Talismane - atelier pratique

Axe CARTEL, UE TAL, Master LITL

Assaf Urieli

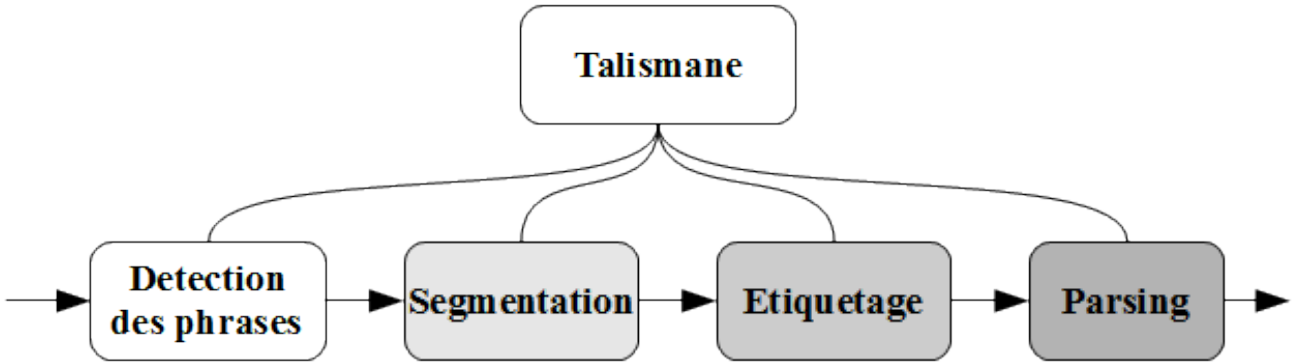
CLLE-ERSS - UMR 5263
Université Toulouse Jean Jaurès

24 janvier 2017

1 Talismane, niveaux d'annotation

Traitement **A**utomatique des **L**angues par **I**nférences **S**tatistiques **M**oyennant l'**A**nnotation de **N**ombreux **E**xemples

- Développé à CLLE-ERSS
- Apprentissage automatique supervisé
- Quatre modules, tous statistiques



Talismane : détails

- `http://redac.univ-tlse2.fr/talismane`
- Sources : `https://github.com/joliciel-informatique/talismane`
- Wiki : `https://github.com/joliciel-informatique/talismane/wiki`
- Logiciel libre (open source) en Java
- Prêt à l'emploi pour le français et l'anglais
- Paramétrable :
  - Filtres pour parser le XML, HTML, et autres formats
  - Descripteurs configurables avec syntaxe expressive
  - Règles : pour forcer ou interdire des décisions locales
- Avantages :
  - Il atteint (voire dépasse) les autres analyseurs statistiques actuellement disponibles pour le français
  - Syntaxe très expressive pour l'écriture des descripteurs et des règles
  - Options de configuration (ex. faisceau + propagation)

Segmentation en mots (Tokenisation)

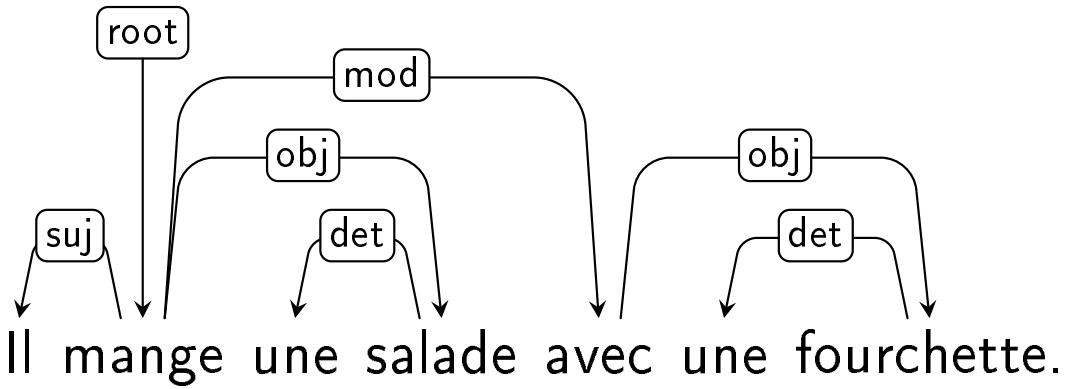
- **Méthode :**
  - Reperage des unités polylexicales
  - Module statistique (pas simplement des listes des mots composés)
  - Liste de patrons : pour une unité reperée par un patron, séparer ou non ?
- **Exemple :**
  - « Tu sais, *bien_que* je ne travaille pas le dimanche, je ferai une exception. »
  - « Tu sais *bien que* je ne travaille pas le dimanche. »

Etiquetage morphosyntaxique (Pos-tagging)

- **Méthode :**
  - Analyse séquentielle, avec descripteurs :
    - N-grammes
    - Catégories grammaticales du lexique pour les mots *n*, *n+1*, *n+2* (cf. Denis et Sagot, 2009)
    - Traits ad-hoc (ex. "que" après "ne")
  - Post-traitement : Ajout des lemmes et des informations morphosyntaxiques sous-spécifiées (genre, nombre, . . .)
- **Exemple :** *La prison ferme ses portes.*
  - La(DET,le,fs) prison(NC,prison,fs) ferme(V,fermer,P3s) ses(DET,son,3s_p) portes(NC,porte,fp).

Parsing : Analyse syntaxique en dépendances

- Parsing par transitions (Nivre, 2008)
- Complexité linéaire
- Pour prendre une décision entre un gouverneur et un dépendant potentiel, descripteurs sur :
  - Les deux mots, leur POS, leur lemmes, ...
  - La phrase : les mots qui séparent les deux mots, ...
  - Les structures syntaxiques partiellement construites



❷ Apprentissage automatique supervisé

Apprentissage automatique supervisé

- **Corpus annoté** : le linguiste, le guide d'annotateur
- **Transformer en classification** : l'ingénieur TAL
- **Descripteurs** : le linguiste, puis convertis par l'ingénieur en format informatique
- **Classifieurs** : boîtes noires : SVM linéaire, perceptrons, deep learning (?)
- **Evaluation**
  - Corpus d'apprentissage et de test
  - Cross-validation

Apprentissage automatique supervisé : Commandes

- `train` : entraînement
  - entrées : corpus annoté, descripteurs, ressources
  - sorties : modèle statistique
- `evaluate` : évaluation
  - entrées : corpus annoté, modèle statistique
  - sorties : fichiers d'évaluation
- `analyse` : analyse
  - entrées : corpus non (ou partiellement) annoté
  - sorties : corpus annoté
- `process` : traitement de corpus
  - entrées : corpus annoté
  - sorties : corpus annoté différemment
- `compare` : comparaison de deux corpus
  - entrées : deux corpus annotés
  - sorties : fichiers d'évaluation

❸ Ressources

Ressources de base pour le français

- **Corpus d'apprentissage** : Corpus français SPMRL (Seddah et al. 2013)
- **Lexique** : LeFFF (Sagot et al., 2006), GLàFF (Sajous et al. 2013)

Ressources : Filtres

- Filtres de texte :
  RegexMarkerFilter OUTPUT <Title>.*?</Title>
  RegexMarkerFilter SKIP (<Bold>|</Bold>|<Sent>|</Sent>)
  RegexMarkerFilter SENTENCE_BREAK </Sent>
- Filtres de tokenisation :
  TokenRegexFilter(posTag=NPP)

  \b(\pWordList(Countries-en))('s)? Civil Aviation (Safety)?Authority\b
- Filtres de casse :

  LowercaseKnownFirstWordFilter

Ressources : Règles

- !P Not(LexiconPosTag("P"))
- !NC HasClosedClassesOnly()
- !ADV Word("que") & IsNull(HistorySearch(Word("ne"))) & Not(FirstWordInSentence)

Ressources : Descripteurs

- Étiquetage de "que" : le verbe précédant sous catégorise "que" comme objet direct (ex. assurer, penser)
- Parsing de la coordination : Étiquettes mal assorties :
  N ADJ et **N**
  vs
  N **ADJ** et **ADJ**

Apport du lexique

- **Étiquetage morpho-syntaxique** : ajout des lemmes et autres traits en post-traitement
- **Descripteurs** : à base de lexique (étiquetage, parsing)
- **Règles** : pour l'étiquetage morpho-syntaxique

Références bibliographiques :

- Abeillé, A. et Clément, L. (2003). *Building a Treebank for French*, in TreeBanks, Springer.
- Candito M., Crabbé B. et Denis P., (2010) *Statistical French dependency parsing : treebank conversion and first results*, Proceedings of LREC'2010, La Valletta, Malta.
- Candito M., Nivre J., Denis P. et Henestroza Anguiano E., (2010) *Benchmarking of Statistical Dependency Parsers for French*, in Proceedings of COLING'2010, Beijing, China
- Denis P. et Sagot B., (2009) *Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort*, in Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC), Hong-Kong.
- Nivre J. (2008), *Algorithms for Deterministic Incremental Dependency Parsing*, Computational Linguistics, 34(4), 513-553
- Sajous F., Hathout N., et Calderone B. (2013), *Glàff, un gros lexique à tout faire du français*. TALN 2013.
- Seddah D. et al (2013). *Overview of the spmrl 2013 shared task : A cross-framework evaluation of parsing morphologically rich languages*. In Proceedings of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages : Shared Task, Seattle, WA.