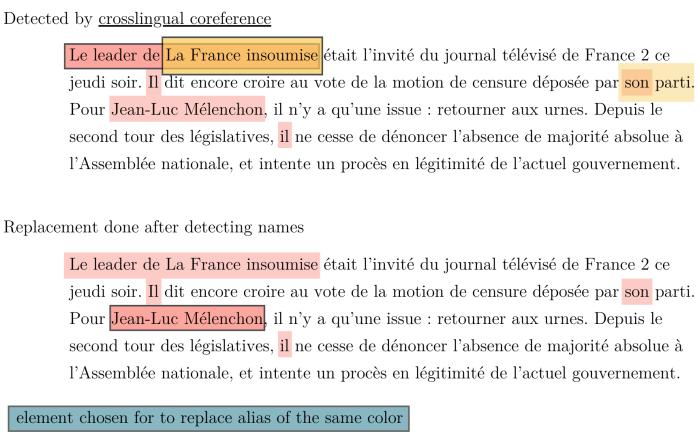

Figure 1:

3.4 Reported speech detection

Reported speech in french is composed of a common name, a quotation verb and the reported words. According to De La Clergerie et al. 2009, this triptych has a limited number of possible combinations, especially in news feeds. During the week, we tried two approaches to detect reported speech.

Naive approach In the naive approach, SpotTheNews detects the three types of clues: People, quotation verb, and punctuation. People are found by NER module of SpaCy. Punctuation is found thanks to a RegEx. SpaCy analyses the sentence and gives to words several attributes including the lemma of the verb. SpotTheNews relies on this lemma to find the infinitive form. Verb of the sentence and their infinitive form are compared to the whitelist of quotation verb extracted from the website Wiktionary. At this point, there are three lists containing each a type of clue and their position in the text. Theses three list are matched to find the closest name n_i, verb v_i and quotation q_i. The objective function is defined as:

$$\min \sum_{\substack{a=0 \\ b=0 \\ c=0}}^{\substack{N \\ V \\ Q}} d(n_a, v_b) + d(n_a, q_c) + d(v_b, q_c)$$

N is the number of detected names, V the number of detected verbs and Q the number of detected quotations. The function $d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is the Euclidean distance. The result is a tuple list of tuple of the form (n_i, v_i, q_i).

Advanced approach In the advanced approach, SpaCy resolves all the references or linguistic anaphoras in the text (see section 3.3). Stanza analyses resolved sentences and gives several tags and a parent word to each word. SpotTheNews then builds a graph based on the dependency tree extracted by Stanza (see figure 2). We consider that the name, the verb and the reported speech are all included in the same sentence.

If the sentence contains quotation marks, SpotTheNews gets the closest quotation verb of the quotation mark in the dependency graph. Otherwise, it searches for a quotation verb without taking into account a quotation mark. Then, it searches for the common name the closest to the verb. To extract the full name of the person, it searches for part of first name or last name using BFS[1] approach. SpotTheNews considers that the found triplet is a reported speech. Theses are filtered to keep only quotations made by people in the whitelist.

[1] Breadth-first search is an algorithm for searching a tree data structure for a node that satisfies a given property.

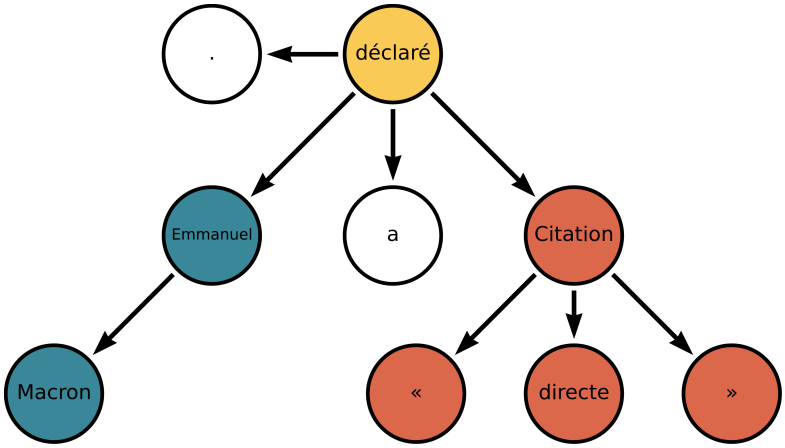

Figure 2: Syntactic tree given by Stanza of the french sentence 'Emmanuel Macron a déclaré « Citation directe ».' (*Emmanuel Macron declared « Direct quotation ».*)

4 Final deliverable

At the end of this week, we proposed the architecture shown in figure 3. Debates are cleaned and analysed to obtain the matching score with each news article. Articles are also cleaned and, for each article reported speech are extracted and identified using methods presented in the section 3.4 and the section 3.3 respectively. Article that match sufficiently a debate (i.e. match score is above a threshold) are selected with their reported speech and presented to the Spot users.

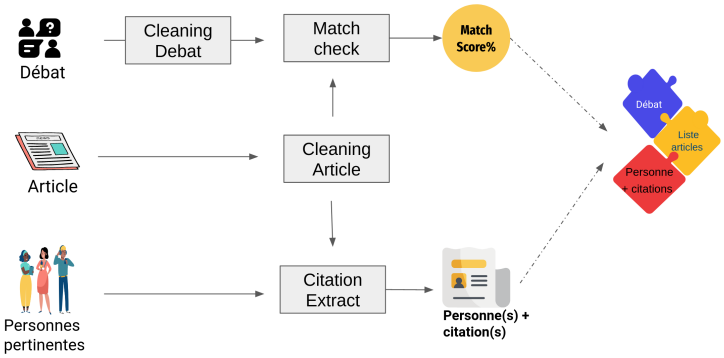

Figure 3: Final workflow proposed

The source code of the reported speech detection is available on Github.

5 Future work

In order to improve our work, we could try to detect reported speech without quotation verbs, removing the need to maintain an updated list of such quotation verbs that may have several meanings and favor mistakes.

We also faced some issues with the co-referencing of small particles, which allows the algorithm to replace "il a dit" with "Emmanuel Macron a dit" and help the association of the reported speech to the author of the citation. The solutions proposed in this reports have room for improved accuracy in these replacements.

6 Acknowledgements

The authors would like to thank the organizers of SEEDS@MADiCS, Myriam BERTRAND and Frédéric BERTRAND, as well as Samy MONNIER, founder of the association SPOT and supervisor of the Project. We also acknowledge and thank Aya SAHBI and Muhammad ARSLAN for their active participation in the project.

References

Berenstein, David (Sept. 2022). *Crosslingual Coreference - a multi-lingual approach to AllenNLP CoReference Resolution along with a wrapper for spaCy.* (Cit. on p. 6).

De La Clergerie, Éric, Sagot, Benoît, Stern, Rosa, Denis, Pascal, Recourcé, Gaëlle, and Mignot, Victor (2009). "Extracting and Visualizing Quotations from News Wires". In: *4th Language and Technology Conference* (cit. on p. 6).

Grootendorst, Maarten (2020). *KeyBERT: Minimal keyword extraction with BERT.* Version v0.3.0. DOI: 10.5281/zenodo.4461265 (cit. on p. 5).

Lee, Kenton, He, Luheng, Lewis, Mike, and Zettlemoyer, Luke (2017). "End-to-end Neural Coreference Resolution". en. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Copenhagen, Denmark: Association for Computational Linguistics, pp. 188–197. DOI: 10.18653/v1/D17-1018 (cit. on p. 6).