Fouille de Textes : de l'utilité d'une approche « orientée linguistique »

Yves Kodratoff
CNRS, LRI Bât. 490, Université Paris-Sud
91405 Orsay, yk@lri.fr

1 . Quelques définitions (dont une pour la FdT)

2. Approche 'en comptage' (stats pures) / approche 'en compréhension' (traitement linguistique préalable)

3. TREC ('Text Retrieval Conference')

4. Notre système

1. DEFINITIONS

1.1. Apprentissage non supervisé/ supervisé

--- Apprentissage non (ou partiellement) supervisé : données non annotées (ou incorrectement annotées) fournies à l'algorithme d'induction.

Exemple : étiquetage 'grammatical'

Phrase d'origine (sans étiquetage) fournie au système .

Aigner also confirmed that in the 1999\2000 season, the champions league would have four qualifying groups of six teams instead of the present six groups of four teams. @/@ 15-APW19980825.1034

****

Phrase obtenue par étiquetage d'un étiqueteur entraîné sur un corpus de textes financiers (+ 'nettoyage' préalable) → phrase incorrectement étiquetée fournie au système

Aigner/NP also/RB confirmed/VBD that/IN in/IN the/DT 1999\2000/CD season/VBP ,/, the/DT champions/NNS league/VBP would/MD have/VB four/JJ qualifying/NN groups/NNS of/IN **six/JJ** teams/NNS **instead_of**/IN the/DT present/NN six/JJ groups/NNS of/IN four/JJ teams/NNS ./. @/@ 15-APW19980825.1034/CD

****

--- Apprentissage supervisé : données correctement étiquetées (ou classées) fournies à l'algorithme d'induction → problèmes liés aux mesures d'efficacité d'un algorithme d'induction :

- problèmes 'faciles' (p.ex. the/DT) (en général précision de 100%)
- problèmes 'difficiles) (p. ex. acts/VBZ ou acts/NNS, p.ex. that/IN ou that/WDT, etc.)

(précision possible : 0%)

Phrase 'correctement' étiquetée fournie au système.

Aigner/NP → NPP also/RB confirmed/VBD that/IN in/IN the/DT 1999\2000/CD ?→? JJ season/VBP → NN,/, the/DT champions/NNS league/VBP → NN would/MD have/VB four/JJ ?→? CD qualifying/NN → JJ (ou VBG ?) groups/NNS of/IN six/JJ ?→? CD teams/NNS instead_of/IN the/DT present/NN → JJ six/JJ ?→? CD groups/NNS of/IN four/JJ ?→? CD teams/NNS ./. @/@ 15-APW19980825.1034/CD

1.2 Recherche documentaire (‘information retrieval’) / Extraction de connaissances (‘information extraction’)

--- La recherche documentaire traite le problème de la pertinence [définie par un ‘dirigeant’] d’un groupe de textes, d’un texte, d’une partie de texte

Exemple 1 : **Reconnaissance du langage utilisé, reconnaissance d’un auteur, reconnaissance de l’état d’esprit d’un auteur.**

Exemple 2 : (**problème de classification, « clustering »)**
Fournir tous les textes relatifs aux relations entre FIFA et American League.

Exemple 3 : **(problèmes de pertinence)**
La phrase ci-dessus est-elle pertinente relativement aux relations entre FIFA et American League ?

--- L’extraction de connaissances

Recueillir des données, des informations ou des connaissances contenues dans un texte, repérer des formes présentes dans les textes.

Exemple 1 : **Systèmes de questions-réponses.**
Exemple 2 : **Lister les propositions du président de la FIFA relativement aux relations entre FIFA et American League**.
Exemple 3 : **Découvrir et annoter des interactions**

regle_6 : The/DT **salm_gene [acts-independently of] abd_A/INTERNULL** ./.

regle_0 : DNA/NN binding/NN studies/NNS in/IN **[yeast_systems suggest-that-the-homeodomain-is_necessary-for]**

abd_A::Ecol_LexA_proteins/*INTERPOSMOU12* to/TO bind/VB to/TO Ubx_sites/NNNT ,/, but/CC the/DT homeodomain/NN does/VBZ not/RB contact/VB DNA/NN exactly/RB like/IN bacterial/JJ helix_turn_helix/NN proteins/NNS **./.**

1.3 La fouille de textes

La fouille de textes est constituée d'un ensemble de méthodologies qui modifient le texte afin d'en préciser le sens dans le but d'améliorer la solution d'un problème spécifique.

Classiquement, on essaie de transformer le texte en tableau de données afin d'appliquer des techniques de fouille de données.

2. COMPTAGES OU LINGUISTIQUE ? : un faux problème

Les pour et les contre d'une approche faisant usage des propriétés linguistiques des textes.

Question réelle : quelles propriétés linguistiques est-il désirable d'utiliser ?

- Au nettoyage : Locutions ou pas ?

123 215 locutions issues de WordNet

P. ex. : La plupart des locutions verbales sont vraies de temps en temps

- Etiquetage grammatical ou pas ?

novel/NN par défaut, et en biologie novel/JJ quasi exclusivement

- Etiquetage fonctionnel ou pas ?

Linguistique de type grammatical ou de type fonctionnel ?
Ford/NomPropre ou /NomGénérique_d'organisation ?

- Terminologie ou pas ?

These/DT observations/NNS suggest/VBP that/IN nuclear/JJ targeting/NN of/IN CBF1/FRM is/VBZ itself/PRP a/DT component/NN of/IN **CBF1-mediated_gene_regulation/NN** and/CC etc.

CBF1-mediated_gene_regulation ou (texte brut) : **CBF1-mediated gene regulation**
ou encore **CBF1-mediated gene_regulation** ?
Etiquetage: NN ou NNP = 'composant des réactions en chimie de la biologie' ?

- Analyse syntaxique, superficielle, profonde, ou pas du tout ?

Exemple d'une phrase toute simple de biologie moléculaire :
These observations suggest that nuclear targeting of CBF1 is itself a component of CBF1-mediated gene regulation and that in the absence of signalling, CBF1 enters the nucleus precommitted to a transcriptional repression function.

- Résolution des coréférences ou pas ?

Si non, alors statistiques sur 'this molecule', 'this site', 'it', etc.

Exemple de l'importance du probleme : TREC, demande par les participants de textes dans lesquelles les coréférences ont été résolues (!)

- Concepts ou pas ?

…

3. TREC (‘Text Retrieval Conference’)

ou : pourquoi faut-il participer aux ‘challenges’ internationaux (nationaux !)

4.1 **Statistiques instantanées** (peut-être non significatives ?)

TREC ‘information retrieval’ : 7 thèmes différents, 103 équipes participantes. Participation française en 2004 : 3 : IRIT (‘novelty’, t.1, t2, t3, t4) , LAMSADE (‘web’), LRI (‘novelty’, t1 et 2),

TREC ‘video’ 1 seul thème, 20 équipes participantes. Participation française : 4.

Problème de financement ?

4.2 **Les thèmes de TREC**

Genomics : 2002- → (**pertinence de textes + d’extraction de connaissances (abandonné !)**)
Novelty : 2002-2004 (**T1 : pertinence et T2 : nouveauté de phrases (les pertinentes étant données)**)
Question-Answering : 1999- → (**extraction de connaissances sur des sujets très variés (p. ex. Place des Club Med aux USA)**)
Web : 1999 – 2004, continue en 2005 avec Industry (**web de compagnies industrielles, au lieu de .gov**)
Terabyte 2004 - → (**en fait _ terabyte**)
Robust 2003 - → (**tâches mal traitées par les autres thèmes**)
HARD (High Accuracy Retrieval from Documents) 1994 - → (**quel type d’information doit on ajouter au textes pour améliorer la précision en pertinence**)
Spams → (2005)

4.3 Les autres compétitions

NTCIR : Compétition organisée par les Japonais :

NTCIR : National Test Collection for Information Retrieval avec

- Cross-Lingual Information Retrieval, multilingual, (Chinois, Japonais, Coréen, Anglais), bilingual, avec langage pivot (intermédiaire entre 2 langages)

- Question Answering challenge (en Japonais)

- Web challenge

CLEF (2000-2005)

Cross-Language Evaluation Forum (2005: DELOS Network of Excellence)

8 competitions , images inclues.

PASCAL – Pattern Analysis, Statistical Modelling and Computational Learning

KDD cup

DEFT 05 - 06

4.4 Pourquoi concourir ?

Permet de savoir

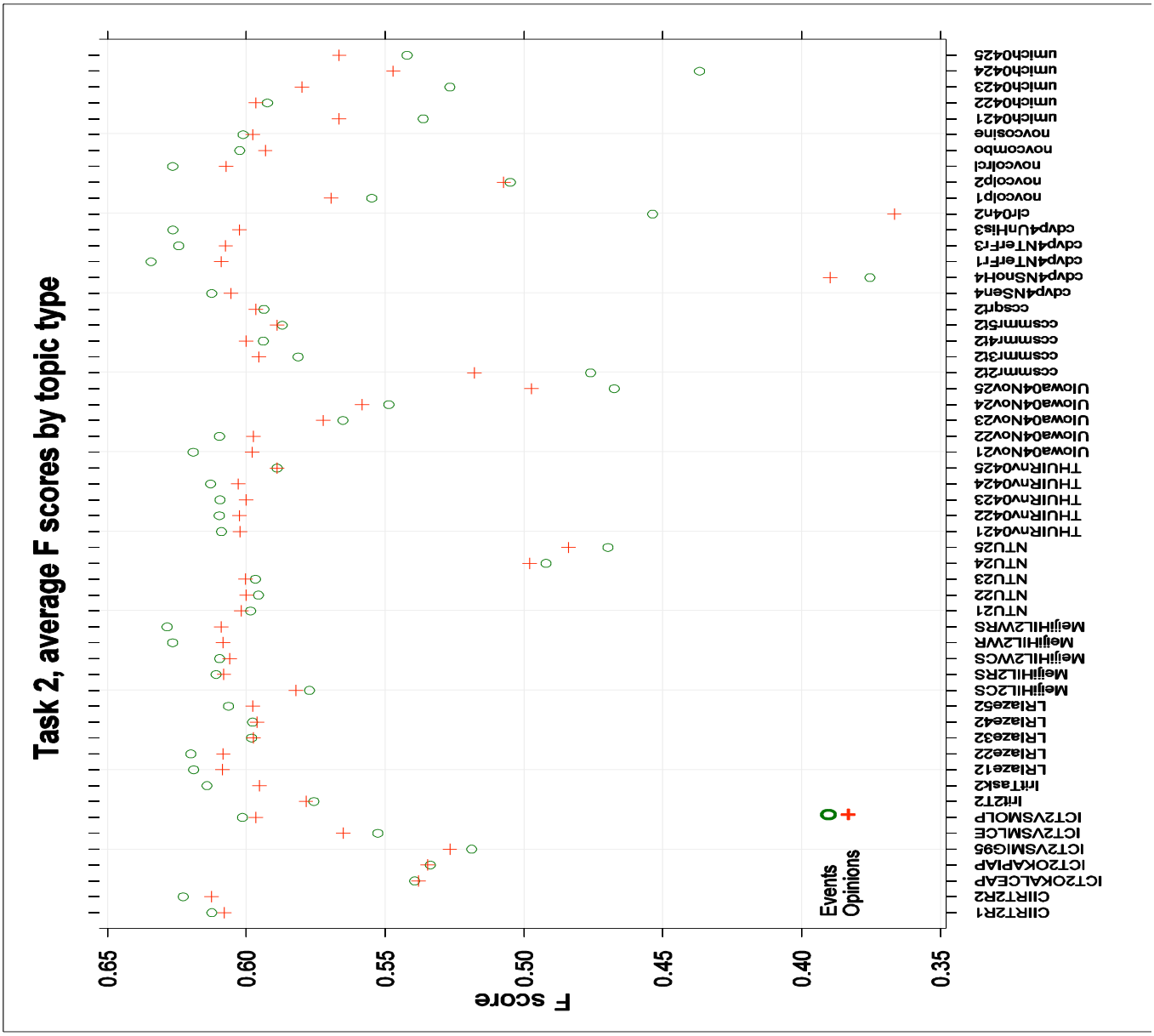
1. si on est loin de l'état de l'art ou non

2. si sa méthode est assez flexible pour s'adapter à des tâches voisines.

3. Au cours des années, savoir si on améliore ou non son système.

Devrait être obligatoire pour les projets financés à une certaine hauteur.

Etrange en particulier que les projets européens n'y soient pas soumis !



4. NOTRE APPROCHE

4.1 Nécessité d'un étiquetage grammatical et 'minimalement sémantique'

- CorTag

Un langage permettant à l'expert du domaine d'écrire 'facilement' des centaines d'expressions régulières relatives au contexte des mots.

Principe : le langage est conçu de sorte que les règles s'auto-commentent.

- ETIQ

Aide à la création d'un corpus 'parfaitement étiqueté' pour que les (puissantes) techniques d'apprentissage n'apprennent pas les fautes d'étiquetage du corpus de départ (!)

On part d'un corpus étiqueté à l'aide d'un étiqueteur généraliste ***Corpus0*** qui deviendra ***SureCorp***$_{i-1}$ dans le traitement iteratif.

L'expert écrit des règles (en CorTag) qui corrigent certains défauts qu'il a remarqué, il applique ses règles à ***Corpus0*** → ***ExpRulCorp***$_0$

Ensuite ETIQ travaille sur 4 versions successives du même corpus: *SureCorp*$_{i-1}$, *ExpRulCorp*$_i$, *IndRulCorp*$_i$ et *SureCorp*$_i$.

• ***ExpRulCorp***$_i$ est obtenu par application règles écrites en CorTag.

- Le système induit des règles à partir des 'contre-exemples' de *SureCorp*$_{i-1}$, et des 'exemples' de *ExpRulCorp*$_i$. (p. ex. C4.5rules)

• ***IndRulCorp***$_i$ = résultat de l'application des règles induites à *SureCorp*$_{i-1}$.

- L'expert examine les differences d'étiquetage entre *ExpRulCorp*$_i$ et *IndRulCorp*$_i$.

• ***SureCorp***$_i$ est le corpus qui conserve les étiquettes issues des règles expertes (si elles ont 'gagné'), les étiquettes induites (si l'induction a 'gagné') ou imposées par l'expert (s'il voit des erreurs en passant – surtout si ces erreurs de contexte expliquent l'erreur de ses règles).

- En principe, l'expert réagit à ses erreurs en modifiant ses règles → *ExpRulCorp*$_{i+1}$

Apprentissage de l'expert : s'il juge que l'induction a 'gagné' de façon judicieuse, il a accès immédiat à la règle induite appliquée et peut l'inclure dans ses nouvelles règles.

4.2 Reconnaissance de présence de concepts dans le texte

Du texte vers les applications en passant la reconnaissance des traces linguistiques de concepts 'importants' présents dans le texte. Se fait en trois étapes.

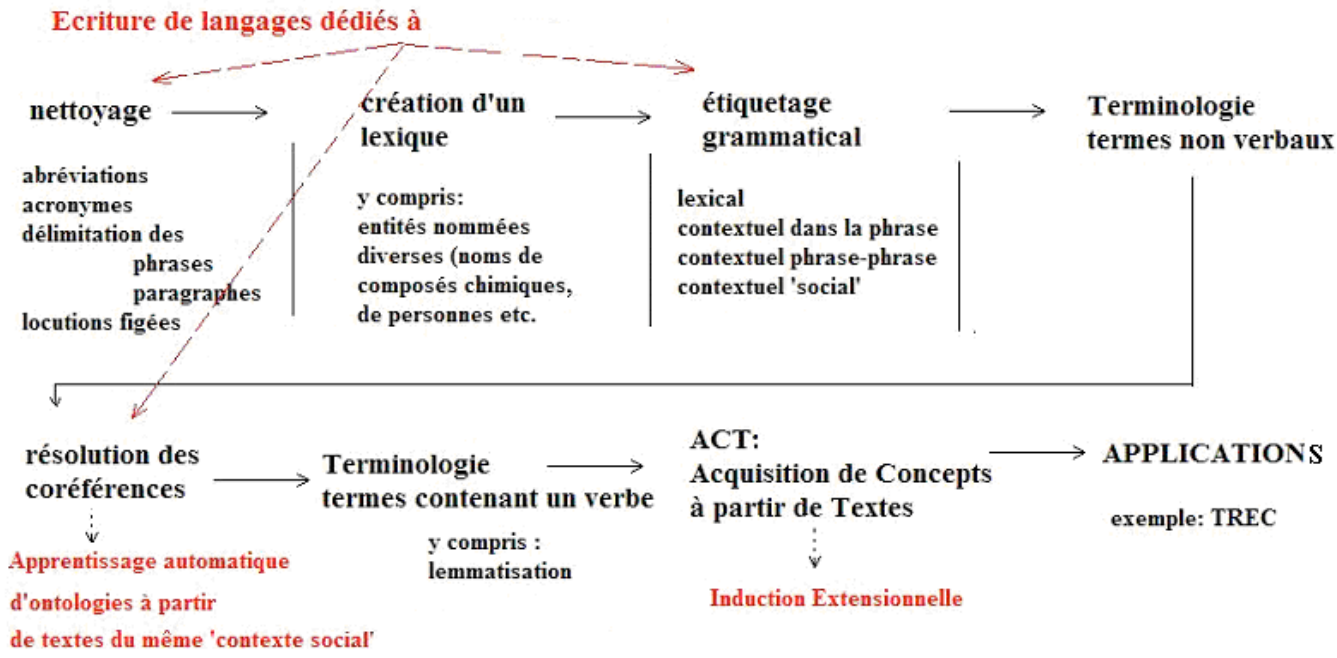
1. L'expert du domaine définit la notion de concept important

P. ex. concepts pertinents pour la tâche considérée

2. Il fournit le plus grand nombre possible d'instances de traces linguistiques de chaque concept :

- **termes nominaux**
- **(termes verbaux)**
- **collocations particulières non repérées à l'étape de terminologie**

3. L'induction extensionnelle complète les traces pour chaque concept



INDUCTION EXTENSIONNELLE

1. Espace des hypothèses

Ensemble de toutes les collocations associées à chaque lexie du texte regroupées par lexie. Chaque groupe doit contenir plus de n collocations, n fixé.

Ex. traces linguistiques d'un concept potentiel inconnu ('relations' ?)

0 (ambiance:Nom,participatif:Adjectif) Classe-4731
0 (atmosphère:Nom,participatif:Adjectif) Classe-4731
0 (esprit:Nom,participatif:Adjectif) Classe-4731
0 (groupe:Nom,participatif:Adjectif) Classe-4731

2. Stratégie de parcours de l'espace des hypothèses : exhaustive.
3. Critère d'optimisation : une mesure de distance composée de 3 distances différentes.
    a. Une distance entre les traces linguistiques fournies par l'expert et un concept potentiel (dépend de l'expert)
    b. Une distance entre mots associés à la lexie de base

(dépend des textes seulement)

c. Une distance évaluant le degré de contradiction lié à l'attribution de la même relation à deux concepts différents.

4. critère de validation : utilité dans les tâches effectuées avec ou sans annotation conceptuelle.