Analyse et expansion des textes en question-réponse

Bernard JACQUEMIN

b_jacquemin@yahoo.fr

7èmes Journées internationales d'Analyse statistique de données textuelles

Introduction

- Des constatations
  - augmentation exponentielle des données textuelles
  - manque de structure dans l'information qu'ils contiennent
  - accès difficile à une information précise
- Solution actuelle : *question answering* par expansion de requête
- Solution proposée :
  - construction d'une structure informationnelle fournissant un accès
  - expansion de l'information plutôt que de la question

Introduction – Exemple

Question : *Qui est le général des Perses ?*

Enrichissement de la question :

universel

chef

Qui est le **général** des Perses ?

supérieur

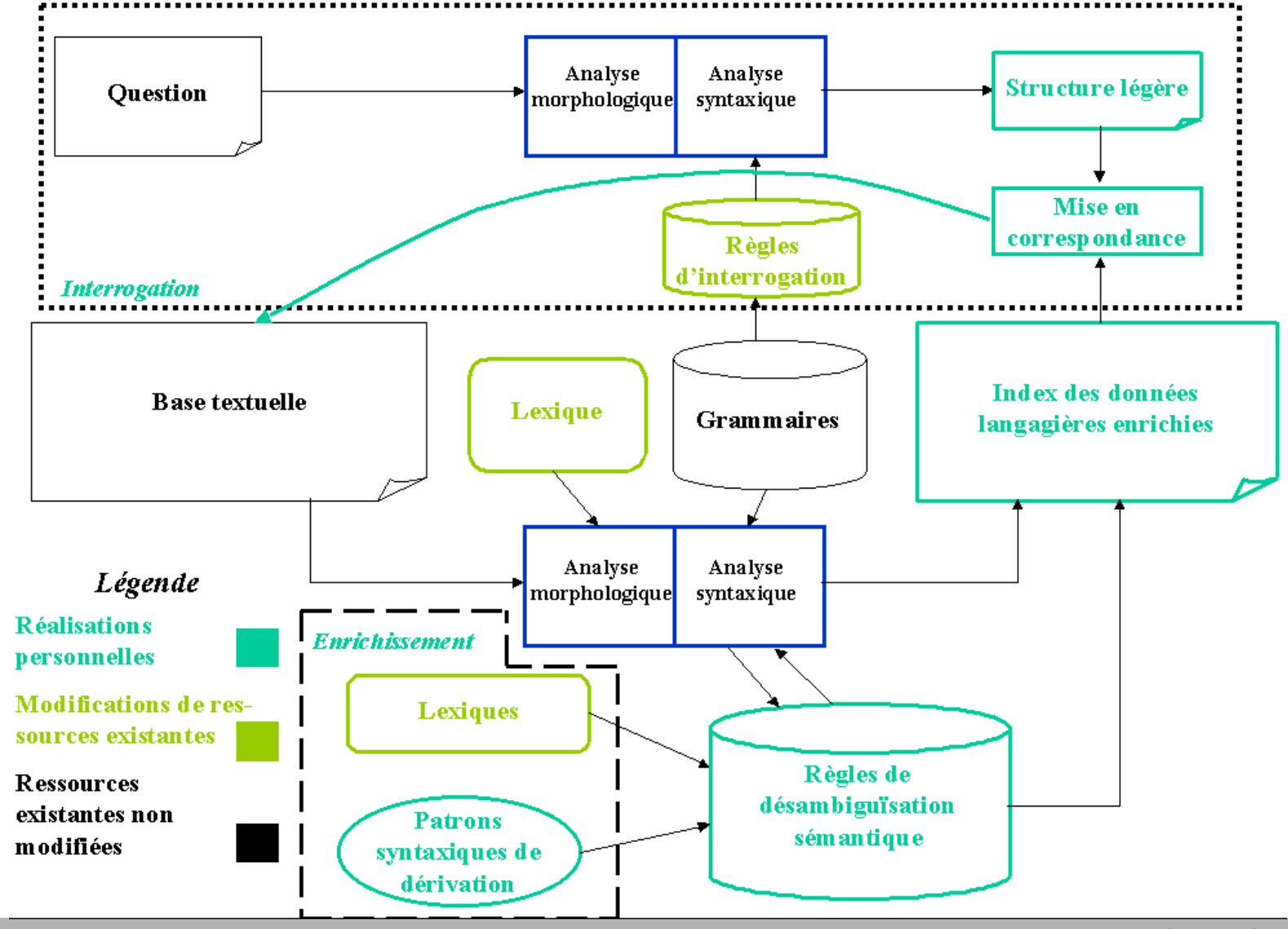
vague

« Dans le chef des Perses, la stratégie était (. . .) »

Plan

- Pourquoi traiter les documents plutôt que la question ?
- Traitements appliqués aux documents
  - analyse morpho-syntaxique
  - désambiguïsation sémantique
  - enrichissement synonymique
  - application de patrons de dérivation
  - indexation des résultats
- Interrogation de la structure
- Évaluation du système
- Perspectives

Architecture du système



Traiter la base documentaire

De quel chef *Domitien est – il le successeur ?*

Document contenant le réponse : article *Domitien* :

Second fils de Vespasien, Domitien succéda à l'empereur Titus et poursuivit la remise en ordre de l'État.

Traiter la base documentaire

	commandant		$héritier$
$De\ quel$	*chef*	$Domitien\ est-il\ le$	$successeur\ ?$
	cuisinier		$dauphin$

Document contenant le réponse : article *Domitien* :

Second fils de Vespasien, Domitien succéda à l'empereur Titus et poursuivit la remise en ordre de l'État.

$\Rightarrow$ utilisation du contexte plus large des documents

$\Rightarrow$ indexation des traitements des documents en une structure de l'information

Traitements – Analyse morpho-syntaxique

- Segmentation et analyse morphologique : NTM (Xerox)
  - découpage du texte en mots
  - proposition d'analyses morphologiques
  - étiquette sémantique pour chacune des analyses proposées
- Désambiguïsation catégorielle et analyse syntaxique : XIP (Xerox)
  - choix contextuel de la catégorie grammaticale
  - construction des dépendances syntaxiques entre lexèmes qui en sont les arguments
  - affectation des étiquettes sémantiques aux lexèmes sous forme de traits
- Stockage du « squelette informationnel » dans l'index sous forme des dépendances entre les mots, et de traits morphologiques, syntaxiques et sémantiques

Traitements – Désambiguïsation sémantique

- Règles de désambiguïsation extraites du dictionnaire *Dubois*
- Conditions d'application syntaxiques et lexicales ou sémantiques
  - patrons lexico-syntaxiques extraits des exemples
  - patrons syntaxiques extraits de la sous-catégorisation
  - généralisation sémantique des patrons lexico-syntaxiques
- Assignation d'un trait de sens au lexème désambiguïsé
- Construction d'une règle de désambiguïsation

Verbe « remporter » au sens « gagner » ***remporter une victoire***

Dépendance de « remporter » : `VARG[DIR](remporter,victoire)`

Règles de désambiguïsation (lexicale – sémantique avec étiquette sémantique MIL pour victoire) :

`remporter : VARG[DIR](remporter,victoire)` $\Rightarrow$ `sens « gagner » n°2`

`remporter : VARG[DIR](remporter,[MIL])` $\Rightarrow$ `sens « gagner » n°2`

Pompée à remporté des ***batailles*** en Orient $\Rightarrow$ *remporté[sens=2]*

Traitement – Enrichissement synonymique

- utilisation des synonymes correspondant au numéro de sens
- adjonction des synonymes dans la structure
- utilisation de dépendances syntaxiques disjonctives

*Second fils de Vespasien, Domitien succéda à l'**empereur** Titus*

Dépendances originales impliquant empereur :

`VARG[INDIR](succéda,empereur)`

`NN(empereur,Titus)`

Synonymes du sens n°1 d'empereur (déterminé par l'analyse sémantique) :

chef, monarque

Dépendances disjonctives :

`VARG[INDIR](succéda,empereur` **OU chef OU monarque**`)`

`NN(empereur` **OU chef OU monarque**`,Titus)`

Traitement – Dérivation morphologique

- Dérivés liés au sens du lexème dont ils dérivent
  - « tracteur » dérive de « tirer » au sens « remorquer »
  - « tracteur » ne dérive pas de « tirer » au sens « faire feu »
- Le dérivé est d'une catégorie grammaticale différente du mot originel
  - table de correspondance des schémas syntaxiques en fonction :
    - de la catégorie originelle
    - de la catégorie du dérivé
    - du type de dérivation (suffixe)

Schéma syntaxique originel de *Domitien* ***succéda*** *à l'empereur Titus* :

`SUBJ(succéda,Domitien) VARG[indir](succéda,empereur)`

Dérivé : « successeur »

Nouvelles dépendances :

$\texttt{NMOD}(\texttt{Domitien}, \mathbf{successeur})$ $\texttt{NARG}[\texttt{INDIR}](\mathbf{successeur}, \texttt{de}, \texttt{empereur})$

Correspondance textuelle :

« Domitien, successeur de l'empereur Titus »

Interrogation

- Analyse morpho-syntaxique
  - identification des lexèmes et des dépendances syntaxiques
  - élimination des caractéristiques propres à la question
  - catégorisation de l'objet de la question
- Construction d'une structure compatible avec la structure informationnelle
- Mise en correspondance de la structure informationnelle avec celle de la question

« De quel chef Domitien est-il le successeur ? »

Structure de la question (dépendances) :

```
NMOD[SPRED](Domitien,successeur)
NMOD[INDIR](successeur,de,chef)
SUBJ(est,Domitien)
FOCUS(chef)
```

Interrogation

~~SUBJ(est,Domitien)~~
~~FOCUS(chef)~~
NMOD[SPRED](Domitien,successeur)
NMOD[INDIR](successeur,de,chef[FOCUS,SOC,COM,HER,humain])

Document contenant la réponse : article *Domitien*

« [. . .] Domitien succéda à l' empereur Titus [. . .] »

SUBJ(succéda,Domitien)
VARG[INDIR](succéda,à,empereur)
NN(empereur,Titus)

SUBJ(succéda OU remplacer,Domitien)
VARG[DIR](remplacer,empereur OU chef OU [.
NN(empereur OU chef OU souverain,Titus)

NMOD[INDIR](successeur,de,empereur OU chef OU [...])
NMOD(Domitien,successeur)

Évaluation

- 200 questions par 8 externes (TREC-8)
- fenêtre : la phrase
- base documentaire : 50 articles d'encyclopédie
- une réponse au moins dans les documents
- baseline :
  - mots-clefs sans enrichissements
  - mots-clefs et synonymes bruts

Enrichissement	Score	Pas de réponse
Plancher	0.295	139
Synonymes (sans sémantique)	0.303	137
Synonymes (avec sémantique)	0.487	100
Tous les enrichissements	0.504	97

Conclusions et perspectives

- Besoins de plus synonymes et d'un moteur logiques
- Dégradation de la requête
  - pondération de la correspondance question-réponse
  - pondération du type de dépendance
- L'approche est généraliste et linguistique
- L'approche fournit une structure de l'information

Merci pour votre attention
Avez-vous des questions ?