It’s complicated!

On Natural Language Processing Tools and Digital Humanities

Thierry Poibeau
Tool Criticism Workshop 3.0
DH2020, 20 July 2020

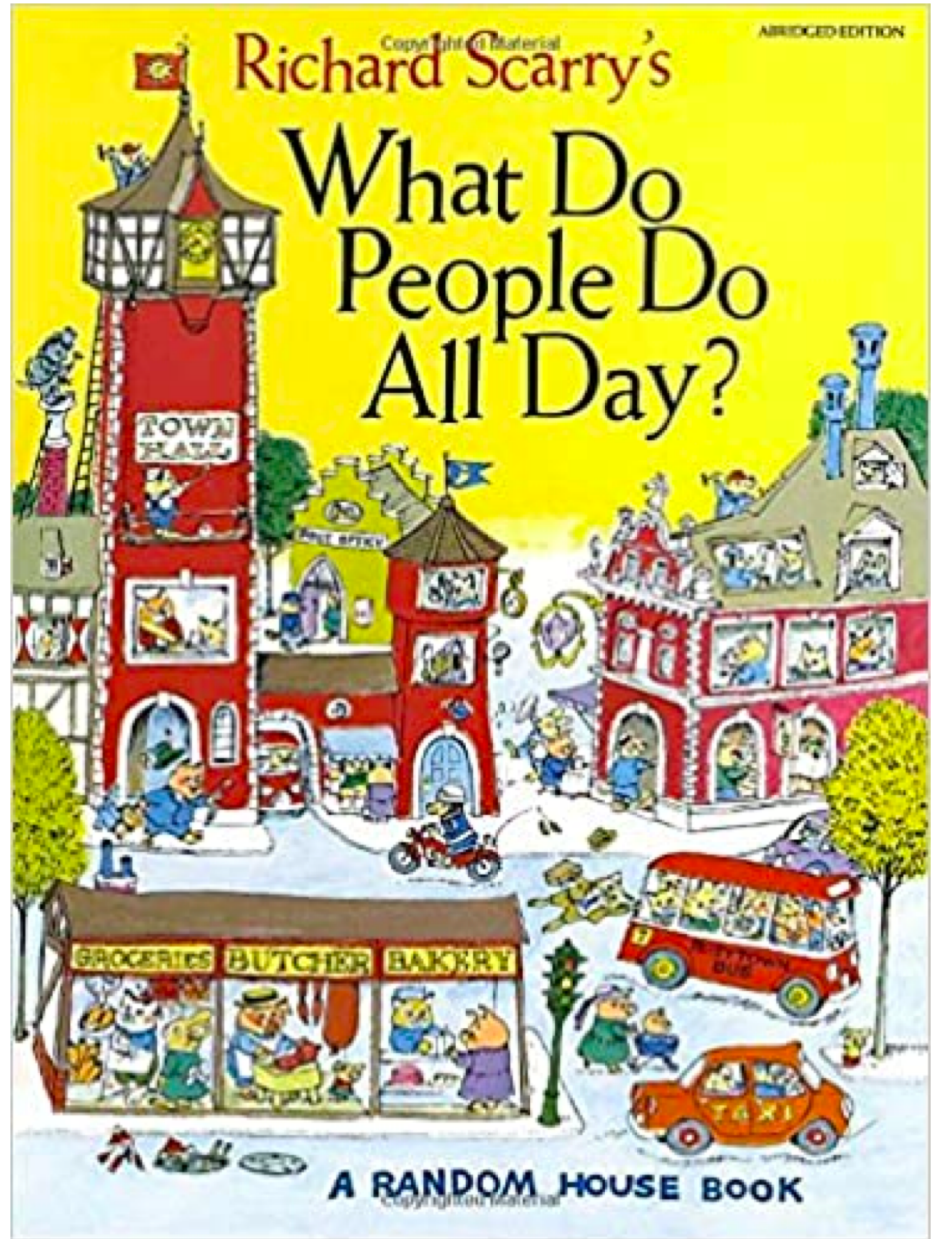
PR[AI]RIE
PaRis Artificial Intelligence Research InstitutE

ENS | PSL

cnrs

Structure of the presentation

- What do people do when they do NLP?
- What do people do when they do DH?
- Some practical experiments
- Conclusion


Richard Scarry's
ABRIDGED EDITION
What Do People Do All Day?
TOWN HALL
POST OFFICE
GROCERIES
BUTCHER
BAKERY
TAXI
A RANDOM HOUSE BOOK

Natural Language Processing

- Goal: Analyse text (aka produce annotations on texts)
  - Part-of-speech tagging
  - Syntactic and semantic parsing
  - Named entity recognition and linking
  - etc.
- Approach: Evaluation-based
  - Specific metrics and reference corpora
  - Main goal is to beat a baseline + beat previous systems

Digital Humanities

- Often based on textual corpora > Use of NLP tools
- Goal: provide a better understanding of a complex question / concept / process
- Approach: Not evaluation-based!
  - Each problem is unique > no standard tasks
  - Generally, no baseline, no reference corpus

- What tools? For what task?
- How can we assess the quality if we can’t evaluate?

NLP tools are not DH tools!

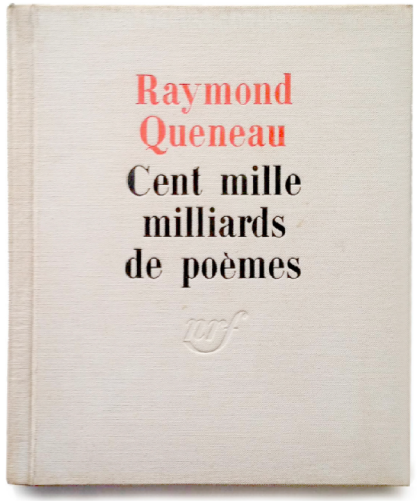
- NLP tools are not trained on DH data > Often not robust on real data
- They do not always provide directly useful annotations in a DH context

- What to do then?

- Some examples

Poetry (rhyme) analysis in French

- Oupoco project: produce new sonnets from the recombination of verses extracted from a corpus of 19th century French poetry
- This project required a proper analysis of rhymes so that new sonnets following rhyming rules can be produced!



Poetry (rhyme) analysis in French

- Dictionaries with phonetic transcriptions are useful
- But not enough!
  - Some sounds are different but can rhyme ([ɛ] ~ [e] ; [ɔ̲] ~ [o])
  - Some pronunciations are the same but words don't rhyme (rimes féminines vs masculines, e.g. words with *–ée* compared to *–é*, *aimé* vs *aimée*)
  - Pronunciation has changed a lot over time
- Conclusion
  - Generic resources are useful
  - But also require some adaptation to the task (poetry analysis)

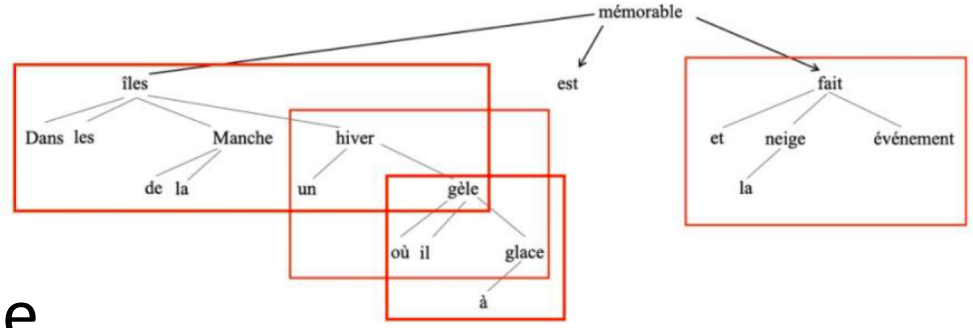
Named Entity Recognition

- Named entity recognition is a key tool for many DH applications
- However, NER systems are highly domain-dependent, i.e. not robust
- Frequent need to retrain on domain specific data
- Active learning can help

« À l'instar de Roberto Martinez PER , il nous reste un goût amer dans la bouche. « C'est vraiment dommage qu'on n'ait pas un stade national pour pouvoir célébrer cette incroyable génération de joueurs pendant l'Euro… ». Et le journal de poursuivre : « C'est même particulièrement navrant. Tête de série, la Belgique LOC aurait pu revendiquer au minimum deux matchs à domicile dans un « Eurostadium LOC » flambant neuf, sis sur le parking C du Heysel LOC . Un projet qui a coûté plus de 25 millions d'euros MONEY pour, in fine, échouer lamentablement, entre querelles à différents niveaux de pouvoir et à relents communautaires. « Retirée par l'UEFA des villes hôtes en décembre 2017 DATE , Bruxelles LOC n'aura donc pas de stade national… et n'en aura sans doute jamais. Aucun gouvernement – à condition déjà d'en avoir un – ne semble près à s'accorder à ce niveau. (…) Bref, une véritable « histoire belge », notre royaume étant, à ce stade, la risée de l'Euro(pe). À charge désormais des Diables LOC de nous éviter de l'être sur le terrain. Y compris à l'autre bout du continent », conclut le journal.

Stylistics, from a syntactic point of view

- Goal: Characterize the style of an author at the sentence (i.e. syntactic) level
  - How regular is the style of an author?
  - How similar / dissimilar to other authors?
  - What are the most typical patterns?



- Parsers provide “acceptable” performance for large scale data
  - For French: UDPipe, Stanford parser, etc.
  - Acceptable = random samples checked manually + formal evaluation of a specific sample

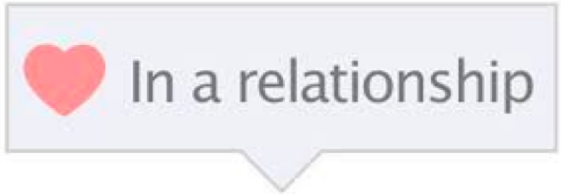
Stylistics, from a syntactic point of view

- What can we do with a parsed corpus?
  - Define and calculate sentence / syntactic complexity > involve defining relevant measures
  - Find specific patterns in a subcorpus > involve defining techniques to extract relevant patterns
  - Check sentence structures > involve transforming syntactic trees into more high-level representations (constituant-based)
- Lots of possibilities / opportunities, but nothing ”given”
  - Transform and re-interpret annotation for the task
  - This is at the same time challenging and rewarding!

Conclusion: What tools, for what purpose?

- Use NLP tools!
  - Provide large scale analysis that would be intractable manually
  - Provide quick and generally accurate results (but error rate must be checked)
- But do it wisely!
  - Tools nearly always need some kind of adaptation
  - More accurate tools appear every year
- Be critical!
  - Tools should be adapted to tasks, not the opposite!
  - Contribute to the Tool Criticism workshop!

Thank you for your attention!



PR[AI]RIE
PaRis Artificial Intelligence Research InstitutE

ENS | PSL

cnrs