Towards the Automatic Processing of Language Registers: Semi-supervisedly Built Corpus and Classifier for French

Gwénolé Lecorvé, Hugo Ayats, Benoît Fournier, Jade Mekki, Jonathan Chevelu, Delphine Battistelli, Nicolas Béchet

Context and objectives

- Socio-linguistics : community-specific langage (Ure, 1982 ; Biber & Conrad, 2009)
- NLP would be useful for text analytics & for natural langage generation
- Proximity with style processing and sub-language studies
  - Authorship attribution (Stramatatos, 2009 ; Iqbal et coll., 2013)
  - New medias (Gianfortoni et coll., 2011 ; Cougnon et Fairon, 2014)

→ Very few on language registers/formality

- Considered registers: **casual, neutral, formal**
- How to build a first : - Labelled dataset?
  - Classifier?

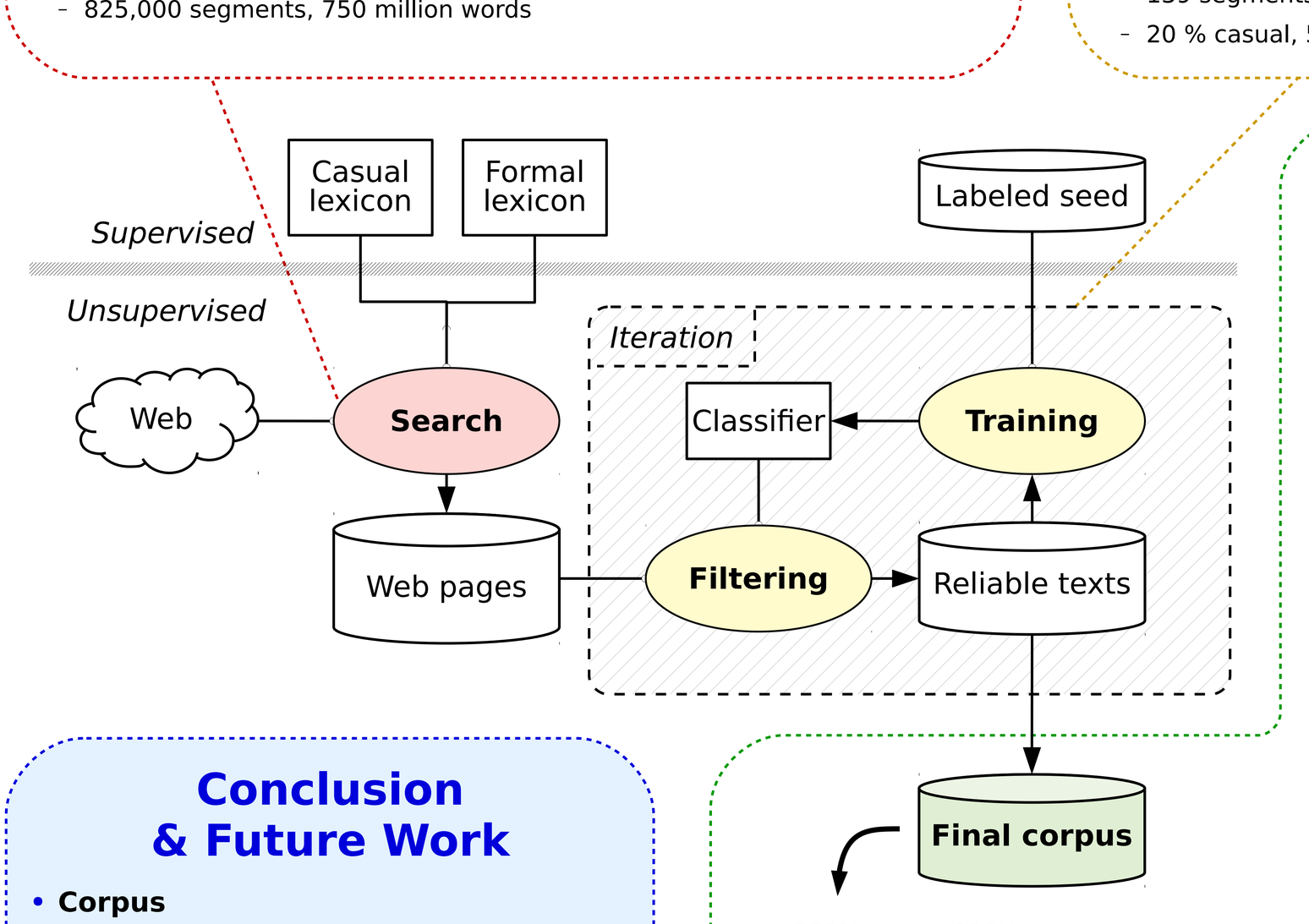
→ Joint iterative semi-supervised construction:
Self-training based on a manually labelled seed and a large collection of web pages

1. Web page collection

- **Specialized lexicons extracted from wiktionary**
  - Simple terms and expressions
  - Selection of unambiguous terms
  - ~6,000 casual, ~300 formal
- **Queries** = 2-6 words → 12,000 queries in total
- **Pages collected**
  - Bing API, 50 pages max. / queries
  - Only 24% of requests with no hit
  - 400,000 pages (33 pages / query)
- **Pre-processing**
  - HTML cleanup + normalization
  - 5000-character segments (paragraph boundaries)
  - Exclusion of non-French segments
  - 825,000 segments, 750 million words

2. Self-training

- **1 segment → 46 global features**
  - Relative frequencies of lexical, phonetic, morphological, morphosyntactic, syntactic phenomena
  - E.g.: casual/formal terms (strictly or with weights), word endings, syntactic errors, etc.
- **Classifier** = feed-forward neural network (2 hidden layers)
- **Filtering** = threshold on the predicted class probability (from 0.7 to 1.0)
- **Manually annotated seed**
  - Romans, journaux, pages web
  - 435 segments, 440 000 mots
  - 33 % of each register (joint labelling of 2 annotators)
  - Training/development/test : 40 / 20 / 40 %
- **2nd test set : subset of the web pages**
  - 139 segments, manually annotated
  - 20 % casual, 50 % neutral, 30 % formal



- 750M words, 825K segments
- 28 % formal, 48 % neutral, 24 % formal

3. Results

- **Classifier accuracy** (best model, i.e. best threshold)

Test

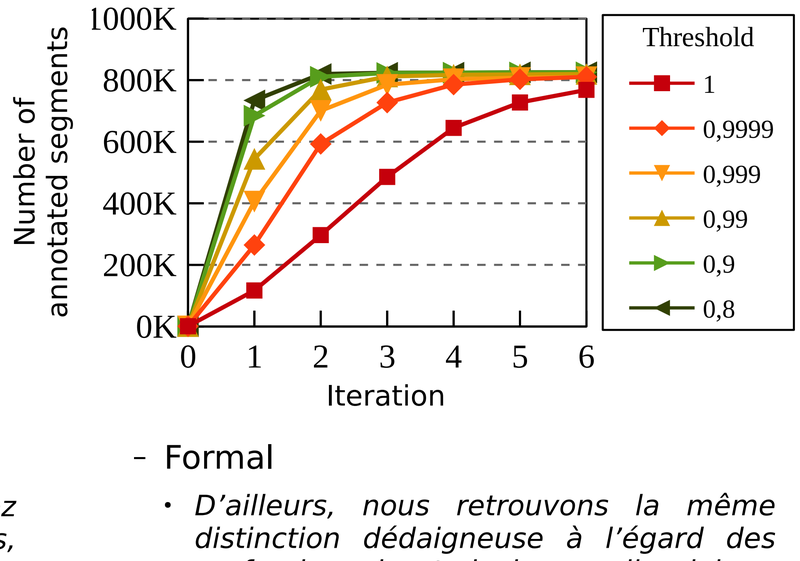
	Casual	Neutral	Formal
Recall	.90 ☺	.78 ☹	.93 ☺
Precision	.84 ☺	.90 ☺	.87 ☺
F-measure	.87	.83	.90

Difference : seed is stereotypical

Labelled web subset

	Casual	Neutral	Formal
Recall	.53 ☹	.72 ☺	.45 ☹
Precision	.52 ☹	.64 ☺	.61 ☺
F-measure	.52	.68	.52

- **Corpus size after each iteration**



- **Excerpts**
  - Neutral
    - *Oui, Monsieur Adrien Richard, si vous aimez mieux, le directeur de l'usine, mais nous, nous ne l'appelons que Monsieur Adrien, parce qu'on a été à l'école ensemble et qu'il nous appelle aussi par notre prénom.*
    - Yes, Mr. Adrien Richard, if you like better, the director of the factory, but we only call him Mr. Adrien because we went to school together and he also calls us our first name.
  - Formal
    - *D'ailleurs, nous retrouvons la même distinction dédaigneuse à l'égard des professionnels et de leur « vil salaire » qui ne les empêche pas de mourir « ès hôpitaux », chez le docte Muret.*
    - Moreover, we find the same disdainful distinction with regard to professionals and their "vile salary" which does not prevent them from dying "in hospitals", at the doctor Muret's.

Conclusion & Future Work

- **Corpus**
  - Large enough to train advanced models
  - Noise in automatic annotations: low enough to train first models
- **Classifier**
  - Good on easy data
  - Worse on ordinary data (especially on casual texts), but better than random
- **Perspectives**
  - Move from binary to real-valued annotations → Reannotate the seed
  - Augment the set of features
  - Define a better filtering criterion (based on risk estimation?)
  - Train sequential models

This work has been benefited from the financial support of the French National Research Agency (ANR) through the TREMoLo project (ANR-16-CE23-0019).