Text as Data

Renouveaux de l’analyse textuelle

Mars 2019

Etienne Ollion – CNRS / SAGE

L’analyse quantitative des textes, d’hier à aujourd’hui

- Une initiative précoce (Berelson & Lazarsfelf 1940s, Lebart & Salem 1988, et aussi Benzécri) → « Lexicométrie ».
- Critiques et renouveaux récurrents

L’analyse quantitative des textes, d’hier à aujourd’hui

- Numérisation du quotidien, capacité de traitement,et IA, font qu’est l’objet d’un intérêt renouvelé.

> Cointet & Parasie, « Ce que le big data fait à l’analyse sociologique des textes », *Revue Fr de Sociologie*, 2018.

- Analyse lexicographique « massive » (Michel et al., 2010)
- Dictionnaires
- Réseaux sémantiques (Leskovec 2009, Rule et al, 2015)
- Topic Modelling (Mohr, DiMaggio in *Poetics*)
- Word Embedding (Evans, Kozlowski et al., à paraître)

L’analyse quantitative des textes, d’hier à aujourd’hui

Des usages qui peuvent varier
- Le texte comme « fin » : on l’analyse pour soi
- Le texte comme « moyen » : on crée des variables à partir du texte.

L’analyse quantitative des textes, d’hier à aujourd’hui

C’est que, outre le traitement, on peut se servir de ces méthodes pour produire des données.

- Idée : utiliser le texte comme moyen d’accès à des pratiques, des représentations.

Particulièrement important pour l’étude des médias et de leurs effets potentiels : ne travailler que sur l’écrit occulte radio et télévision.

L’analyse quantitative des textes, d’hier à aujourd’hui

Possibilités :

- Transcrire (signal son → texte)

- Étiqueter du texte
  Topic modelling sur corpus pour lier thèmes & positions
  (ex. Jelveh *et al.*, 2014 : « Political languages in Economics »)

  - Détermination des schèmes du commentaire politique
  Boelaert & Ollion, projet en cours

> Voir Evans & Aceves 2016 ; Gentkow, Kelly & Taddy 2017

Un exemple : le projet CaCoPol

Catégoriser le commentaire politique

Motivation

Une étude des discours <u>sur la</u> politique

– De nombreux travaux sur le discours politique (linguistique, infocom, science politique)

– Mais relativement moins de travaux sur la manières d'énoncer la politique, alors que :

- Le discours politique est très souvent médié
- Les cadres (formels, intellectuels) du journalisme contraignent son expression

> étude du journalisme politique et de sa manière de narrer l'activité politique.

Motivation

Une approche quantitative du discours journalistique

– Objectiver certaines intuitions de la sociologie du journalisme

– Faire une contribution à la littérature en sciences sociales / machine learning.

Recherche en cours

Evolution des schèmes de description de la politique

Au-delà de l'hypothèse de la montée de l'infotainment, quels schèmes (structures sémantiques récurrentes & cohérentes) se sont diffusés, quels schèmes ont perdu en visibilité?

Détour méthodologique

Déterminer des “schèmes”, un tâche complexe

Exemple : le schème du “commentaire sportif” (politique comme stratégie, rapport de entre camps)

- « Un sondage montre un François Hollande de plus en plus affaibli ».
- « Les sarkozystes règlent leurs comptes »
- « Si vous voulez, l’une joue aux échecs, l’autre jette les cartes sur la table »
- « Mais quel calcul politique ! »
- « Mais comment réagit l’Élysée à cette offensive vallsiste ? »
- « Son camp devrait plutôt se réjouir »

Détour méthodologique

Déterminer des “schèmes”, un tâche complexe

Pour les méthodes existantes, une tâche très difficile

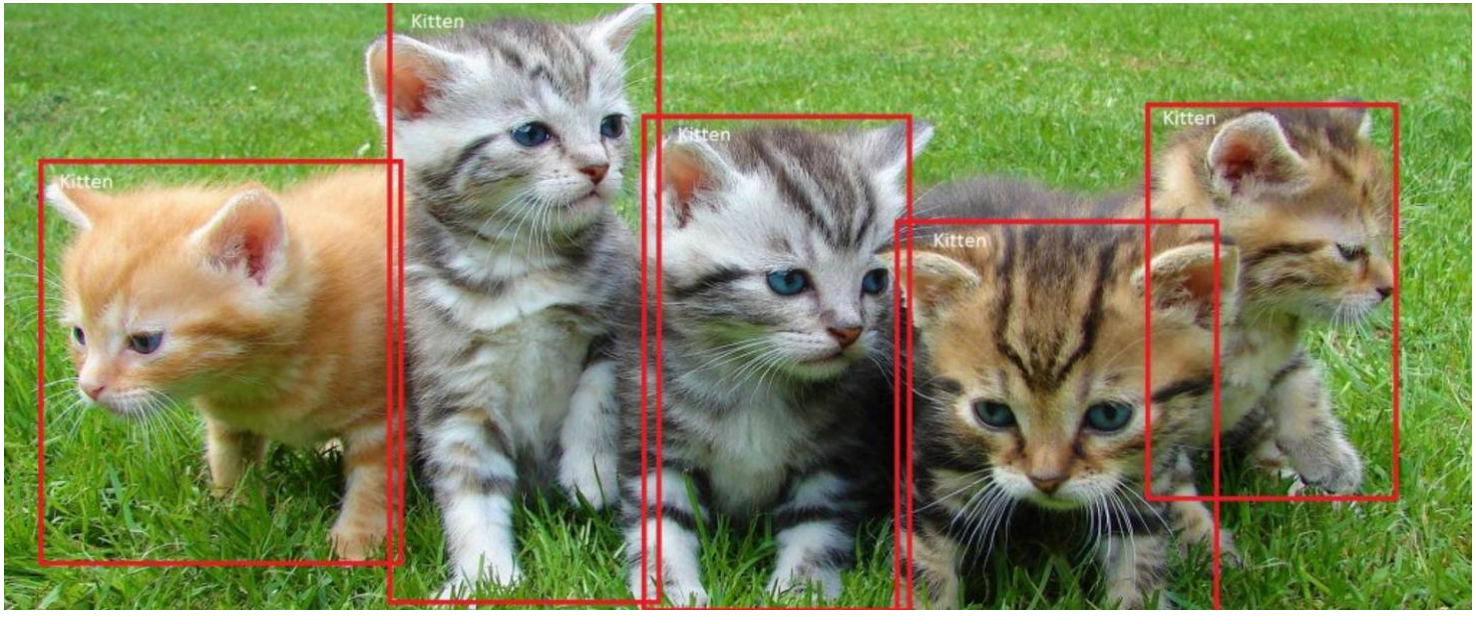
- Pas possible de recourir à des comptes de mots, à des dictionnaires
- Il faut repérer des phrases de structure et avec des mots différents (pas de n-grams, probablement pas de réseaux sémantiques)

→ Il faut une méthode qui ne repose pas sur l’hypothèse « bag of word »

Détour méthodologique

Solution: Recourir à l'apprentissage profond (deep learning)

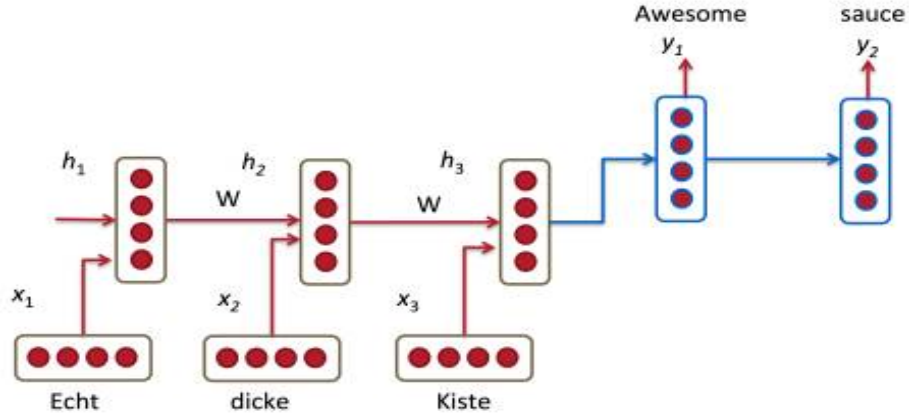
A condition qu'il ait « vu » suffisamment d'exemples, un classifieur pourra reconnaître un schème qu'il n'a jamais vu s'il ressemble aux autres.



Les réseaux de neurones récurrents

Des modèles capables de gérer une chaine de caractères

- Recurrent (RNN, LSTM, GRU…) : des réseaux de neurones avec une boucle
- Technologie qui a permis les avancées récentes en traitement du langage (traduction automatique)
- Karpathy 2015: « The unreasonable effectiveness » of a simple model



Les réseaux de neurones récurrents

Quelques exemples récents

Shakespeare (4.4 MB txt)

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

Algebraic geometry (LaTeX, 16 MB)

For $\bigoplus_{n=1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the compariconly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, **??** and the fact that any U affine, see Morphisms, Lemma **??**. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma **??** we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. □

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, and $\mathcal{T}_i$ is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and $\mathcal{F}_p$ exists and let $\mathcal{F}_i$ be a presheaf of $\mathcal{O}_X$-modules on $\mathcal{C}$ as a $\mathcal{F}$-module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1}\mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)^{opp}_{fppf}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longmapsto (U, \mathrm{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.

Proof. See discussion of sheaves of sets. □

The result for prove any open covering follows from the less of Example **??**. It may replace S by $X_{spaces,étale}$ which gives an open subspace of X and T equal to S_{Zar}, see Descent, Lemma **??**. Namely, by Lemma **??** we see that R is geometrically regular over S.

Les réseaux de neurones récurrents

Une application récente : l'annotation d'entités

- Détecter des catégories dans une chaine de caractères
- Marquer leurs frontières (B: start, I: continue, 0: nothing)

```
John  lives in New   York  and works for the European Union
B-PER O     O  B-LOC I-LOC O   O     O   O   B-ORG    I-ORG
```

Le modèle est entrainé à partir d'exemples, puis reproduit le codage. Pourrait reproduire cela sur des schèmes politiques (si entrainé spécifiquement)

Les schèmes recherchés

(liste provisoire et révisable)

Commentaire sportif

La politique comme stratégie et les rapports de force

Prophétie

L’auteur annonce le futur (probable) du champ politique

Absence d’alternative

L’action est (était) inévitable ou nécessaire

Avis autorisé

Le journaliste donne explicitement son avis

Porte-parolat

Le journaliste enrôle des entités pour soutenir son argumentation

En pratique

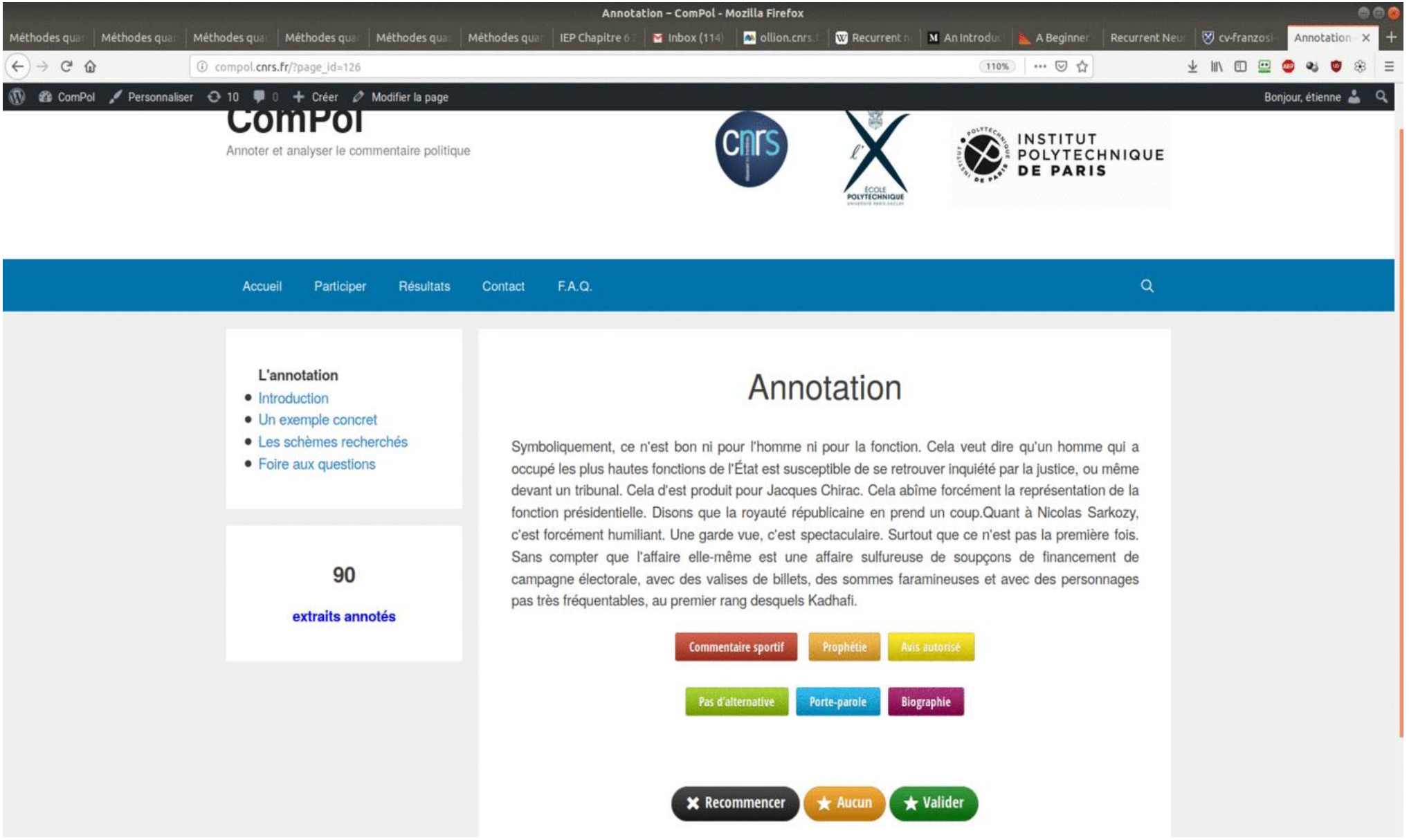
– Collection de 2100 éditoriaux radio transcrits (RTL, Europe 1, France Inter, France Culture)
  – Le choix de l'oral, le choix de l'éditorial
  – Période 2013 à 2018

– Découpage en phrases d'environ 200 caractères.

- Requiert un volume d'étiquetage important
→ Création d'une interface d'annotation en ligne (Ben Dhiaf, Garinet, Landreau, Meyer, Vital)

CaCoPol

Mars 2019



Nos questions

– L’idée vous paraît-elle intéressante/pertinente?
– Quels travaux, quelle littérature consulter
– Les catégories sont-elles pertinentes
– Annotation: quelle stratégie?
  - Qui annote?
  - Quelle stratégie d’enrôlement (crowdsourcing?)

– La plateforme sera, *in fine*, agnostique (n’importe quel corpus, n’importe quelle catégorie, n’importe quelle langue): utile pour vous?