Présentation d'article par Nathan Galard

Marion Baranes. Vers la correction automatique de textes bruités: Architecture générale et détermination de la langue d'un mot inconnu.

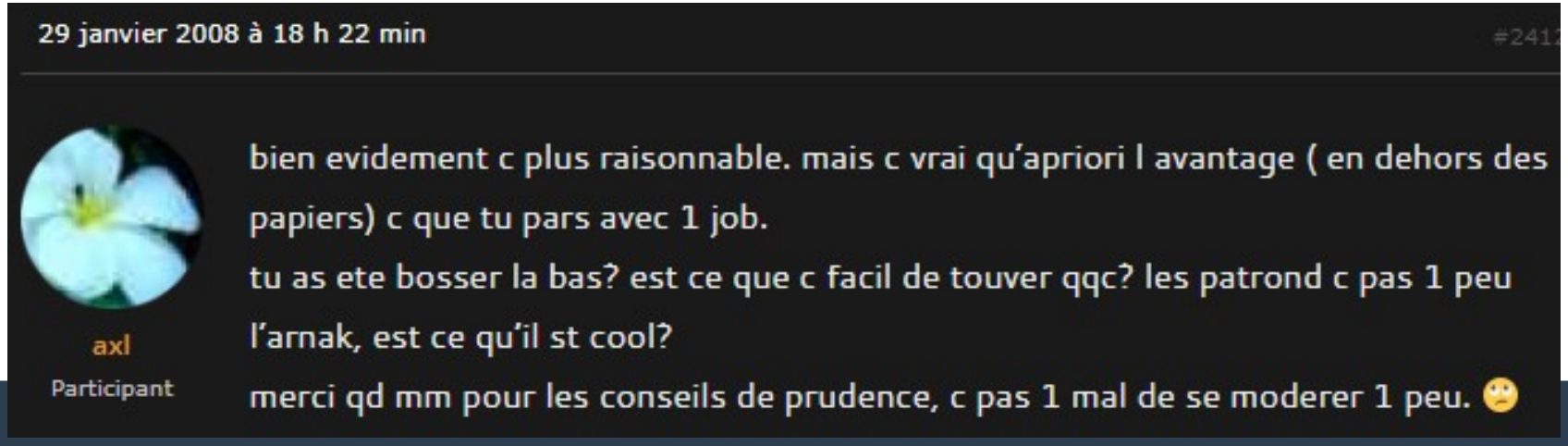
RECITAL'2012 - Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, Jun 2012, Grenoble, France. pp.95-108, 2012. <hal-00701400>

Sujet/objectif 1

- Correction orthographique et typographique automatisée sur textes de qualité très dégradée
- Multilingue
- Contextuel (id est « s'adaptant »)

Prérequis pour nombreux outils

/!\ sur-correction versus sous-correction



www.australia-australie.com/sujet/odysse-agri/
Consulté le 28/01/19

Sujet/objectif 2

- Plusieurs modules, un doit déterminer si mot inconnu étranger ou pas :
  - Faute d'orthographe
  - Emprunt
  - Néologisme
  - Mot étranger
  - Entité nommée

agreement :'(

OMG !

reseter

jumper

surkiffant

Mâchouillon

Tro klasssssss

Jyväskylä

Objet

- **Textes bruités provenant du web**
  - Forums
  - Sites d’avis
  - Réseaux sociaux
  - Blogs
  - (Mails)

Corpus messages clients
(canaux variés)

État de l’art de la correction automatique 1

- **Type/provenance de corpus (ex : OCR vs SMS)**
- **Faute lexicale vs faute grammaticale (vert-vers)**
  - Traitées séparément (hors de pri)
  - Modèles de langage n-grammes (tjs hor de pri)
  - n-grammes phonétiques [ou] les deux
- **Phonétisation (sms)**
- **Ressources lexicales spécifiques**
- **Modèles entraînés sur textes bruités + nettoyée**

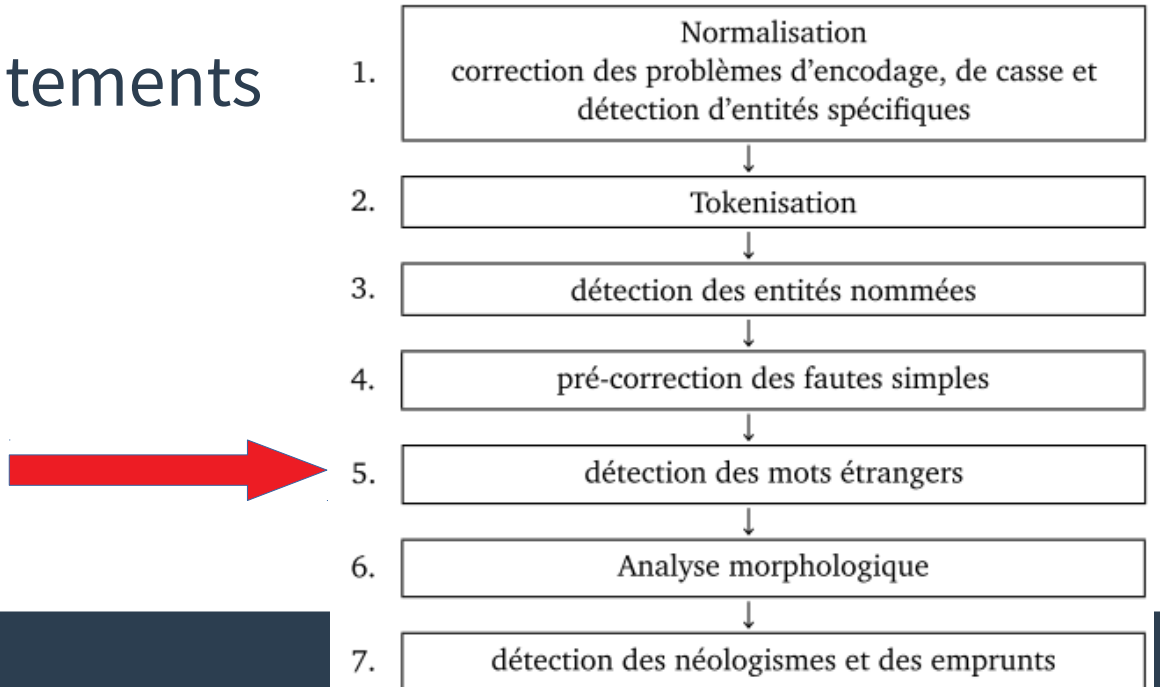
État de l’art de la correction automatique 2

- **Méthode enrichie par résultats de requêtes**
- **Mots inconnus**
  - Lexiques spécifiques
  - Contexte

- **Même logique : détecter > liste candidats > choisir**

Moyens généraux

- **Implémentation d'une architecture de correction**
  - En plusieurs modules
  - Grammaticale et lexicale
  - Combinaison de plusieurs méthodes très variées
  - Série de prétraitements



Les prétraitements n°1 ; 2 ; 3 et 4

- **Normalisation** (encodage, casse, entités spécifiques)
- **Tokenisation (segmentation en unités)**
- **Détection des entités nommées**(noms propres, acronymes)

  → « Nettoyage » ; SxPipe (chaîne modulaire et paramétrable qui applique des traitements de surface)
- **Pré-correction des fautes univoques**
  - Noooooooon ! > Non !

Module de détection des mots étrangers

(Mots inconnus ~~entités nommées~~ > correction ?)

- **≠ détection langue (contexte)**

- **Étranger = emprunt non-adapté morphologiquement**

≠ inconnu français ni néologisme

- **Étranger ≈ anglais → distinguer inconnus fra. / anglais**

- **Classifieur : 2 classes**
  - Corpus d'entrainement
  - Syst. de clas. pouvant apprendre modèle probabiliste
  - Corpus d'évaluation représentatif

Les corpus d'entrainement et système de classification

- **Ressources lexicales** (**Le*fff* et EnLex**)
- **Corpus de textes propres** (**journaux, Wiki, livres**)
- **Corpus bruités produits par l'utilisateur** (**WaCKy > frWaC et ukWaC**)
  - Contiennent mots étrangers > approximation peu dérangeante
- **Baseline, class. naïve > fréquence brute tokens**
- **Système proposé**
  - N-grammes
  - Fréquence-ratio
  - t-test

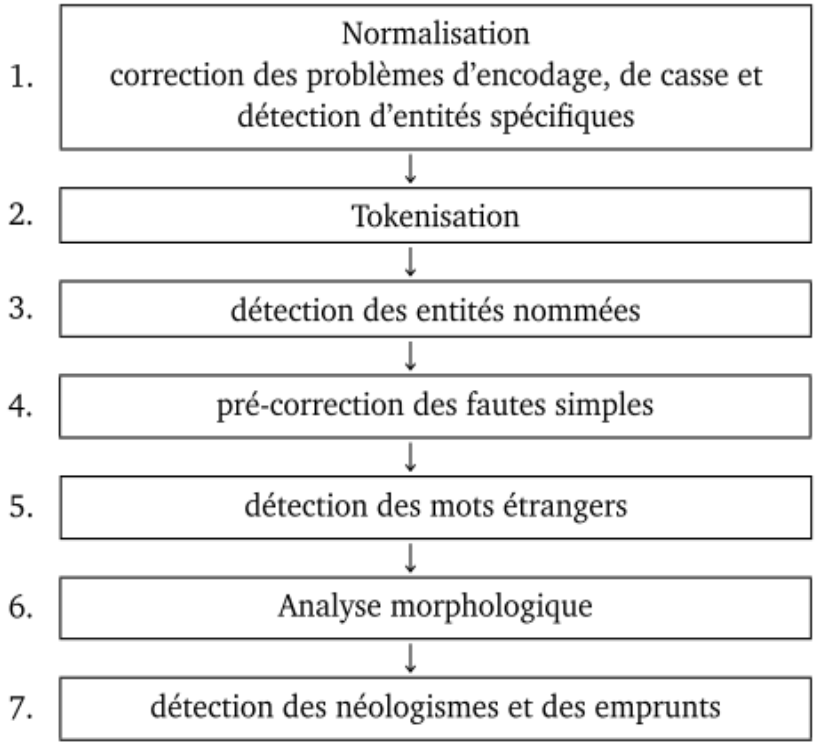
Évaluation

- **Données de réf. récoltées et annotées manuellement 564 x2**
- **Évaluation isolée**
- **Taux d'erreur :**
  - 0,136 Baseline
  - 90 % de bonnes classifications → « satisfaisant »

    (2-grammes et t-test)

Recontextualisation, modules 6 et 7

- **Module 5 : dois-je corriger ?**
  - Oui : analyse morphologique 6 → détection 7 (faute, emprunt, néologisme)
  - Si encore inconnu → « étranger »
- **Prétraitements → correction**



Conclusion

- **User-Generated Content et TAL : outils mal adaptés**
- **Besoin de normalisation et correction**
- **Danger : sur-correction**
- **Module détection mots absents dictionnaire de réf.**
- **Résultat satisfaisants : >90% bonne classification**
  - Sans contexte !

Perspectives

- **Autres modules de prétraitement**
  - Analyseur morphologique
    - Déterminer nature de ces termes

Merci de votre attention !