FRACAS: a FRench Annotated Corpus of Attribution relations in newS

Richard, Ange (1,2) Alonzo Canul, Laura C. (1) Portet, François (1)
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
(2) Univ. Grenoble Alpes, CNRS, Sciences Po Grenoble, Pacte

Quote extraction and source attribution

- Applications for both Social science and Natural Language Processing
- A complex task
- Little available corpora

Our contributions:

1. FRACAS, manually annotated corpus of 10,965 attribution relations in newswire texts
2. Experiments with recent relation extraction architectures and comparison with rule-based baseline

Task definition

- Quotation extraction:
  - Sequence classification: finding quote, speakers and cue spans in a text
  - Quote types: direct, indirect, mixed

- Source attribution:
  - Relation extraction
  - Linking a quote span to a speaker span

Related Works

- Quote extraction systems:
  - Many rule-based
  - Different approaches:
    - Sentence classification (Brunner, 2013)
    - Sequence classification (some rule-based, some neural network-based)
      - Pouliquen *et al.*, 2007, Salway *et al.*, 2017, Soumah *et al.*, 2023
      - Scheibe *et al.*, 2016, Papay and Pado, 2019; Pareti, 2015
- Relation extraction systems have greatly improved with recent architectures and the use of Large Language Models
- Few available labelled corpora:
  - Some for English (PARC3 (Pareti, 2012), Quotebank (Vaucher *et al.*, 2021)
  - None for French

Our corpus

Our Corpus: FRACAS

French newswires from Reuters agency

Part of a multilingual corpus:

RCV2 Corpus, distributed by the National Institute of Standards and Technology

August 1996 to August 1997

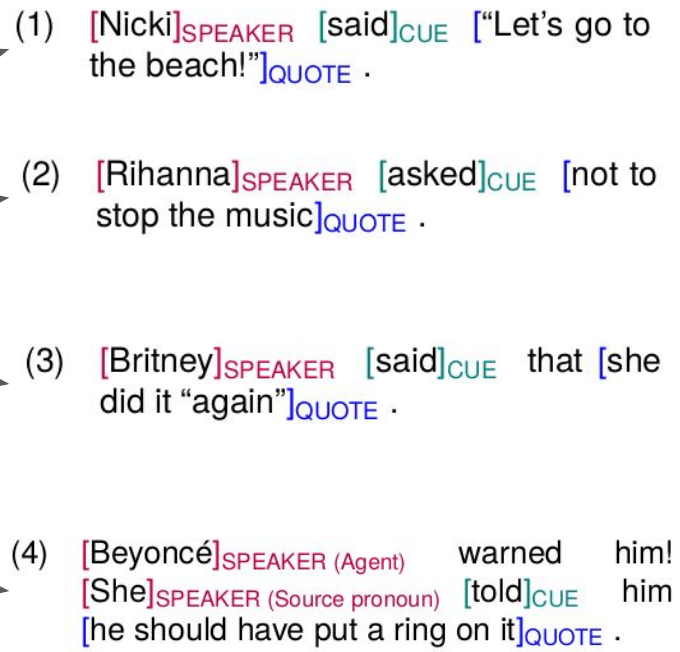
1,676 texts picked at random from the 85,710 texts of the French part of the corpus

Partition	#docs	#tokens	Mean #tok. per doc
train	1,114	436,150	391.51
dev	281	131,202	466.91
test	281	137,083	487.83
TOTAL	1,676	704,435	448.75

Table 2: Number of documents and tokens in the FRACAS

Annotations

- **Quotation types:**
  - Direct quotation
  - Indirect quotation
  - Mixed quotation
- Speaker types:
  - Agent
  - Group of People
  - Organization
  - Source Pronoun (+ referent entity)
- Cue



Inter-annotator agreement

- 9 annotators for the first batch (1,436 texts)
- 3 annotators for additional texts to improve speaker gender balance
- BRAT software (Stenetorp et al., 2012)
- γ score to measure IAA (Mathet et al., 2015) as it accounts for both unitizing and categorization
- Final annotations determined by an assessment of the best annotator regarding a gold standard.

Entity label	**γ agreement**
All entities	0.7699
Direct Quotation (Q)	0.8857
Indirect Quotation (Q)	0.6415
Mixed Quotation (Q)	0.7468
Cue	0.8291
Agent (S)	0.8337
Organization (S)	0.7858
Group of people (S)	0.7828
Source pronoun (S)	0.898

Table 5: γ agreement between annotators of first annotation campaign (S = Speaker entity types, Q = Quotation entity types)

Experiments

Modeling approaches

We model quotation extraction as a relation extraction task and solve it using two approaches:

- Radar de Parité (**baseline**, Soumah et al., 2023):
  - A rule-based system designed for quotation extraction in French (and the only one freely available up to the publication of this article)
  - The system extracts speaker, cues and quotes (no subdivision for speaker and quote types)
- REBEL (**SOTA,** Cabot and Navigli, 2021):
  - A system for relation extraction using natural language generation and based on BART (Lewis et al., 2019)
  - Relation types are not constrained
  - A multilingual version of this model pre-trained in French news was recently released (Cabot et al. 2023)

Data processing and evaluation

Our data preparation pipeline follows two main steps:

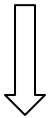
1. **Paragraph building** (i.e. sectioning documents from FRACAS into smaller paragraphs) - Input for Radar de Parité and REBEL
2. **Triplet linearization** - Output needed to train REBEL

We test our models on two common evaluation modes for relation extraction:

- ***Boundaries* evaluation** - A predicted relation is considered correct if entity boundaries and relation label match the gold standard
- ***Strict* evaluation** - A predicted relation is considered correct if entity boundaries, entity labels and relation label match the gold standard

From annotations to paragraphs for natural language generation

fr_XX LONDRES, 30 mars, Reuter - Les Spice Girls, groupe féminin en tête des ventes de disques, ont ouvert dimanche après-midi l'antenne de Channel 5, première télévision hertzienne créée en Grande-Bretagne depuis 15 ans.\nDétenue par les groupes britanniques de communications Pearson Plc et United News & Media, le consortium européen CLT-Ufa et la société américaine d'investissement Warburg Pincus, Channel 5 vise en priorité les moins de cinquante ans.\n"Beaucoup de gens âgés de moins de cinquante ans ne regardent pas les informations télévisées", a déclaré Kirsty Young, présentatrice du journal de cette nouvelle chaîne hertzienne.\nLa journaliste a promis que les programmes d'information de Channel 5 seraient "plus excitants pour les rendre plus accessibles". Ce "ne sera pas simplement un homme en costume assis derrière un bureau", a-t-elle ajouté.



tp_XX <triplet> "Beaucoup de gens âgés de moins de cinquante ans ne regardent pas les informations télévisées" <dirquot> Kirsty Young, présentatrice du journal de cette nouvelle chaîne hertzienne quoted in <dirquot> a déclaré <cue> indicates
tp_XX <triplet> que les programmes d'information de Channel 5 seraient "plus excitants pour les rendre plus accessibles" <mixquot> La journaliste <per> quoted in <mixquot> a promis <cue> indicates
tp_XX <triplet> Ce "ne sera pas simplement un homme en costume assis derrière un bureau" <mixquot> elle <pron> quoted in <mixquot> ajouté <cue> indicates
tp_XX <triplet> elle <pron> La journaliste <per> refers

Example of the 2 stages of data preparation followed to make FRACAS ready for evaluation on the two proposed systems. The picture on the top shows an example from FRACAS (a paragraph built from a newswire text) with three annotated relations highlighted. The picture on the bottom shows our linearization method applied to the sample on the top, to make it suitable for training the multilingual generation model.

Experiment results

System	QI	I	R	Prec	Rec	F1
$mREBEL_{bo}$	66.21	70.83	43.24	70.59	65.15	67.76
$mREBEL_{st}$	62.07	69.09	43.24	67.60	62.40	64.89

Table 6: Relation extraction results on the dev set of FRACAS. Relation scores per relation type (QI: quoted in, I: indicates, R: refers) are given.

Main takeaways

- Language generation models provide an improvement over other RE modeling strategies when it comes to identifying entity boundaries and speakers
- However, they still struggle to identify correct tags for quotations (discriminating between indirect and mixed quotations is a hard task even for human annotators) and referents

Read our paper for more details and links to *FRACAS* and our code !