Modélisation du prétraitement des textes

Thomas Heitz - LRI - heitz@lri.fr - www.lri.fr/~heitz
Université Paris Sud - 91405 Orsay Cedex - France

Définition du prétraitement

① Corriger les incohérences

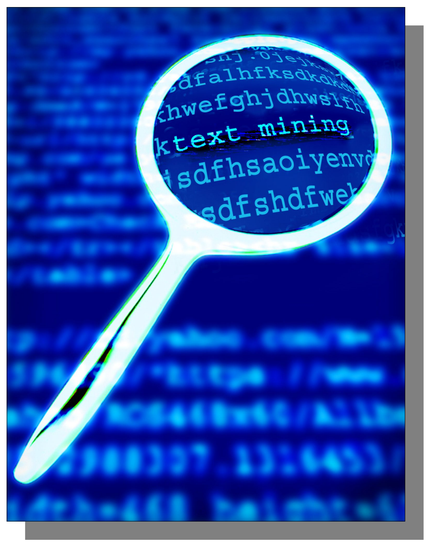
- Format variable d'**encodage**. Ex. *UTF-8 et ISO-8859-1 encodent différemment les accents*.
- **Marqueurs** de présentation ou de sens. Ex. *trouver une partie commune entre des balises XML*.
- Fautes d'**orthographe** ou incohérences typographiques. Ex. *2,3,4,5 points qui correspondent à etc*.

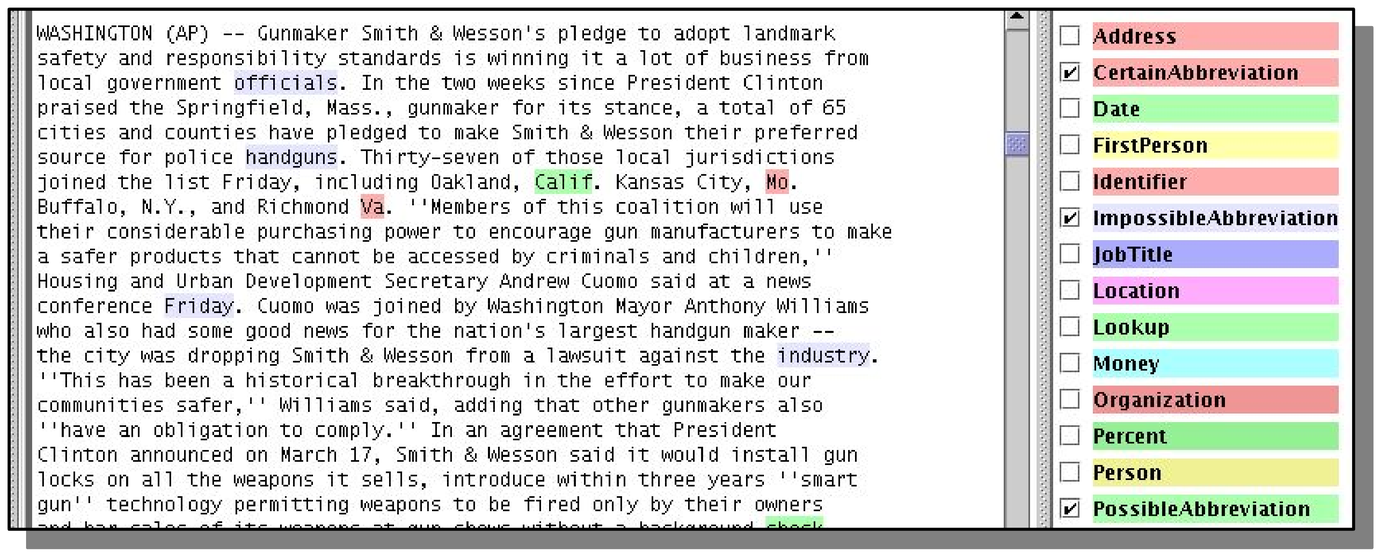
② Expliciter certaines ambiguïtés

- **Ambiguïtés** lexicales. Ex. *développer les abréviations dans leur forme complète*.
- **Structure** des textes. Ex. *utiliser les lignes vides comme marqueurs de fin de paragraphe*.

③ Normaliser les expressions

- **Segmenter**. Ex. *découper un texte en phrases*.
- **Lemmatisation** et groupement des locutions. Ex. *noms d'organismes vivants dans un texte de biologie*.





Dépendances entre les traitements

① Définition

- **Motivation** : décrire les enchaînements de traitements, localiser les maillons faibles de la chaîne de traitements.
- **En pratique** :
  - Rechercher les cycles de **dépendance** entre au moins deux traitements.
  - Vérifier si le cycle est justifié.
  - Trouver la meilleure **condition d'arrêt** du cycle.

② Exemple

- **Segmentation** en phrases d'un texte sur les points suivis d'un espace puis d'un mot capitalisé, c-à-d commençant par une majuscule.
- **Cycle interne** à une étape : recherche des abréviations et des mots toujours capitalisés (seconde étape). Condition d'arrêt : *stabilisation des résultats*.
- **Cycle** entre deux étapes : le traitement des ambiguïtés lexicales (seconde étape) influe sur la segmentation (étape finale). Condition d'arrêt : *un nouveau cycle rajoute plus d'erreurs qu'il n'en corrige*.