Extraction de termes :
présentation des diverses approches

Atelier sur les corpus spécialisés en terminologie
28 avril 2003

Patrick Drouin
OLST-ÉCLECTIK

Plan

- Introduction
- Utilité pour le travail terminologique
- Survol des approches
  - Mécaniques
  - Linguistiques
    - Analyse syntaxique
    - Frontières de termes
  - Statistiques
  - Hybrides
- Conclusion

Introduction

- Présentation des diverses approches
- Explication des techniques utilisées
- Exposé des possibilités et des limites
- Illustration du processus à l'aide de contextes
- Recensement de quelques logiciels
- Références pertinentes
- Description de pistes et non de solutions directement utilisables...

Utilité des extracteurs

- Pourquoi?
  - Rapidité et systématicité
  - Analyse de gros volumes de documentation
- Utilisation potentielle
  - Dépouillement terminologique
  - *Prédépouillement* terminologique
  - Alimentation de dictionnaires électroniques
  - Veille terminologique

Survol : modèles mécaniques

- Technique
  - Objectifs des concepteurs ne sont pas terminologiques
  - Prise en charge de corpus sans prétraitement
  - Recensement des redondances : segments répétés
  - Filtrage des données
- Logiciels
  - http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/lexico3.htm
- Références
  - Choueka, Klein et Neuwtiz (1983); Choueka (1988); Salem (1987); Lebart et Salem (1988, 1994)

Survol : modèles mécaniques

Dans la documentation officielle, lorsqu'on fait état des réserves considérables d'eau souterraine dont dispose le Québec, on souligne également la bonne qualité de cette eau. Cette appréciation semble rallier à peu près tout le monde même si, par ailleurs, on reconnaît qu'on manque de connaissances au sujet de cette ressource.

L'eau naturelle

Premier constat: l'eau naturelle n'est pas toujours pure. En fait, l'eau chimiquement pure ne se rencontre jamais dans la nature. Parmi toutes les nappes d'eau souterraine auxquelles on s'alimente présentement, il en existe une partie, que nous fixerons par hypothèse à 20%, qui ne satisfait pas aux normes physico-chimiques : un ou plusieurs éléments a une teneur qui dépasse la limite acceptable. Elle est donc d’une qualité simplement inacceptable.

Survol : modèles mécaniques

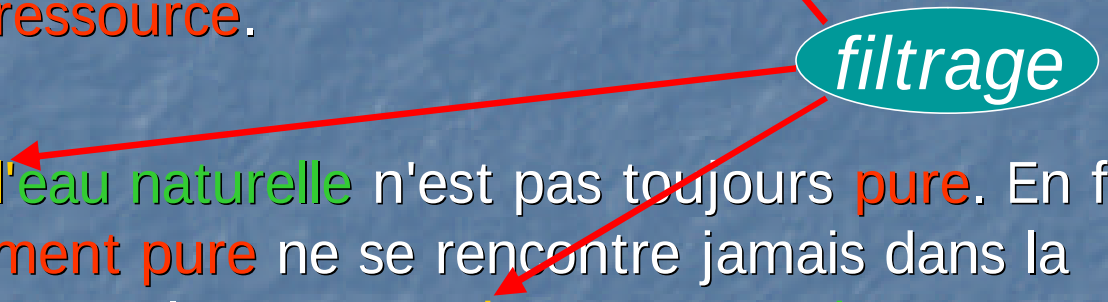
Dans la documentation officielle, lorsqu'on fait état des réserves considérables d'eau souterraine dont dispose le Québec, on souligne également la bonne qualité de cette eau. Cette appréciation semble rallier à peu près tout le monde même si, par ailleurs, on reconnaît qu'on manque de connaissances au sujet de cette ressource.

L'eau naturelle

Premier constat: l'eau naturelle n'est pas toujours pure. En fait, l'eau chimiquement pure ne se rencontre jamais dans la nature. Parmi toutes les nappes d'eau souterraine auxquelles on s'alimente présentement, il en existe une partie, que nous fixerons par hypothèse à 20%, qui ne satisfait pas aux normes physico-chimiques : un ou plusieurs éléments a une teneur qui dépasse la limite acceptable. Elle est donc d'une qualité simplement inacceptable.

Survol : modèles mécaniques

Dans la documentation officielle, lorsqu'on fait état des réserves considérables d'eau souterraine dont dispose le Québec, on souligne également la bonne qualité de cette eau. Cette appréciation semble rallier à peu près tout le monde même si, par ailleurs, on reconnaît qu'on manque de connaissances au sujet de cette ressource.



L'eau naturelle

Premier constat: l'eau naturelle n'est pas toujours pure. En fait, l'eau chimiquement pure ne se rencontre jamais dans la nature. Parmi toutes les nappes d'eau souterraine auxquelles on s'alimente présentement, il en existe une partie, que nous fixerons par hypothèse à 20%, qui ne satisfait pas aux normes physico-chimiques : un ou plusieurs éléments a une teneur qui dépasse la limite acceptable. Elle est donc d'une qualité simplement inacceptable.

Survol : modèles mécaniques

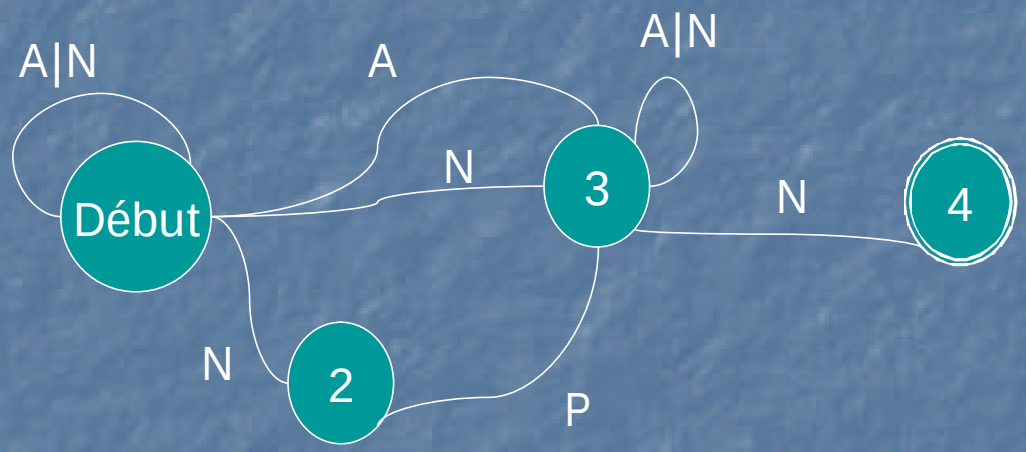
- Avantages
  - Bonne exploitation des corpus
  - Technique très systématique (trop?)
  - Dépistages d'un vaste ensemble de phénomènes plus ou moins linguistiques
  - Extraction indépendante des langues et domaines
- Désavantages
  - Nécessite des corpus de très grande taille
  - Lenteur des traitements
  - Étape non négligeable de filtrage essentielle

Survol : modèles linguistiques (1)

- Technique
  - Recours à l'analyse syntaxique complète ou locale
  - Description morphosyntaxique des structures de termes valides
  - Utilisation de grammaires, d'automates, d'expressions régulières
  - Corpus bruts ou prétraités
- Logiciels
  - Termino/Nomino
  - FASTER
  - NPtool
- Références
  - David et Plante (1990); Plante, Dumas et Plante (2000); Voutilainen (1993); Jacquemin (1997)

Survol : modèles linguistiques (1)

- Automates



- Expressions régulières

((A|N)+ | (A|N)* (N P) (A|N)*) N

- *hard wood*
- *high pressure*
- *high pressure system*
- *government of Canada*
- *research activity*
- *developement of high pressure system*
- *sense based retrieval*
- *...*

Survol : modèles linguistiques (1)

Dans la documentation officielle, lorsqu'on fait état des réserves considérables d'eau souterraine dont dispose le Québec, on souligne également la bonne qualité de cette eau. Cette appréciation semble rallier à peu près tout le monde même si, par ailleurs, on reconnaît qu'on manque de connaissances au sujet de cette ressource.

L'eau naturelle

Premier constat: l'eau naturelle n'est pas toujours pure. En fait, l'eau chimiquement pure ne se rencontre jamais dans la nature. Parmi toutes les nappes d'eau souterraine auxquelles on s'alimente présentement, il en existe une partie, que nous fixerons par hypothèse à 20%, qui ne satisfait pas aux normes physico-chimiques : un ou plusieurs éléments a une teneur qui dépasse la limite acceptable. Elle est donc d'une qualité simplement inacceptable.

Survol : modèles linguistiques (1)

?

Dans la documentation officielle, lorsqu'on fait état des réserves considérables d'eau souterraine dont dispose le Québec, on souligne également la bonne qualité de cette eau. Cette appréciation semble rallier à peu près tout le monde même si, par ailleurs, on reconnaît qu'on manque de connaissances au sujet de cette ressource.

L'eau naturelle

Premier constat: l'eau naturelle n'est pas toujours pure. En fait, l'eau chimiquement pure ne se rencontre jamais dans la nature. Parmi toutes les nappes d'eau souterraine auxquelles on s'alimente présentement, il en existe une partie, que nous fixerons par hypothèse à 20%, qui ne satisfait pas aux normes physico-chimiques : un ou plusieurs éléments a une teneur qui dépasse la limite acceptable. Elle est donc d'une qualité simplement inacceptable.

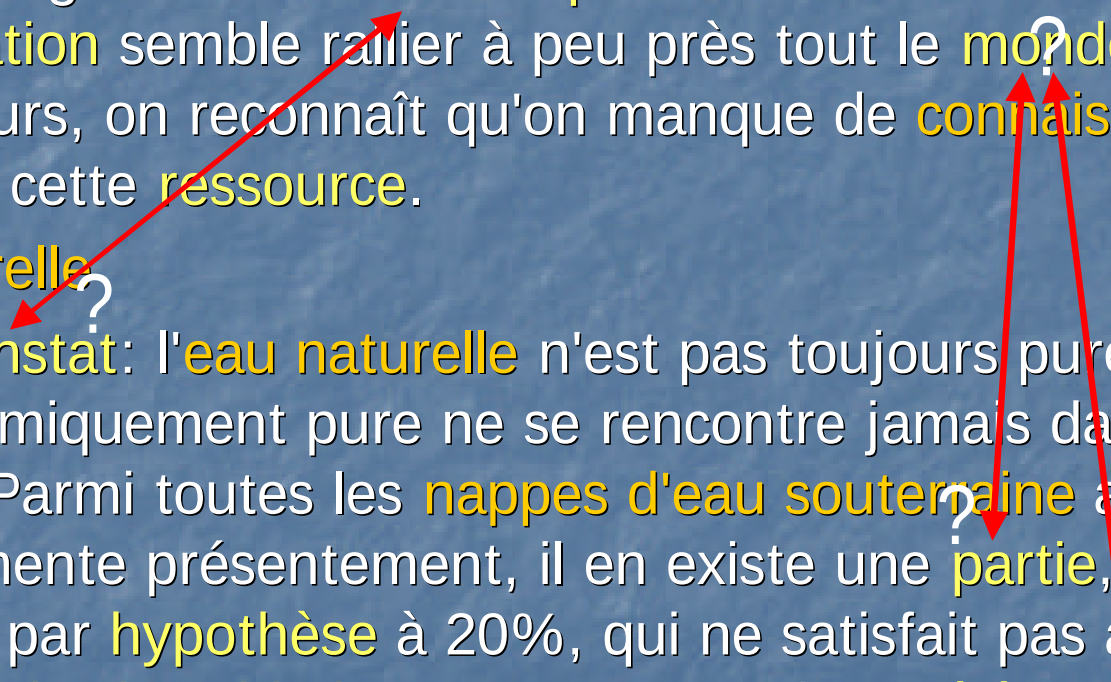
?

Survol : modèles linguistiques (1)

Dans la documentation officielle, lorsqu'on fait état des réserves considérables d'eau souterraine dont dispose le Québec, on souligne également la bonne qualité de cette eau. Cette appréciation semble rallier à peu près tout le monde même si, par ailleurs, on reconnaît qu'on manque de connaissances au sujet de cette ressource.

L'eau naturelle

Premier constat: l'eau naturelle n'est pas toujours pure. En fait, l'eau chimiquement pure ne se rencontre jamais dans la nature. Parmi toutes les nappes d'eau souterraine auxquelles on s'alimente présentement, il en existe une partie, que nous fixerons par hypothèse à 20%, qui ne satisfait pas aux normes physico-chimiques : un ou plusieurs éléments a une teneur qui dépasse la limite acceptable. Elle est donc d'une qualité simplement inacceptable.



Survol : modèles linguistiques (1)

Dans la documentation officielle, lorsqu'on fait état des réserves considérables d'eau souterraine dont dispose le Québec, on souligne également la bonne qualité de cette eau. Cette appréciation semble rallier à peu près tout le monde même si, par ailleurs, on reconnaît qu'on manque de connaissances au sujet de cette ressource.

L'eau naturelle

Premier constat: l'eau naturelle n'est pas toujours pure. En fait, l'eau chimiquement pure ne se rencontre jamais dans la nature. Parmi toutes les nappes d'eau souterraine auxquelles on s'alimente présentement, il en existe une partie, que nous fixerons par hypothèse à 20%, qui ne satisfait pas aux normes physico-chimiques : un ou plusieurs éléments a une teneur qui dépasse la limite acceptable. Elle est donc d'une qualité simplement inacceptable.

?

Survol : modèles linguistiques (1)

- Avantages
  - Prise de décision sur des bases linguistiques
  - Description proche de l'intuition du linguiste ou du langagier
  - Patrons de formation bien documentés
  - Possibilité de prise en charge des diverses réalisations pour une même terme (ex.: *filtration membranaire*)
- Désavantages
  - Description du terme et opposition aux unités nominales non terminologiques
  - Complexité et lourdeur potentielles du traitement
  - Descriptions spécifiques à une langue

Survol : modèles linguistiques (2)

- Technique
  - Recours à des descriptions négatives de structures linguistiques
  - Description de ce qui ne peut être un terme valide
  - Identification des frontières entre les termes potentiels
  - Utilisation de corpus bruts ou prétraités
- Logiciels
  - LEXTER
  - Notions
- Références
  - Bourigault (1992); Drouin et Ladouceur (1994)

Survol : modèles linguistiques (2)

- Frontières hors contexte
  - Ponctuation forte, verbes conjugués, pronoms, etc.

- Frontières contextuelles
  - $V_{part\_passé}$ + P --> coupe + élimination
    vannes de réglage associée à ce générateur
  - $V_{part\_passé}$ + de --> Ø
    vanne motorisée d'isolement

  - DE + Ø + Su--> Ø
    régulation de tempétature
  - DE + LE --> Ø
    relestage du ventilateur principal
  - DE + DET --> coupe + élimination
    influence de ce paramètre

Survol : modèles linguistiques (2)

Dans la documentation officielle, lorsqu'on fait état des réserves considérables d'eau souterraine dont dispose le Québec, on souligne également la bonne qualité de cette eau. Cette appréciation semble rallier à peu près tout le monde même si, par ailleurs, on reconnaît qu'on manque de connaissances au sujet de cette ressource.

L' eau naturelle

Premier constat: l' eau naturelle n'est pas toujours pure. En fait, l'eau chimiquement pure ne se rencontre jamais dans la nature. Parmi toutes les nappes d'eau souterraine auxquelles on s'alimente présentement, il en existe une partie, que nous fixerons par hypothèse à 20%, qui ne satisfait pas aux normes physico-chimiques : un ou plusieurs éléments a une teneur qui dépasse la limite acceptable. Elle est donc d'une qualité simplement inacceptable.

Survol : modèles linguistiques (2)

- Avantages
  - Prise de décision sur des bases linguistiques
  - Descriptions reposent souvent sur des catégories fermées
  - Dictionnaires ou grammaires faciles à élaborer
  - Indépendance face au domaine
  - Rapidité accrue du traitement
- Désavantages
  - Problème des structures de surface identiques
  - Descriptions liées à la langue analysée

Survol : modèles statistiques

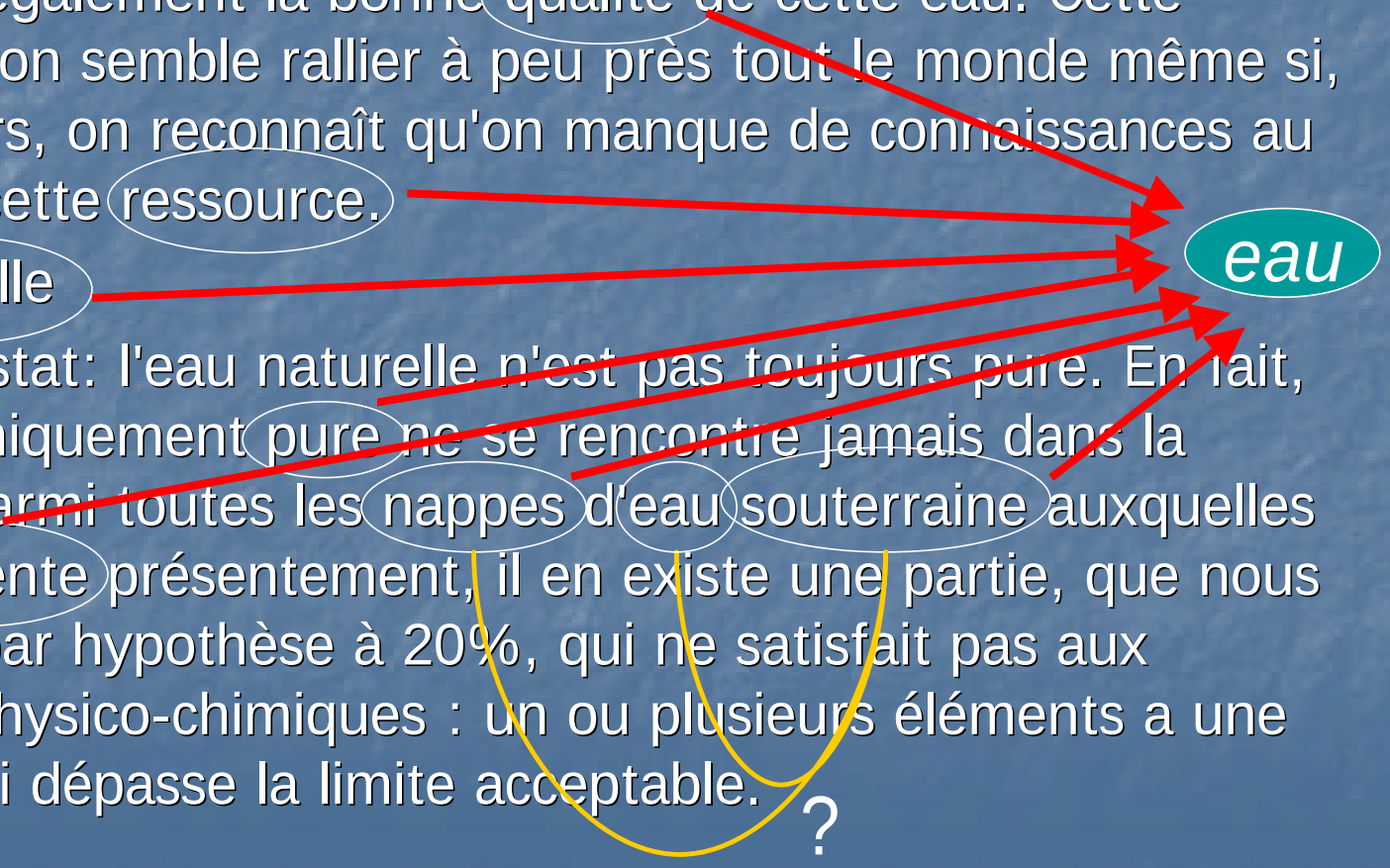
- Technique
  - Indices statistiques qui «mettent en évidence» les tendances à la cooccurrence
  - Filtrage des données
  - Utilisation de corpus bruts ou prétraités
  - Techniques qui ne sont habituellement pas destinées à l'extraction de termes mais qui constituent un bon point de départ pour l'extraction
- Logiciels
  - ANA
  - Divers outils sous forme de scripts
- Références
  - Church et Hanks (1989); Enguehard et al. (1992); Ahmad (1996)

Survol : modèles statistiques

Dans la documentation officielle, lorsqu'on fait état des réserves considérables d'eau souterraine dont dispose le Québec, on souligne également la bonne qualité de cette eau. Cette appréciation semble rallier à peu près tout le monde même si, par ailleurs, on reconnaît qu'on manque de connaissances au sujet de cette ressource.

L'eau naturelle

Premier constat: l'eau naturelle n'est pas toujours pure. En fait, l'eau chimiquement pure ne se rencontre jamais dans la nature. Parmi toutes les nappes d'eau souterraine auxquelles on s'alimente présentement, il en existe une partie, que nous fixerons par hypothèse à 20%, qui ne satisfait pas aux normes physico-chimiques : un ou plusieurs éléments a une teneur qui dépasse la limite acceptable.



Survol : modèles statistiques

- Avantages
  - Identification de phénomènes difficilement observables
  - Rapidité de traitement
  - Extraction indépendante des langues et domaines
  - Exploitation maximale des corpus
- Désavantages
  - Relation très étroite entre le corpus et les résultats
  - Résultats varient pour des corpus similaires
  - Prise de décision sur des bases non linguistiques
  - Étape de filtrage essentielle et importante
  - Corpus analysé doivent être de grande taille (perte d'info)

Survol : modèles hybrides

- Technique
  - Combinaison de méthodes statistiques et linguistiques
  - Jugement sur la pertinence linguistique des résultats statistiques
  - Confirmation statistique de l'importance d'une unité linguistique
  - Utilisation de corpus bruts ou prétraités
- Logiciels
  - ACABIT
  - TERMS
  - XTRACT
  - TermoStat
- Références
  - Daille (1993); Justeson et Katz (1993); Smadja (1993); Lauer (1994); Frantzi et Ananiadou (1995); Drouin (2002)

Survol : modèles hybrides

Dans la documentation officielle, lorsqu'on fait état des réserves considérables d'eau souterraine dont dispose le Québec, on souligne également la bonne qualité de cette eau. Cette appréciation semble rallier à peu près tout le monde même si, par ailleurs, on reconnaît qu'on manque de connaissances au sujet de cette ressource.

L'eau naturelle

Premier constat: l'eau naturelle n'est pas toujours pure. En fait, l'eau chimiquement pure ne se rencontre jamais dans la nature. Parmi toutes les nappes d'eau souterraine auxquelles on s'alimente présentement, il en existe une partie, que nous fixerons par hypothèse à 20%, qui ne satisfait pas aux normes physico-chimiques : un ou plusieurs éléments a une teneur qui dépasse la limite acceptable.

Survol : modèles hybrides

Dans la documentation officielle, lorsqu'on fait état des [réserves considérables d'eau souterraine] dont dispose le Québec, on souligne également la bonne qualité de cette [eau]. Cette appréciation semble rallier à peu près tout le monde même si, par ailleurs, on reconnaît qu'on manque de connaissances au sujet de cette [ressource].

L'[eau naturelle]

Premier constat: l'[eau naturelle] n'est pas toujours [pure]. En fait, l'[eau] chimiquement [pure] ne se rencontre jamais dans la nature. Parmi toutes les [nappes d'eau souterraine] auxquelles on s'alimente présentement, il en existe une partie, que nous fixerons par hypothèse à 20%, qui ne satisfait pas aux [normes physico-chimiques] : un ou plusieurs [éléments] a une [teneur] qui dépasse la [limite acceptable].

Survol : modèles hybrides

- Avantages
  - Mise en évidence de phénomènes difficilement observables
  - Rapidité de traitement
  - Exploitation maximale des corpus
- Désavantages
  - Étape de filtrage essentielle
  - Corpus analysé doivent être de grande taille (perte d'info)
  - Dépendance potentielle à la langue traitée
  - Relation très étroite entre corpus et résultats

Conclusion

- Solution magique inexistante
- Situations avantageuses d'utilisation
- Modération des attentes ...
- Extraction de «candidats termes»
- Travaux sur le caractère terminogène des candidats termes
- Extraction bilingue