# Natural Language Processing and Semantics for Cybersecurity: challenges and approaches to deal with social network data

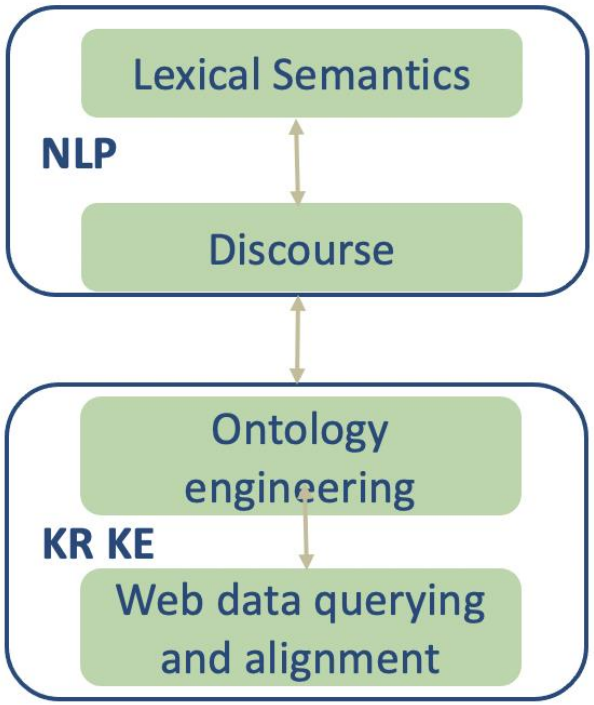
Nathalie Aussenac-Gilles

CNRS, IRIT, MELODI

aussenac@irit.fr

# **ME**thodes et ingénierie des **L**angues, des **O**ntologies et du **DI**scours

14 permanents, 2 sites (UPS et UT2J)
16 thèses en cours ; 3 post-doctorants ; 3 Ingénieurs de recherche
http://www.irit.fr/-Equipe-MELODI-
Farah Benamara farah.benamara@irit.fr
Philippe Muller philippe.muller@irit.fr




## Compétences

- Machine learning / Deep learning
- Extraction d'information
- Ontologies et techniques du web sémantique
- Analyse de corpus
- Annotation sémantique de corpus

# Outline of the talk

- Cybersecurity: challenges for NLP and document processing
- Security issues with Data and text from social networks
- Some NLP techniques to deal with these issues
- Ethic issues and regulations: example with authorship identification

# Cybersecurity: challenges for NLP and document processing

# Cybersecurity and NLP: multiple relationships

## When the end user is the target …

- Cyber harassment, threatening
- Opinion manipulation, fake news
- Trapping people before making physical contact
- Getting his money, his data

## … because the end user can become a « help » in the attack

- Pishing emails
- Asking personal data, leading to fake web sites
- Hooking, ransomware
- Reaching other people or groups

## Natural language is the media, a step towards other kinds of attacks

- Physical attacks, prostitution networks, paedophilia, human trafficking ...
- Cyber attacks either on the user machine, or from his machine towards larger institutions (hospitals, universities, companies, …)

# Cybersecurity and NLP: multiple relationships

## When the end user is the issue

- Author of cyber crimes using natural language (harassment in social networks, fake news on web sites, pishing emails, …)
- Dialogs and posts about crimes in real life: terrorism, prostitution, …
- Exchange of prohibited content: pedopornographic pictures and video, governmental data, stolen technical documents …

## Digital tracks can be a clue

- Text + images + video + icons: various (complementory) media
- Connected persons in the network
- Metadata: log time, log location, alias, message time, …
- Natural language

## Need for Natural Language and Document Processing

# Language in social networks

## Variety of networks

- Forums  structured in closed groups for registered participants: Discord, Slack
- Open forums: Twitter
- FaceBook or Tiktok: mix of personal posts, comments about these posts (that can become discussions), advertisements and commercial posts

## Features of these networks

- You need to register: identity checking or nor
- You can post text: open (for every one) or closed (selected group)
- What I can read: contributions from my groups or algorithmic push
- Nature of contributions: short text, conversations, posts
- Relations between contributions: comment, answer, repost, share …
- …

## Need to adapt NLP techniques / tasks

# Security issues with Data and text from social networks

# Challenges for NLP: example of attack prevention

## The need: attack prevention

- prevent a drama, human or material damages
- caused either naturally or by humans
- by watching information flows, chats or social network posts and conversations

## Potential social impacts

- Defense/security (e.g., intention to comit a crime)
- Health/sanitary crisis (e.g., suicide, virus propagation)
- Civil security (e.g., plan help, preventive evacuation, intervention)

## The issue : How to detect an intent to act?

- Find informative text (vs noise or spam)
- Detect weak signals (comments, critics, calls …)
- Provide a relevant interpretation of these signals
- Identify the source and the author
- Provide a degree of trust of this analysis
- (Suggest potential relevant reactions)

# Challenges for NLP: NLP tasks

- Weak signal identification
- Identification of sexist discourse, dangerous actions or threats
- Authorship identification
- Fake news or spam recognition

- Personal network identification
- Extracting named entities: places, persons, dates
- Extracting events and their relations

# Challenges for NLP

- Understanding coded words or phrases
- Analysing text from various social networks: Discord vs Twitter vs FB
- Multimodality analysis, combining text, images, video
- Multilinguality; few endowed languages

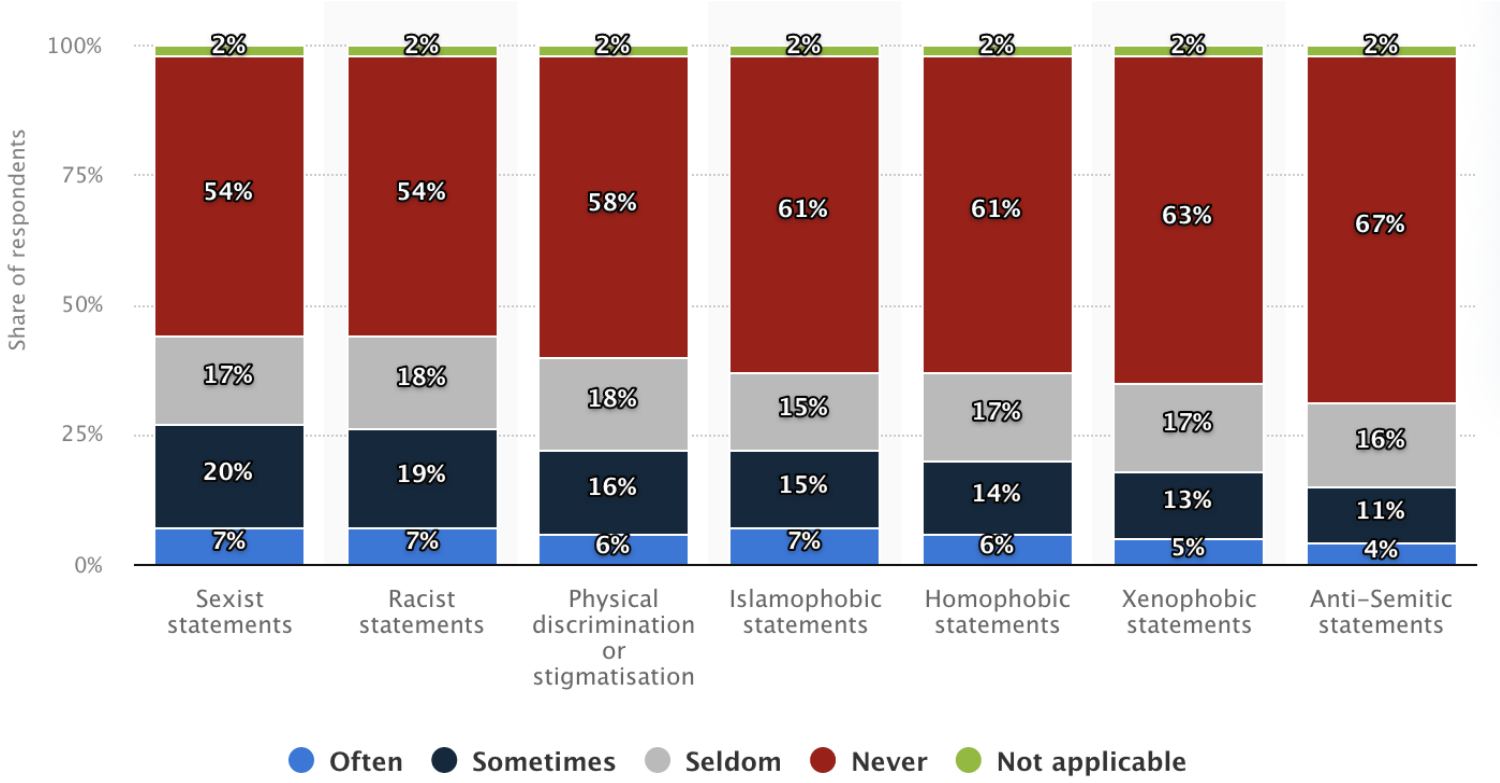
- Volume (either too small or too large)
- Very short time to perform analyses
- Joint analysis of text, document structure and meta-data

… And ensure compliance with ethics and current regulations regarding personal data and AI.

# How important are these phenomenon?

## Hate speech

Source : https://www.statista.com/statistics/421111/france-frequency-hateful-speech-encounter-online/



# How important are these phenomenon?

## Reporting and Denonciation of hate and violent speech acts

Source *Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, Marlène Coulomb-Gully: An Annotated Corpus for Sexism Detection in French Tweets. LREC 2020: 1397-1403*

| Sexist content | | | | Non-sexist | Total |
|---|---|---|---|---|---|
| 4,487 | | | | 7,787 | 12,274 |
| direct | descriptive | reporting | other | | |
| 45 | 780 | 3,222 | 440 | | |

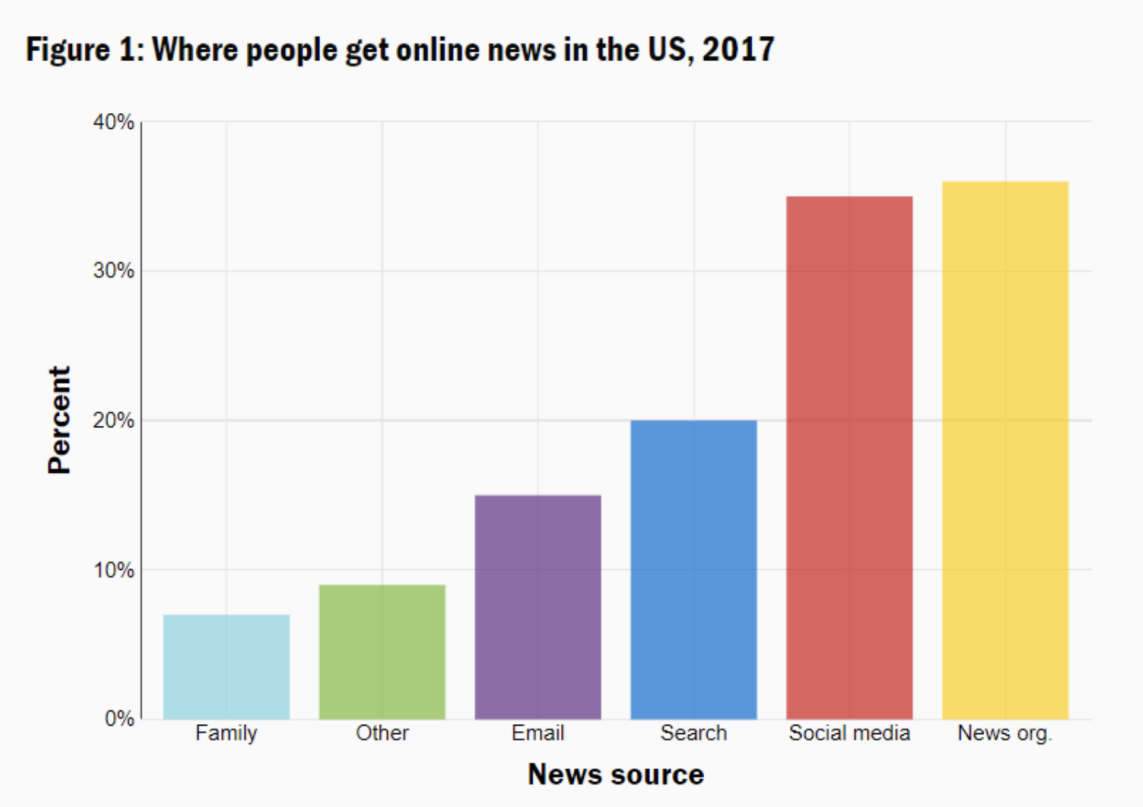
Figure 1.2 – Tweet distribution in our French dataset.

# How important are these phenomenon?

## Fake news

Source : https://europeansting.com/2019/03/06/fake-news-what-it-is-and-how-to-spot-it/

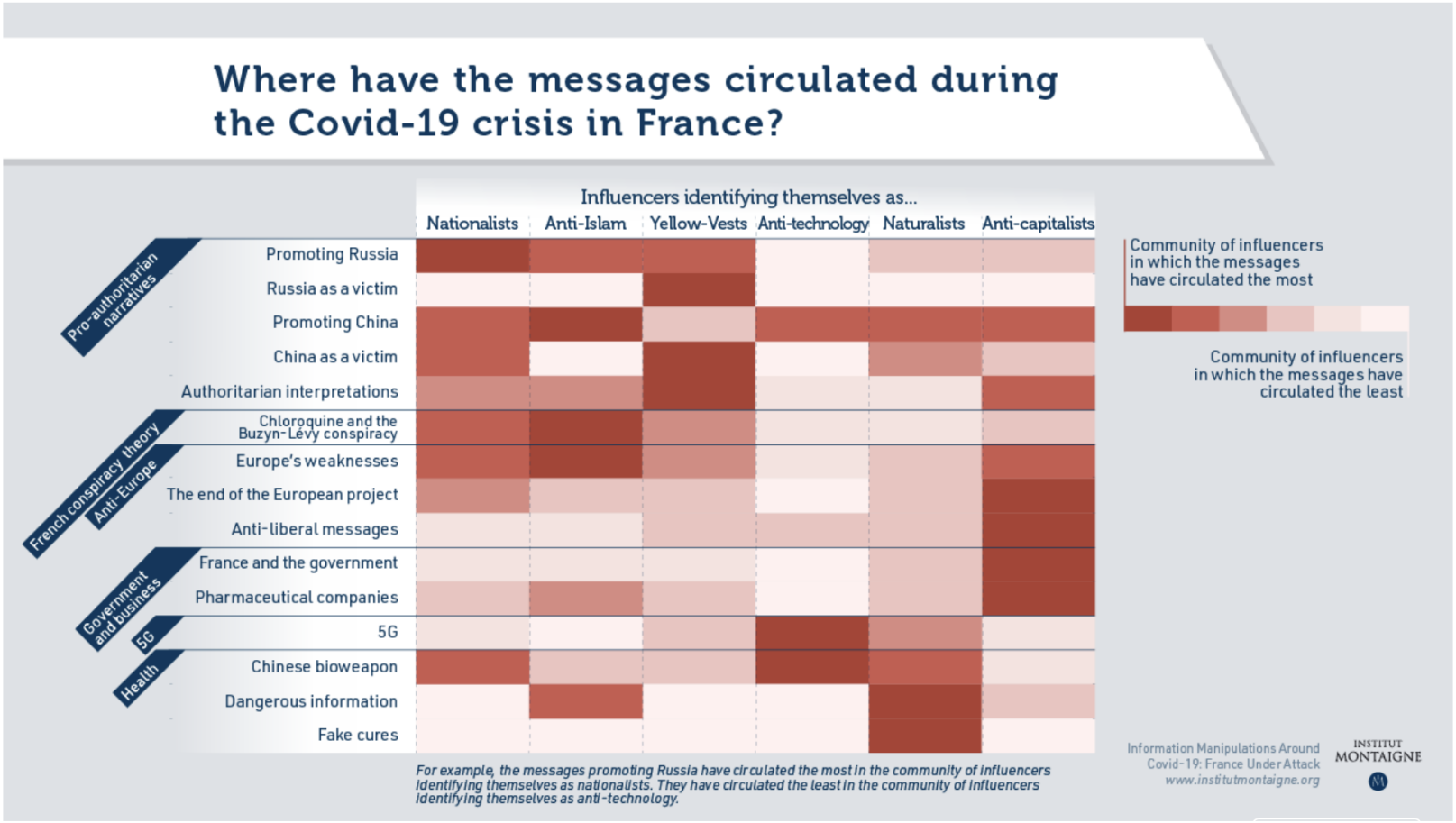
Erosion of public trust
in traditional news sources,
creating a vacuum filled
by misinformation



# How important are these phenomenon?

## Fake news: network and community importance

Source : institutmontaigne.org



# NLP for not so natural language

NLP can be applied to text with « artificial » languages

- Logs
- Programming langages (code)
- Controled languages
- Specifications

… To look for regularities, identify dangerous behaviours, check conformance with writing rules, etc

… To ensure higher security

# NLP in short

# NLP and relations with other disciplines

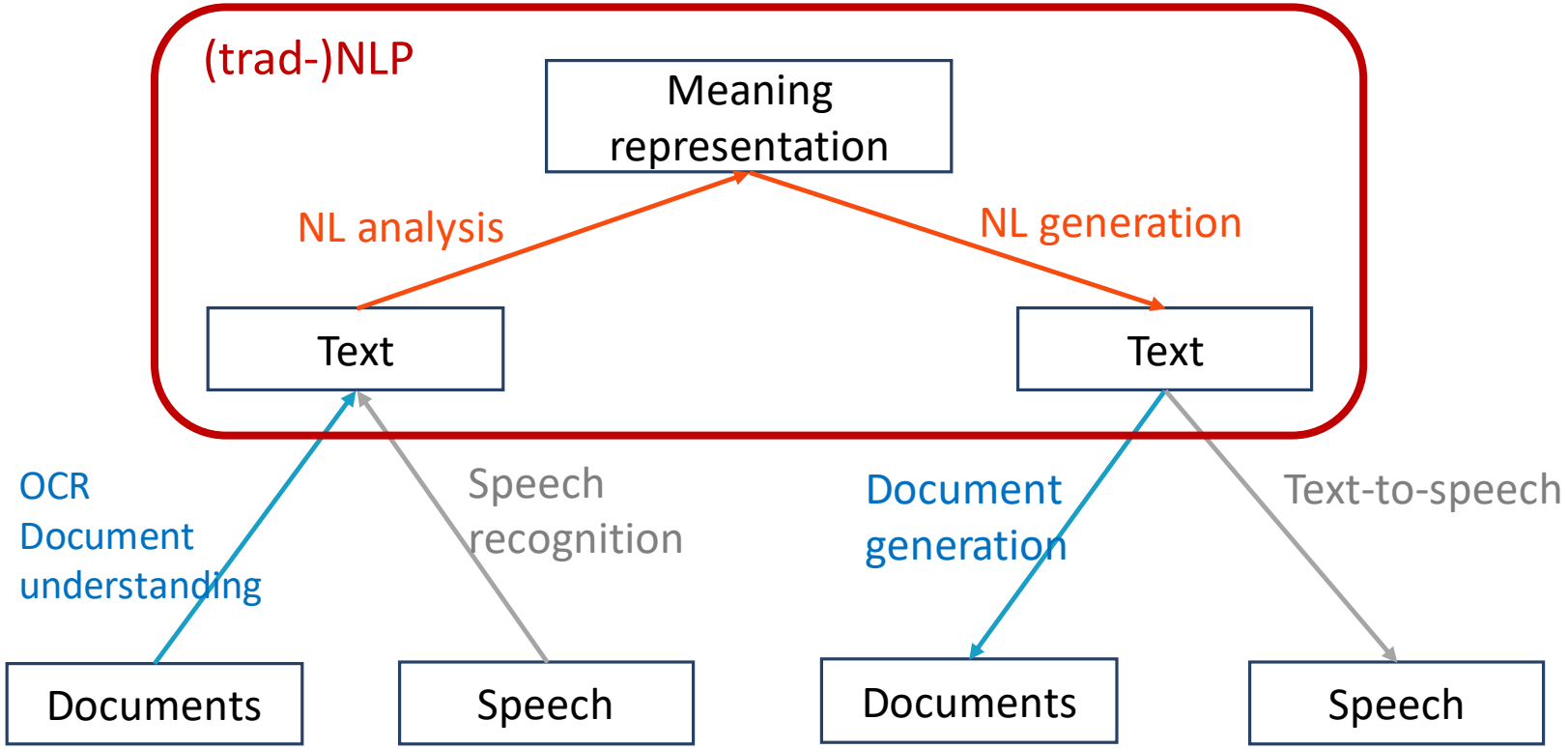
## NLP and friends

- Formal semantics
- Computational Linguistics
- Natural Language Processing
- Text mining (based on statistics, counting word, token or character frequency)
- Information extraction

## A cross-disciplinary domain

- Linguistics (and psycho-linguistics)
- Philosophy
- Computer science (IA)
- Sociology

# NLP, what do we mean?



# Text, an important data and knowledge source

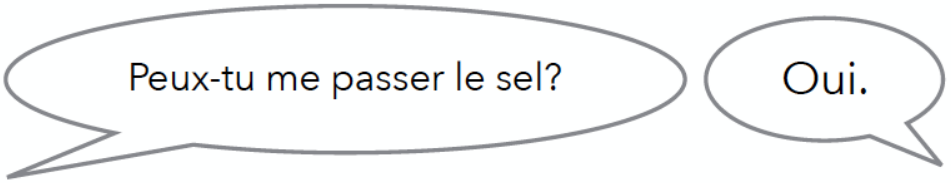
Text is 80% of the big data

| NLP Applications using | or producing text |
| --- | --- |
| • Syntactic analysis, POS tagging | Automatic traduction |
| • Spelling cheking, grammatical checking | Rewriting |
| • Keyword extraction, topic labelling | Text summarization |
| • Dialogue modelling | Conversation management |
| • Fact checking | Chatbot, robotique |
| • Search engines | Question answering |
| • … | |

# Language processing: a layered approach

Discours

Pragmatique

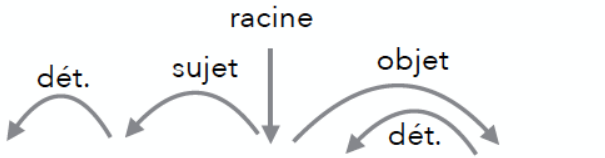



Sémantique

**Le cuisinier**$_{agent}$ **goûte**$_{manger\_goûter}$ **la soupe**$_{patient}$ **avec une cuillère**$_{instrument}$

Phrases

Syntaxe




Syntagms

Mots

Lexicology / terminology

Morphologie

inévitablement = in+évit+able+ment

tweeterait = TWEETER-cond-prs-3sg

Sons

Phonologie

/lɪŋgwɪstɪks/

# Why is NLP difficult?

Sequential data by essence more than « bag of words »

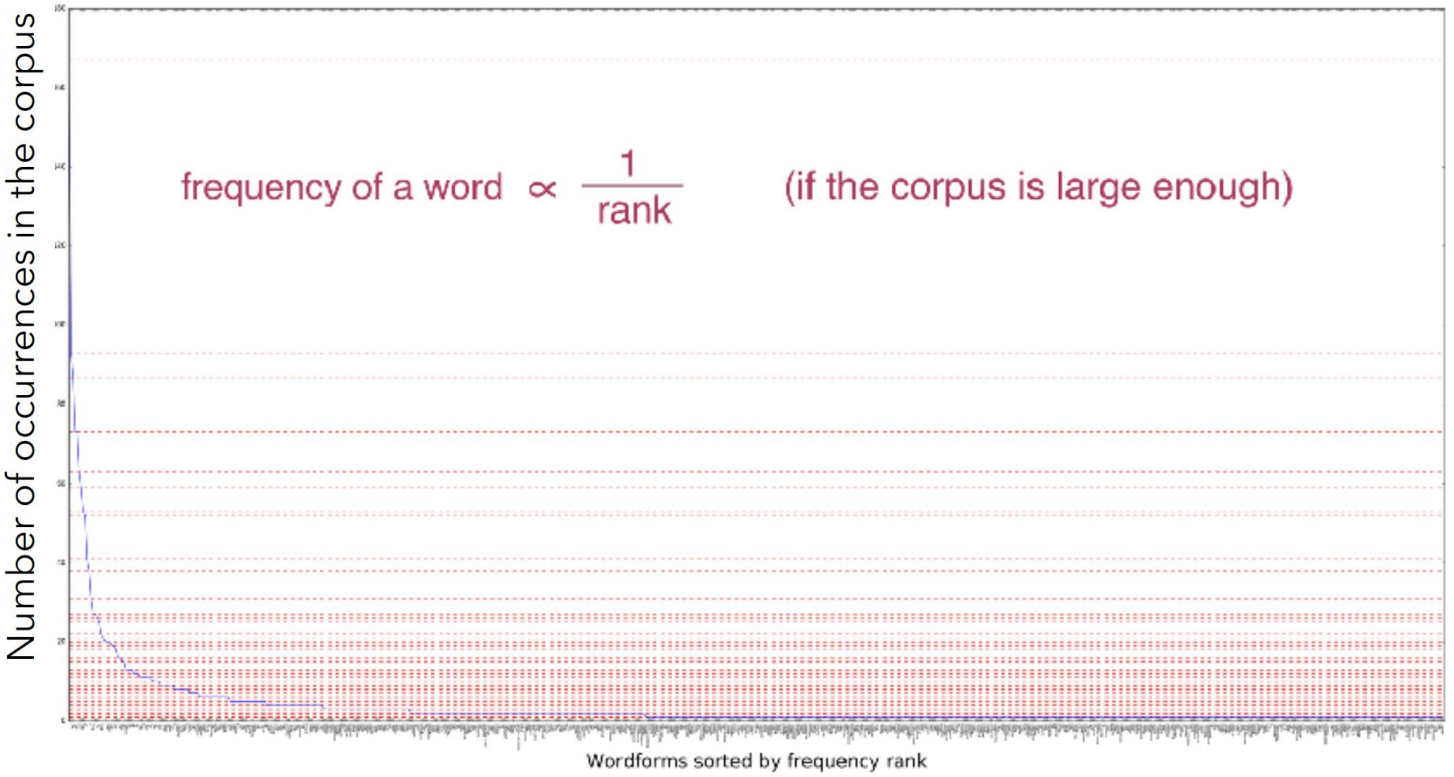
Les plats et l'ambiance de ce restaurant ne sont pas assez bons

- Meaning results from a composition process
  - From words to syntagms
  - From syntagms to sentence
- Syntax and grammatical categories (parts of speech) have a role
  - Adverbes modulate adjective or verb values (assez bon)
  - Nouns and adjectives can characterize other nouns
  - Opinions are based on these features

# Why is NLP difficult?

## Entries are sparse and intedependant

- Large vocabulary, frequence in power law (Zipf law)
- Words are connected with semantic links



# Why is NLP difficult?

Context is required to solve lexical ambiguities and polysemies

For named entities

Détection
+
Linking

Paris (disambiguation)

From Wikipedia, the free encyclopedia

**Paris** is the largest city and capital of France.

**Paris** may also refer to:

Look up *Paris* in W
the free dictionary.

Contents [hide]

1 People
2 Mythological or fictional
3 Places
3.1 Canada
3.2 United States
3.3 Other
4 Film and television
5 Music
5.1 Artists
5.2 Musicals
5.3 Albums
5.4 Songs
6 Science and technology
7 Ships
8 Other uses
9 See also

People [ edit ]

- Paris (surname), a list of people and fictional characters
- Paris (given name), a list of people and fictional characters
- Lucius Domitius Paris (died 67 AD), actor in Rome under the emperor Nero
- Count of Paris, a title held by senior members of the House of Orléans, and a list of the titleholders

Mythological or fictional characters [ edit ]

- Paris (mythology), a prince of Troy in Greek mythology
- Count Paris, in Shakespeare's play *Romeo and Juliet*
- The Great Paris, stage name of a fictional character on the television series *Mission: Impossible*

Places [ edit ]

**Canada** [ edit ]

- Paris, Ontario, a community
- Paris, Yukon, a former community

**United States** [ edit ]

For verbs and nouns

Les poules du couvent couvent.

Il porte la porte.

Meaning is not « additive »
Cf pomme de terre

# Why is NLP Difficult?

Context is required to solve ambiguities at the sentence level

- Domain dependence (how long is long?)

The life duration of this smartphone is not very long.

We had to way a long while between the meals.

It has been raining for a long time.

- Neutral Expressions

This movie surprised me

- Authors'view versus the reader's interpretation

A small restaurant (it is too small or tiny and cute ?)

# Why is NLP difficult?

Context is important at the discourse level

The meaning of the entire discourse is not the « sum » of its parts

[the characters are unsympathetic .]1

[The scenario is totally absurd .]2

[The decor seems to be made of cardboard .]3

[But all these elements make the charm of this TV series.]4

# NLP, an evolving domain – Historical trends

## Early works

- Logics
- Rule based reasonning
- Grammars and patterns
- Human interpretation by language experts: linguists, semanticians, AI language specialists, philosophers
- Layered approach

# Why is NLP Difficult?

Implicit rating

This film will stay a long time in your DVD cabinet.

Implicit characteristics

My new phone lasted 3 days: Durability –

This camera fits in my pocket: Size +

Figurative language

Si Morandini meurt subitement dans son émission« Vous êtes en direct" on pourra dire qu'il est morandirect.

# Current big challenges for NLP

How can we enable machines to understand the meaning of linguistic expressions in the same way as humans, whatever the source of information?

- Need for a cross disciplinary approach: put linguistic at the core of computational models
- Models should adapt to domains and contexts of anunciation.
- Trendy technique: use deep learning or Large Language Models
- Hypotheses:
  - end to end process in a single model
  - Include these models with pre/post processing in more complex and task oriented pipelines
  - Use representations at various detail levels

# NLP, an evolving domain – Historical trends

The ML 1st revolution: feature-based learning

- Turn a sentence into a sequence of features
- Each word may be represented with more than one features
  - POS; tense and negative/positive form for a verb, sing or plural for nouns, capital letters or normal ones, pontuation, etc
  - Features are selected according to the objective
- Language analysis is transformed into a (vector) clustering issue
- Each sentence or word is represented as a feature vector
- Training data help the system learn which vectors may belogn to which cluster

# NLP, an evolving domain – Historical trends

## The ML 2nd revolution: neural networks, transformers

Hypothesis

- Because meaning cannot be fully characterised, it has to learned
- Word meaning can be (statistically) inferred from its use : distributional semantics
- Statistical language models: most probable word after a sequence of words
- Guess the missing word

The functional interplay of philosophy and ... should, as a minimum, guarantee..
The rapid advance in ... today suggests...
...calculus, which are more popular in ... -oriented schools.

# NLP explosions with LLMs

Large available Data volume

- Not for all languages, domains, tasks

Processing time and storage capacity

- How to design lighter and « greener » models with reduced size, less test runs, and less computation effort?
- How to design models that require less computation capacity at run time?

Scientific papers

- NLP conference have moved from 200 or 300 submission papers to several thousands (4 or 5 000 submissions)
- Ex : ACL confrence accepts around 800 papers to keep a 20% selection rate

# Challenges for NLP at the era of LLMs

## Coarse grain analysis

- Current trend: LLMs make it all!
- The question becomes: How to train / fine-tune LLMs to perform NLP on cybercrime data?

## Closer look at LLMs' performance

- Personal data is kept in neural networks > compliance with RGPD restricts their use
- Many practical / contextual issues
  - Volume (either toofew text or too large flow)
  - short period (3 days) during which processing should be completed
  - very different characteristics of contributions on each network
- LLMs do not perform well
  - In specific domains
  - With poor endowed languages
  - For very precise tasks
- Training and fine-tuning requires (anotated) text: not easy to collect

# LLMs

Neural networks

. sequence reading of sentences

. Words are turned into vectors

. Parameters = number of layers, size of each layer, size of input and output vectors

Transformers

. Can be bi-directional (read the sentence in 2 directions)

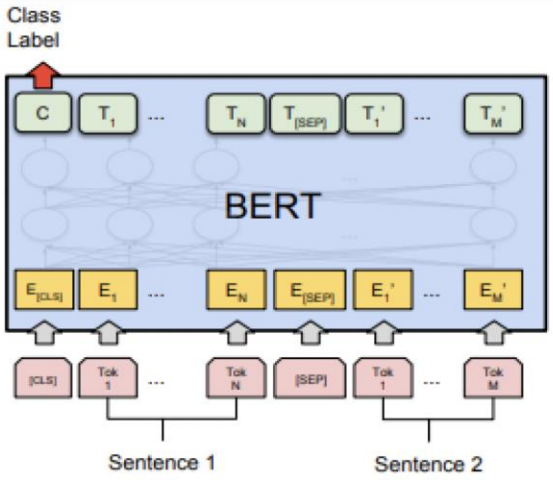
. 2 steps : encoding (from NL to vectors) and decoding (from vector to NL)

. Can include attention mechanisms,

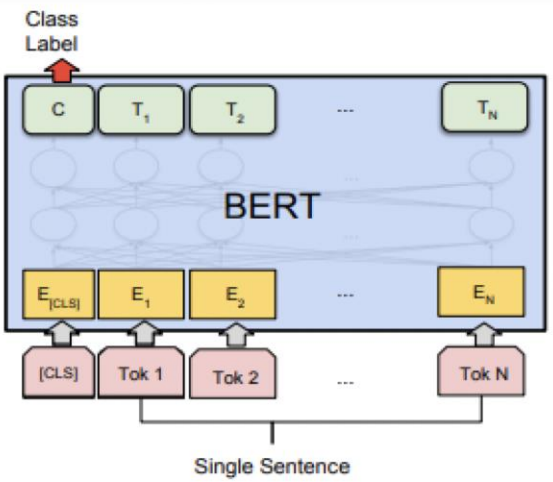
- like a filter that focuses only on some items in the sentence (like verbes, left or right words etc)
- the representation of a sequence (or sentence) is computed by relating different words in the same sequence

Language models can either be a NN, a transformer, an encoder or a decoder

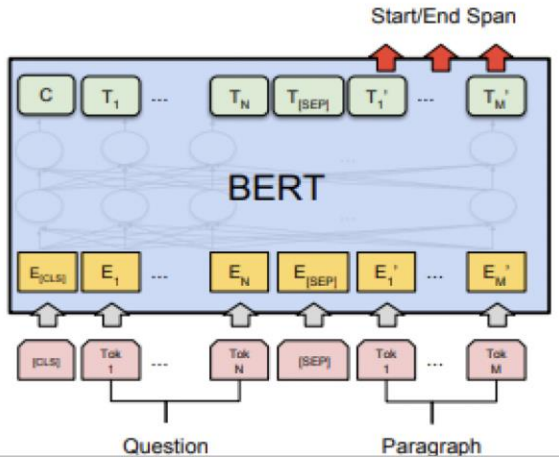
# Example of encoder: BERT

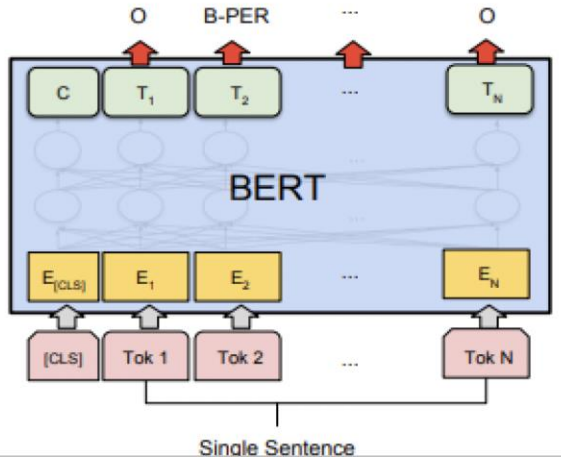



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG




(b) Single Sentence Classification Tasks: SST-2, CoLA



# LLMs: advances and limitations

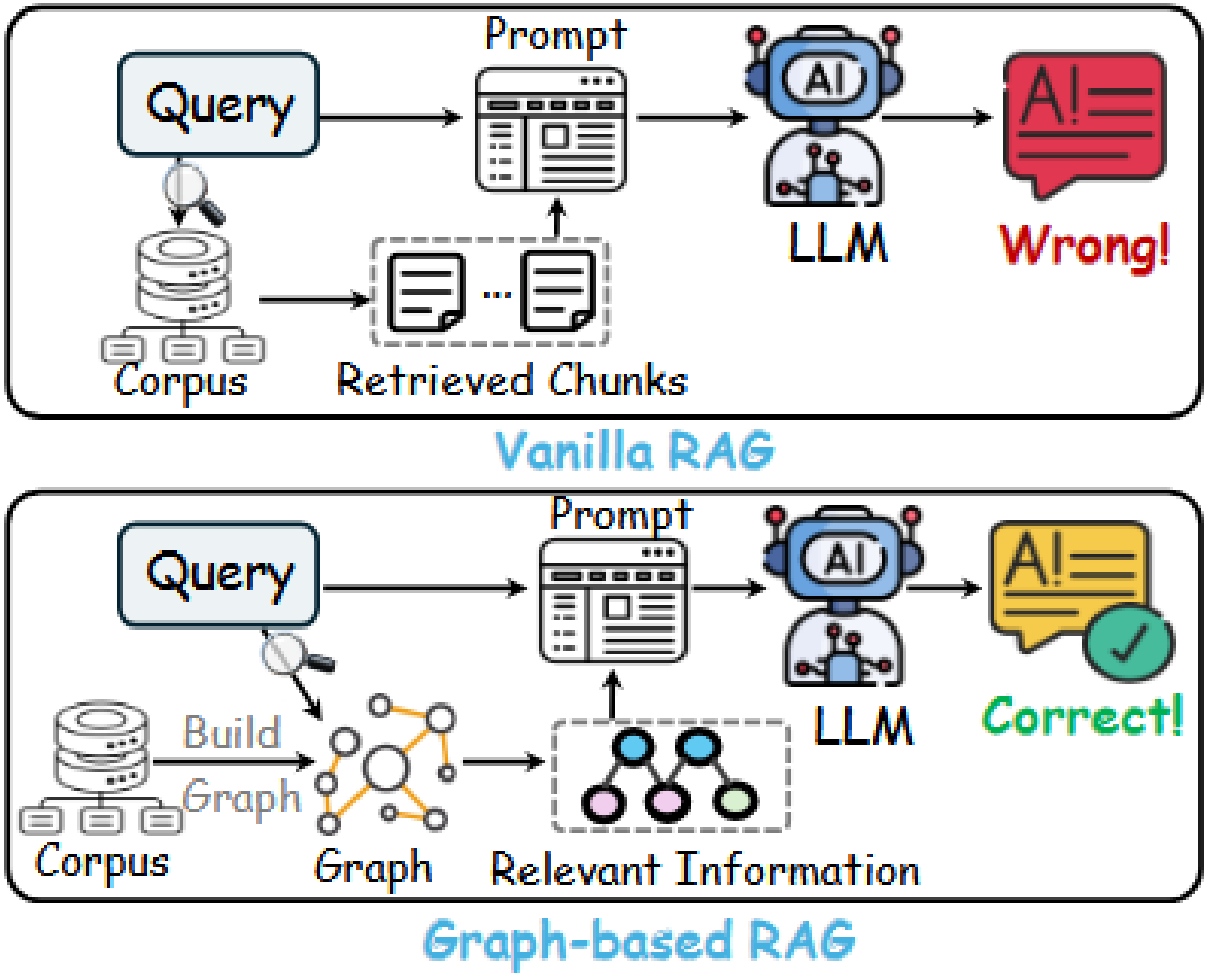
## Advances

- Big progress in tasks like relation extraction (from 30% with patterns to 50% with trad ML to 85% with LLMs and 90% now)
- Seem to adapt to almost any task
- New approaches to reduce costs and errors

## Limitations

- Same as all NN
  - black boxes
  - Biaises, not very robust (ex : table analysis works until the 5th line)
  - Need very large sets of training data
- Language specific issues

# Knowlegde Graphs and LLMs: graph RAGs




## RAG limitations

. Chunks are not connected

. Chunks are small

. Query is small > desambiguation errors

## Graph are expected to

. Contribute to select more relevant chunks

. Provide connected context

. Improve desambiguation

. Add information that is not in the corpus

# Graphs and LLMs for fake news

from KGFakeNet: A Knowledge Graph-Enhanced Model for Fake News Detection (Kumar et al., GenAIK 2025)

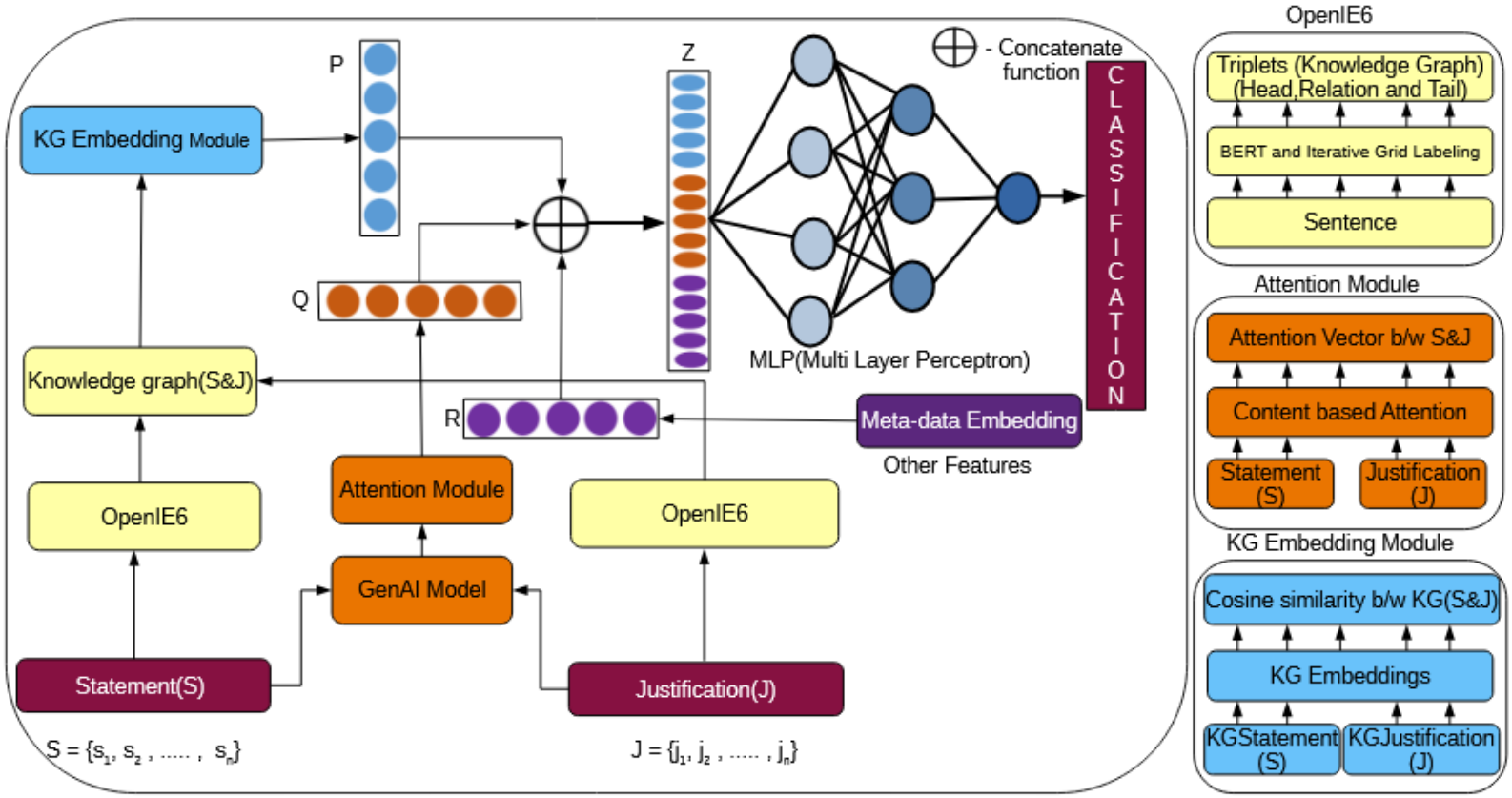



Figure 1: This framework represents the architecture of KGNewsNet computes value P (from TransE embeddings) and Q (from GPT (Black et al., 2022) embeddings) using attention mechanisms. These, along with metadata vector R, are concatenated into $Z$ and passed through an MLP for final classification.

# Example: authorship identification

# Authorship identification, what is the question?

Early works

. identify authors/speakers in political debates /discourses

. Is this document wirtten by the same author as this one?

. Automatically identify citations/quotes in scientific documents

Recent works

. Identify speakers during a meeting (automatic meeting reports)

. Guess who wrote hate speeches

Social needs that require authorship identification

. Identify the person behind various alias

. Find the same author on various social networks

. Identify if one or several people is behind one alias: cf Qanon (https://arxiv.org/pdf/2303.02078)

# Ethic issues using LLMs

Using social Social Network Data when working with the Police (French law)

- No right to access personal data just for surveillance or anticipation purposes
- All rights to access the personal data of suspects in the frame of an investigation.
- You can collect/ analyse a suspect's personal data only during the police custody which lasts a maximum of 3 days.

Training data is hidden in the LLM

- Some layers become « specialized » after various training rounds and can can render the data they have lof the answerearned
- At query time, they can provide pieces of this data in the generated text of their answer.

# Efficiency of basic statistical metrics

Efficient metrics according to the state of the art

. Efficient = green, energy saving AND good results AND easy to compute

. N-grams of characters with n from 1 to 3

. Metadata, like the volume of data posted each hour of the day, whatever the day of the week (or taking into account the day)

. Place and location of the post

. 10 most Frequent keywords

The state of the art deals with one social network at a time

. Need to compare aliases from various social networks

# Using basic statistical metrics

Evaluation of each metric: is it a good comparison criteria?

. Using alter-ego: the contribution set of each alias is splitted into 2 subsets (random selection of the contributions)

. The alter-ego data should be computed as the most similar one

How to combine or agregate results from various criteria ?

. Various possible combination algorithms

- Vote
- Clustering
- Linear combination
- Vector concatenation

. Experimental setting to evaluate each solution

# Autorship attribution: remaining challenges

Impact of the language

. Test in progress with the Czech language (not used to train the LLM)

. Few volume of data is available

Data from various social networks

. Have different types of content: nature, volume, length of each post

. The propotype is able to integrate this data thanks to an ontology

. It is possible to query this integrated data

Test new criteria to get better results

. Alter-ego may be ranked in the 20 first most similar aliases

. Stylistic criteria, relations in the social network etc

# Conclusion – key points to take away

Text and natural language can be

. The target of (social) security attacks

. The means to attack people and their data

NLP has made major advances thanks to

. the increasing volume of digital text

. LLMs, Nureal networks and RAGs

NLP still needs to improve

. For specific tasks

. For poorly endowed languages

. To manage phenomenons

Implementing efficient NLP systems requires expertise

. In natural language, in linguistics and now in machine learning

# Further readings

## LLMs for fake news or hate speech

## Online courses

## Papers

J. Su, C.Cardie, and P. Nakov. 2024. Adapting Fake News Detection to the Era of Large Language Models. In *Findings of the ACL: NAACL 2024*, 1473–1490, Mexico City, Mexico. ACL.

## software

Authorship identification

Online courses

Papers

Software