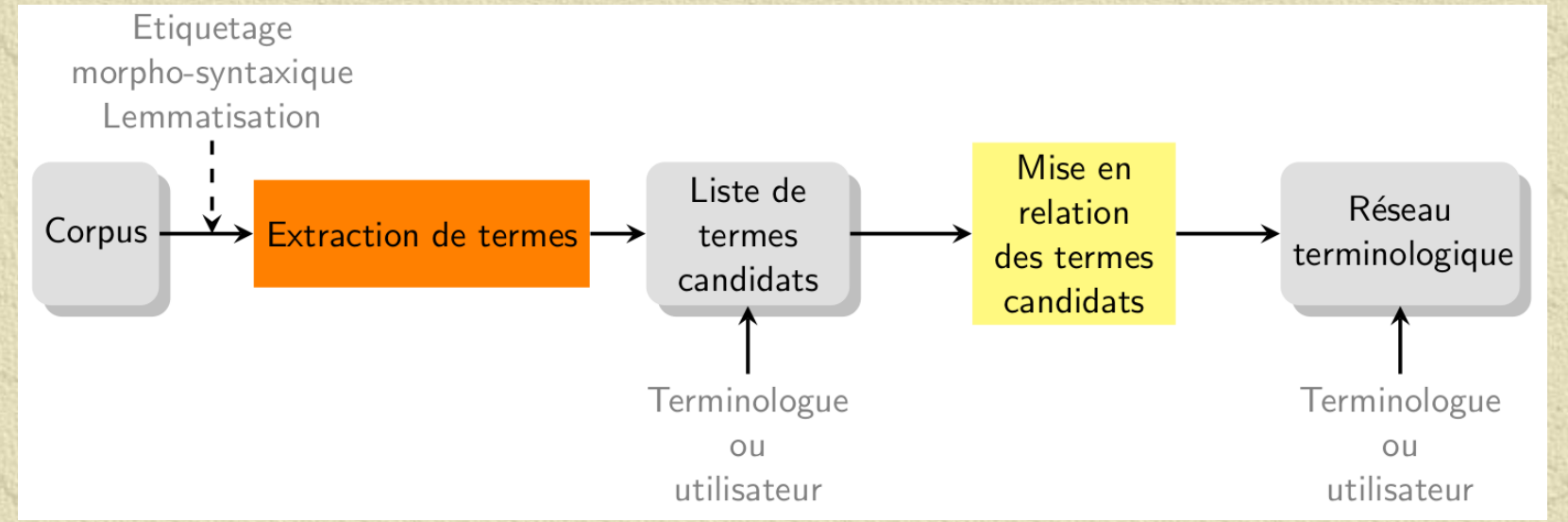
Extraction des connaissances à partir des corpus

Démarche


Etiquetage
morpho-syntaxique
Lemmatisation
Corpus
Extraction de termes
Liste de termes candidats
Mise en relation des termes candidats
Réseau terminologique
Terminologue ou utilisateur
Terminologue ou utilisateur

Concepts

Terminologie :

Est l'ensemble des termes qui identifient un sujet spécifique.

Le Terme:

Une entité syntaxique composée d'un mot ou d'un groupe de mots considéré comme une notion importante dans le domaine étudie et qui peut être associe à un concept dans une base de connaissance

Concepts

L'extraction terminologique:

L'extraction terminologique est une application du Traitement Automatique du Langage Naturel qui consiste à extraire automatiquement une liste de termes à partir d'un corpus spécialisé. Les logiciels réalisant l'extraction terminologique sont appelés extracteurs de termes.

Les outils d'extraction des terminologies :

On peut envisager d'utiliser deux catégories d'outils :

- outils d'extraction de terminologie exploitant les caractéristiques linguistiques du terme.
- Et ceux basés sur une étude statistique du corpus, exploitant principalement le phénomène d'occurrence.

Les outils d'extraction des terminologies :

- **Outils à vocation terminologique**

1. • outils mettent en œuvre une méthode descendante
2. • outils mettent en œuvre une méthode ascendant

Les outils d'extraction des terminologies

- **Méthodes Ascendantes :**

Corpus (Texte) ---Extraction-----But.

- **Méthodes Descendantes :**

But ---Projection--- Corpus (texte)

Méthodes ascendantes

- L'approche ascendante consiste à recueillir le maximum de données verbales auprès d'un expert et les regrouper pour former un modèle.

- La méthode KOD a été utilisée.

Méthodes Descendantes

- L'approche descendante se focalise rapidement sur la définition du modèle d'expertise afin de filtrer les connaissances acquises et de guider efficacement le processus d'acquisition de ces connaissances.

- La méthode KADS a été utilisée.

Avantage et inconvénient

- Avec des outils de type descendant, on doit gérer à la fois du bruit (résultats erronés) et du silence (données pertinentes oubliées).

Avantage et inconvénient

Du silence, parce que certains éléments ne sont pas repérés car ils ne sont pas connus a priori alors qu'ils seraient pertinents.

Du bruit, à cause de l'absence de contrainte sur les contextes dans lesquels sont pris les éléments considérés comme pertinents.

Avantage et inconvénient

- Avec des outils de type ascendant, on doit gérer essentiellement du bruit.

Modèle Mixte

- Étapes du processus d'acquisition des connaissances:
  1. Recueillir les connaissances de base pour la construction du modèle d'expertise (ascendante)
  2. Choix d'un modèle générique (descendante)
  3. Instanciation du modèle conceptuel (descendante)
  4. Opérationnalisation du modèle conceptuel (descendante).

Les outils d'extraction des terminologies :

- **Outils d'analyse de corpus**
  - En effet, ces outils permettent de constituer des dictionnaires de fréquences, des index sélectifs ou systématiques.
  - Mais si ces outils peuvent jouer un rôle de soutien, d'assistance au repérage de termes, c'est surtout parce qu'ils facilitent la recherche de concordances (liste des mots avec leurs contextes), le repérage de parties de mots et de collocations, c'est-à-dire des cooccurrences de termes.
  - Nombreux sont les outils de ce type (Sato, Tact, Hyperbase, etc

Quelques systèmes d'extraction de la terminologie

Systèmes	*Linguistiques*	*Statistiques*	*Références*
TERMINO	**X**		[David et Plante 1990]
LEXTER	**X**		[Bourigault 1993]
FASTR	**X**		[Jacquemin 1996]
INTEX	**X**		[Silberztein1994 ; Ibekwe-SanJuan 2001]
ANA		**X**	[Enguehard 1993]
MANTEX		**X**	[Frath *et al.* 2000]
XTRACT	**X**	**X**	[Smadja 1993]
ACABIT	**X**	**X**	[Daille 1994]
CLARIT	**X**	**X**	[Evans et Zhai 1996]
TERMIGHT	**X**	**X**	[Daga et Church 1997]
SYNTEX	**X**	**X**	[Bourigault et Fabre 2000]
C/NC VALUE	**X**	**X**	[Frantzi *et al.* 2000]
WASPBENCH	**X**	**X**	[Kilgariff et Tugwel 2001]
FIPS	**X**	**X**	[Nerima *et al.* 2003]
ESATEC	**X**	**X**	[Biskri *et al.* 2004]
EXIT	**x**	**x**	[Roche *et al.* 2004]

Lexter

Didier Bourigault (1993), Analyse syntaxique locale pour le repérage de termes complexes dans un texte.

Analyse endogène (pas de connaissance du domaine)

L'acquisition des termes est effectuée en trois étapes :

- Extraction des groupes nominaux maximaux
- Décomposition des groupes nominaux

Lexter

- Méthode linguistique
- Trois étapes :
  - Extraction des groupes nominaux maximaux
  - Décomposition des groupes nominaux maximaux
  - Présentation des résultats sous forme d'un réseau sémantique

Quelques systèmes d'extraction de la terminologie

- **Extraction des groupes nominaux maximaux:**

  - Repérage de frontières syntaxiques : verbes conjugués, pronoms, conjonctions, préposition + adjectif possessif, etc.

  - Extraction de syntagmes nominaux maximaux.

Lexter

- **Extraction des groupes nominaux maximaux « Exemple »**
  - En entrée : Texte initial (étiqueté)

<Prep>En <NomFS>présence <Prep>de <NomFS>sténose <Adj?S>sévère <Prep>du <NomMS>tronc <Adj?S>commun <Prep>de <Det?S>l' <NomFS>artère <Adj?S>coronaire <Adj?S>gauche <Typo>, <Det?S>on <Pro>se <VCONJ>contente <Prep>d' <Det>un <Nom?S>minimum <Prep>d' <NomFP>injections <Typo>,

  - Groupes nominaux maximaux

sténose sévère du tronc commun de l'artère coronaire gauche

minimum d'injections

Lexter

- **Décomposition des groupes nominaux maximaux**

  - **Hypothèse :** tout terme complexe est composé d'une tête et d'une expansion.
  - Deux règles classiques de décomposition

    - nom1 adjectif :

      Tête : nom1

      Expansion : adjectif

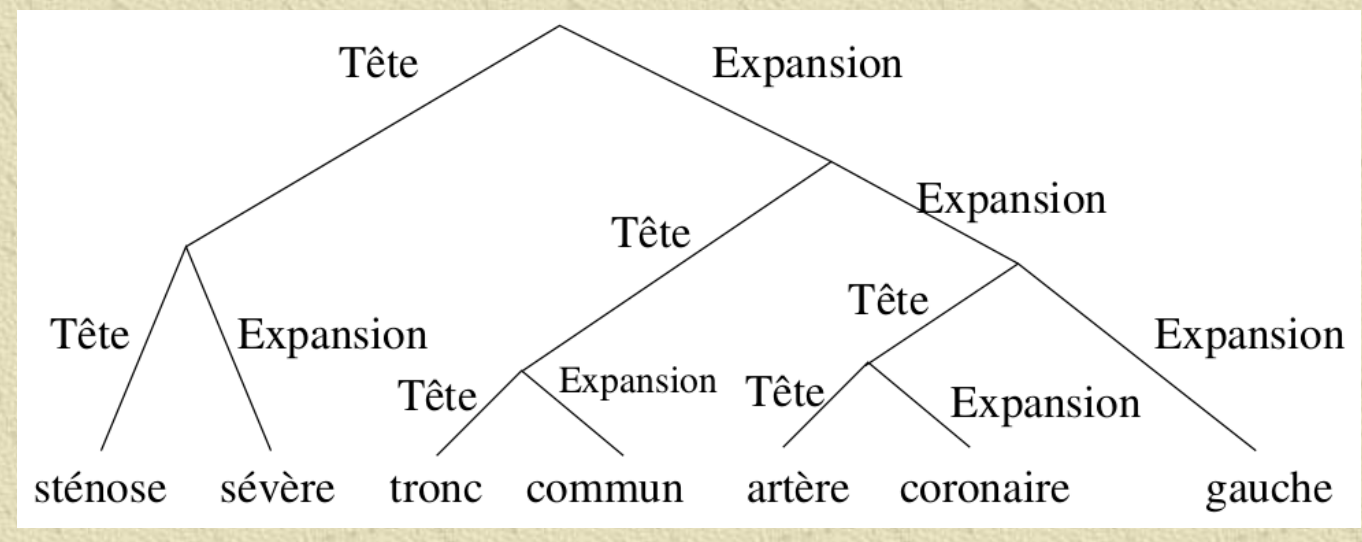
    - nom1 de nom2 :

      Tête : nom1

      Expansion : nom2 (de)

Lexter

sténose sévère du tronc commun de l'artère coronaire gauche



Lexter

- **Décomposition des groupes nominaux maximaux**
  - **Problème d'ambiguïtés**
  - **Problème de rattachement s'il y a absence d'informations sur le genre ou le nombre**

Par exemples, les groupes nominaux de type Nom1 de Nom2 Adjectif (corps français) : ***centre de tourisme équestre***

Deux types de décompositions pour «centre de tourisme équestre»

Tête : centre
Expansion: tourisme équestre

Tête : centre de tourisme
Expansion : équestre

Acabit

Béatrice Daille (1995), Repérage et extraction de terminologie par une approche mixte statistique et linguistique.

- Approche mixte linguistique et statistique
- Bitermes et leurs variantes
- Extraction de candidats termes à partir d'un corpus préalablement étiqueté et désambiguïsé

Acabit

L'acquisition terminologique dans Acabit se déroule en deux étapes :

1) Analyse linguistique et regroupement de variantes :

Corpus étiqueté

- Transducteurs pour la recherche de séquences nominales
- Extraction de candidats termes :
  - N Adj : station terrienne
  - N 1 prep N 2 : liaison par satellite
  - N 1 N 2 : diode tunnel
- Décomposition en candidats termes binaires :

réseau de transit à satellite

→ réseau de transit

→ réseau à satellite

Acabit

L'acquisition terminologique dans Acabit se déroule en deux étapes :

2. Filtrage statistique :
    - Mesures statistiques pour le tri de candidats termes binaires
    - Calcul de scores et de distances sur les composants des candidats termes basés sur les fréquences
    - log-likehood ratio (Dunning, 1993)

le mieux pour retenir les termes candidats sans être sensible aux fréquences.