Influence des domaines de spécialité dans l'extraction de termes-clés

Adrien Bougouin Florian Boudin Béatrice Daille

Université de Nantes, LINA

2 Juillet 2014

Introduction

Contexte

Termes-clés (mots-clés)

- Mots ou expressions polylexicales
- Aperçu d'un document
- Donnés par les auteurs, des lecteurs ou des documentalistes

Applications

- Indexation de documents
- Résumé automatique
- Classification de documents
- Expansion de requêtes
- Ciblage (*marketing*)
- etc.

Introduction

Contexte

Termes-clés (mots-clés)

- Mots ou expressions polylexicales
- Aperçu d'un document
- Donnés par les auteurs, des lecteurs ou des documentalistes

Applications

- Indexation de documents
- Résumé automatique
- Classification de documents
- Expansion de requêtes
- Ciblage (*marketing*)
- etc.

Introduction

Contexte

Termes-clés (mots-clés)

- Mots ou expressions polylexicales
- Aperçu d'un document
- Donnés par les auteurs, des lecteurs ou des **documentalistes**

Applications

- **Indexation de documents**
- Résumé automatique
- Classification de documents
- Expansion de requêtes
- Ciblage (*marketing*)
- etc.

Introduction

Contexte (suite)

Indexation de documents scientifiques

Création de notices bibliographiques :

- Titre
- Auteurs
- Résumé
- Codes de classement
- **Descripteurs conceptuels $\Leftrightarrow$ termes-clés**

Introduction

Exemple

Variabilité du **Gravettien** de Kostienki (bassin moyen du Don) et des territoires associés

Dans la région de Kostienki-**Borschevo**, on observe l'expression, à ce jour, la plus orientale du modèle européen de l'évolution du **Paléolithique supérieur**. Elle est différente à la fois du modèle Sibérien et du modèle de l'Asie centrale. Comme ailleurs en **Europe**, le **Gravettien** apparaît à Kostienki vers 28 ka (Kostienki 8 /II/). Par la suite, entre 24-20 ka, les techno-complexes **gravettiens** sont représentés au moins par quatre faciès dont deux, ceux de Kostienki 21/III/ et Kostienki 4 /II/, ressemblent au **Gravettien** occidental et deux autres, Kostienki-**Avdeevo** et Kostienki 11/II/, sont des faciès propres à l'**Europe** de l'Est, sans analogie à l'Ouest.

Descripteurs (termes-clés) : **Europe**, Kostienko, **Borschevo**, variation, typologie, industrie osseuse, industrie lithique, Europe centrale, **Avdeevo**, **Paléolithique supérieur**, **Gravettien**.

Archéologie

Introduction

Problèmatique

Mais. . .

L'assignation de termes-clés est une tâche coûteuse.
$\Rightarrow$ Il faut extraire les termes-clés automatiquement.

Extraction automatique de termes-clés

- Supervisée/**non-supervisée**
- Extraction des termes-clés **contenus dans** le titre/résumé
- Terme-clé = unité textuelle **importante** dans le titre/résumé
  - Fréquente et spécifique (TF-IDF)
  - Centrale (Mihalcea et Tarau, 2004, TextRank)

Introduction

Problèmatique

Mais. . .

L'assignation de termes-clés est une tâche coûteuse.
$\Rightarrow$ Il faut extraire les termes-clés automatiquement.

Extraction automatique de termes-clés

- Supervisée/**non-supervisée**
- Extraction des termes-clés **contenus dans** le titre/résumé
- Terme-clé = unité textuelle **importante** dans le titre/résumé
  - Fréquente et spécifique (TF-IDF)
  - Centrale (Mihalcea et Tarau, 2004, TextRank)

Introduction

Hypothèse

Il est plus difficile d'extraire les termes-clés pour certaines disciplines que pour d'autres.

$\Rightarrow$ Quels sont les facteurs qui influent sur cette difficulté ?

Plan

1. Données
2. Extraction de termes-clés
3. Expérience
4. Conclusion et perspectives

Plan

1 Données

2 Extraction de termes-clés

3 Expérience

4 Conclusion et perspectives

Données

- Cinq disciplines :
  - Archéologie
  - Linguistique
  - Sciences de l'information
  - Psychologie
  - Chimie
- Entre 700 et 800 notices par discipline :
  - Titre
  - Résumé
  - Descripteurs (termes-clés de références)
    - Contrôlés : appartenant au vocabulaire de la discipline
    - Non-contrôlés : choisis librement

Données

- Cinq disciplines :
  - Archéologie
  - Linguistique
  - Sciences de l'information
  - Psychologie
  - Chimie
- Entre 700 et 800 notices par discipline :
  - Titre
  - Résumé
  - Descripteurs (termes-clés de références)
    - Contrôlés : appartenant au vocabulaire de la discipline
    - Non-contrôlés : choisis librement

Données (suite)

	Archéologie	Linguistique	Sciences de l'information	Psychologie	Chimie
Notices	718	715	706	720	782
Mots/doc.	219,1	156,7	119,7	185,7	105,2
Termes-clés/doc.	16,6	8,0	8,5	11,6	12,8
Mots/terme-clé	1,3	1,8	1,7	1,6	2,2
Diversité des termes-clés	25,5 %	23,0 %	25,0 %	17,4 %	40,6 %
Termes-clés contrôlés	79,8 %	86,9%	85,8 %	90,9%	83,0 %
Termes-clés non contrôlés	20,2 %	13,1%	14,2 %	9,1%	17,0 %
Termes-clés extractibles (Rappel max.)	62,9 %	38,8 %	32,4 %	27,1 %	23,7 %

- Peu de contenu
- Différence d'organisation du discours
- Différence de complexité des termes-clés
- Diversité plus importante en chimie
- Difficulté a priori très importante

Données (suite)

	Archéologie	Linguistique	Sciences de l'information	Psychologie	Chimie
Notices	718	715	706	720	782
Mots/doc.	219,1	156,7	119,7	185,7	105,2
Termes-clés/doc.	16,6	8,0	8,5	11,6	12,8
Mots/terme-clé	1,3	1,8	1,7	1,6	2,2
Diversité des termes-clés	25,5 %	23,0 %	25,0 %	17,4 %	40,6 %
Termes-clés contrôlés	79,8 %	86,9%	85,8 %	90,9%	83,0 %
Termes-clés non contrôlés	20,2 %	13,1%	14,2 %	9,1%	17,0 %
Termes-clés extractibles (Rappel max.)	62,9 %	38,8 %	32,4 %	27,1 %	23,7 %

- Peu de contenu
- Différence d'organisation du discours
- Différence de complexité des termes-clés
- Diversité plus importante en chimie
- Difficulté a priori très importante

Données (suite)

Variabilité du Gravettien de Kostienki (bassin moyen du Don) et des territoires associés

Dans la région de Kostienki-Borschevo, on observe l'expression, à ce jour, la plus orientale du modèle européen de l'évolution du Paléolithique supérieur. Elle est différente à la fois du modèle Sibérien et du modèle de l'Asie centrale. Comme ailleurs en Europe, le Gravettien apparaît à Kostienki vers 28 ka (Kostienki 8 /II/). Par la suite, entre 24-20 ka, les techno-complexes gravettiens sont représentés au moins par quatre faciès dont deux, ceux de Kostienki 21/III/ et Kostienki 4 /II/, ressemblent au Gravettien occidental et deux autres, Kostienki-Avdeevo et Kostienki 11/II/, sont des faciès propres à l'Europe de l'Est, sans analogie à l'Ouest.

Descripteurs (termes-clés) : Europe, Kostienko, Borschevo, variation, typologie, industrie osseuse, industrie lithique, Europe centrale, Avdeevo, Paléolithique supérieur, Gravettien.

Archéologie

Etude d'un condensat acide isocyanurique-urée-formaldéhyde

La synthèse d'un condensat acide isocyanurique-urée-formaldéhyde utilisant la pyridine en tant que solvant a été effectuée par réaction sonochimique.

Descripteurs (termes-clés) : Réaction sonochimique, hétérocycle azote, cycle 6 chaînons, ether.

Chimie

Données (suite)

	Archéologie	Linguistique	Sciences de l'information	Psychologie	Chimie
Notices	718	715	706	720	782
Mots/doc.	219,1	156,7	119,7	185,7	105,2
Termes-clés/doc.	16,6	8,0	8,5	11,6	12,8
Mots/terme-clé	1,3	1,8	1,7	1,6	2,2
Diversité des termes-clés	25,5 %	23,0 %	25,0 %	17,4 %	40,6 %
Termes-clés contrôlés	79,8 %	86,9%	85,8 %	90,9%	83,0 %
Termes-clés non contrôlés	20,2 %	13,1%	14,2 %	9,1%	17,0 %
Termes-clés extractibles (Rappel max.)	62,9 %	38,8 %	32,4 %	27,1 %	23,7 %

- Peu de contenu
- Différence d'organisation du discours
- Différence de complexité des termes-clés
- Diversité plus importante en chimie
- Difficulté a priori très importante

Données (suite)

	Archéologie	Linguistique	Sciences de l'information	Psychologie	Chimie
Notices	718	715	706	720	782
Mots/doc.	219,1	156,7	119,7	185,7	105,2
Termes-clés/doc.	16,6	8,0	8,5	11,6	12,8
Mots/terme-clé	1,3	1,8	1,7	1,6	2,2
Diversité des termes-clés	25,5 %	23,0 %	25,0 %	17,4 %	40,6 %
Termes-clés contrôlés	79,8 %	86,9%	85,8 %	90,9%	83,0 %
Termes-clés non contrôlés	20,2 %	13,1%	14,2 %	9,1%	17,0 %
Termes-clés extractibles (Rappel max.)	62,9 %	38,8 %	32,4 %	27,1 %	23,7 %

- Peu de contenu
- Différence d'organisation du discours
- Différence de complexité des termes-clés
- Diversité plus importante en chimie
- Difficulté a priori très importante

Données (suite)

	Archéologie	Linguistique	Sciences de l'information	Psychologie	Chimie
Notices	718	715	706	720	782
Mots/doc.	219,1	156,7	119,7	185,7	105,2
Termes-clés/doc.	16,6	8,0	8,5	11,6	12,8
Mots/terme-clé	1,3	1,8	1,7	1,6	2,2
Diversité des termes-clés	25,5 %	23,0 %	25,0 %	17,4 %	40,6 %
Termes-clés contrôlés	79,8 %	86,9%	85,8 %	90,9%	83,0 %
Termes-clés non contrôlés	20,2 %	13,1%	14,2 %	9,1%	17,0 %
Termes-clés extractibles (Rappel max.)	62,9 %	38,8 %	32,4 %	27,1 %	23,7 %

- Peu de contenu
- Différence d'organisation du discours
- Différence de complexité des termes-clés
- Diversité plus importante en chimie
- Difficulté a priori très importante

Données (suite)

Variabilité du Gravettien de Kostienki (bassin moyen du Don) et des territoires associés

Dans la région de Kostienki-Borschevo, on observe l'expression, à ce jour, la plus orientale du modèle européen de l'évolution du Paléolithique supérieur. Elle est différente à la fois du modèle Sibérien et du modèle de l'Asie centrale. Comme ailleurs en Europe, le Gravettien apparaît à Kostienki vers 28 ka (Kostienki 8 /II/). Par la suite, entre 24-20 ka, les techno-complexes gravettiens sont représentés au moins par quatre faciès dont deux, ceux de Kostienki 21/III/ et Kostienki 4 /II/, ressemblent au Gravettien occidental et deux autres, Kostienki-Avdeevo et Kostienki 11/II/, sont des faciès propres à l'Europe de l'Est, sans analogie à l'Ouest.

Descripteurs (termes-clés) : Europe, Kostienko, Borschevo, variation, typologie, industrie osseuse, industrie lithique, Europe centrale, Avdeevo, Paléolithique supérieur, Gravettien.

Archéologie

Etude d'un condensat acide isocyanurique-urée-formaldéhyde

La synthèse d'un condensat acide isocyanurique-urée-formaldéhyde utilisant la pyridine en tant que solvant a été effectuée par réaction sonochimique.

Descripteurs (termes-clés) : Réaction sonochimique, hétérocycle azote, cycle 6 chaînons, ether.

Chimie

Données (suite)

	Archéologie	Linguistique	Sciences de l'information	Psychologie	Chimie
Notices	718	715	706	720	782
Mots/doc.	219,1	156,7	119,7	185,7	105,2
Termes-clés/doc.	16,6	8,0	8,5	11,6	12,8
Mots/terme-clé	1,3	1,8	1,7	1,6	2,2
Diversité des termes-clés	25,5 %	23,0 %	25,0 %	17,4 %	40,6 %
Termes-clés contrôlés	79,8 %	86,9%	85,8 %	90,9%	83,0 %
Termes-clés non contrôlés	20,2 %	13,1%	14,2 %	9,1%	17,0 %
Termes-clés extractibles (Rappel max.)	62,9 %	38,8 %	32,4 %	27,1 %	23,7 %

- Peu de contenu
- Différence d'organisation du discours
- Différence de complexité des termes-clés
- Diversité plus importante en chimie
- Difficulté a priori très importante

Données (suite)

	Archéologie	Linguistique	Sciences de l'information	Psychologie	Chimie
Notices	718	715	706	720	782
Mots/doc.	219,1	156,7	119,7	185,7	105,2
Termes-clés/doc.	16,6	8,0	8,5	11,6	12,8
Mots/terme-clé	1,3	1,8	1,7	1,6	2,2
Diversité des termes-clés	25,5 %	23,0 %	25,0 %	17,4 %	40,6 %
Termes-clés contrôlés	79,8 %	86,9%	85,8 %	90,9%	83,0 %
Termes-clés non contrôlés	20,2 %	13,1%	14,2 %	9,1%	17,0 %
Termes-clés extractibles (Rappel max.)	62,9 %	38,8 %	32,4 %	27,1 %	23,7 %

- Peu de contenu
- Différence d'organisation du discours
- Différence de complexité des termes-clés
- Diversité plus importante en chimie
- Difficulté a priori très importante

Données (suite)

	Archéologie	Linguistique	Sciences de l'information	Psychologie	Chimie
Notices	718	715	706	720	782
Mots/doc.	219,1	156,7	119,7	185,7	105,2
Termes-clés/doc.	16,6	8,0	8,5	11,6	12,8
Mots/terme-clé	1,3	1,8	1,7	1,6	2,2
Diversité des termes-clés	25,5 %	23,0 %	25,0 %	17,4 %	40,6 %
Termes-clés contrôlés	79,8 %	86,9%	85,8 %	90,9%	83,0 %
Termes-clés non contrôlés	20,2 %	13,1%	14,2 %	9,1%	17,0 %
Termes-clés extractibles (Rappel max.)	62,9 %	38,8 %	32,4 %	27,1 %	23,7 %

- Peu de contenu
- Différence d'organisation du discours
- Différence de complexité des termes-clés
- Diversité plus importante en chimie
- Difficulté a priori très importante

Données (suite)

	Archéologie	Linguistique	Sciences de l'information	Psychologie	Chimie
Notices	718	715	706	720	782
Mots/doc.	219,1	156,7	119,7	185,7	105,2
Termes-clés/doc.	16,6	8,0	8,5	11,6	12,8
Mots/terme-clé	1,3	1,8	1,7	1,6	2,2
Diversité des termes-clés	25,5 %	23,0 %	25,0 %	17,4 %	40,6 %
Termes-clés contrôlés	79,8 %	86,9%	85,8 %	90,9%	83,0 %
Termes-clés non contrôlés	20,2 %	13,1%	14,2 %	9,1%	17,0 %
Termes-clés extractibles (Rappel max.)	62,9 %	38,8 %	32,4 %	27,1 %	23,7 %

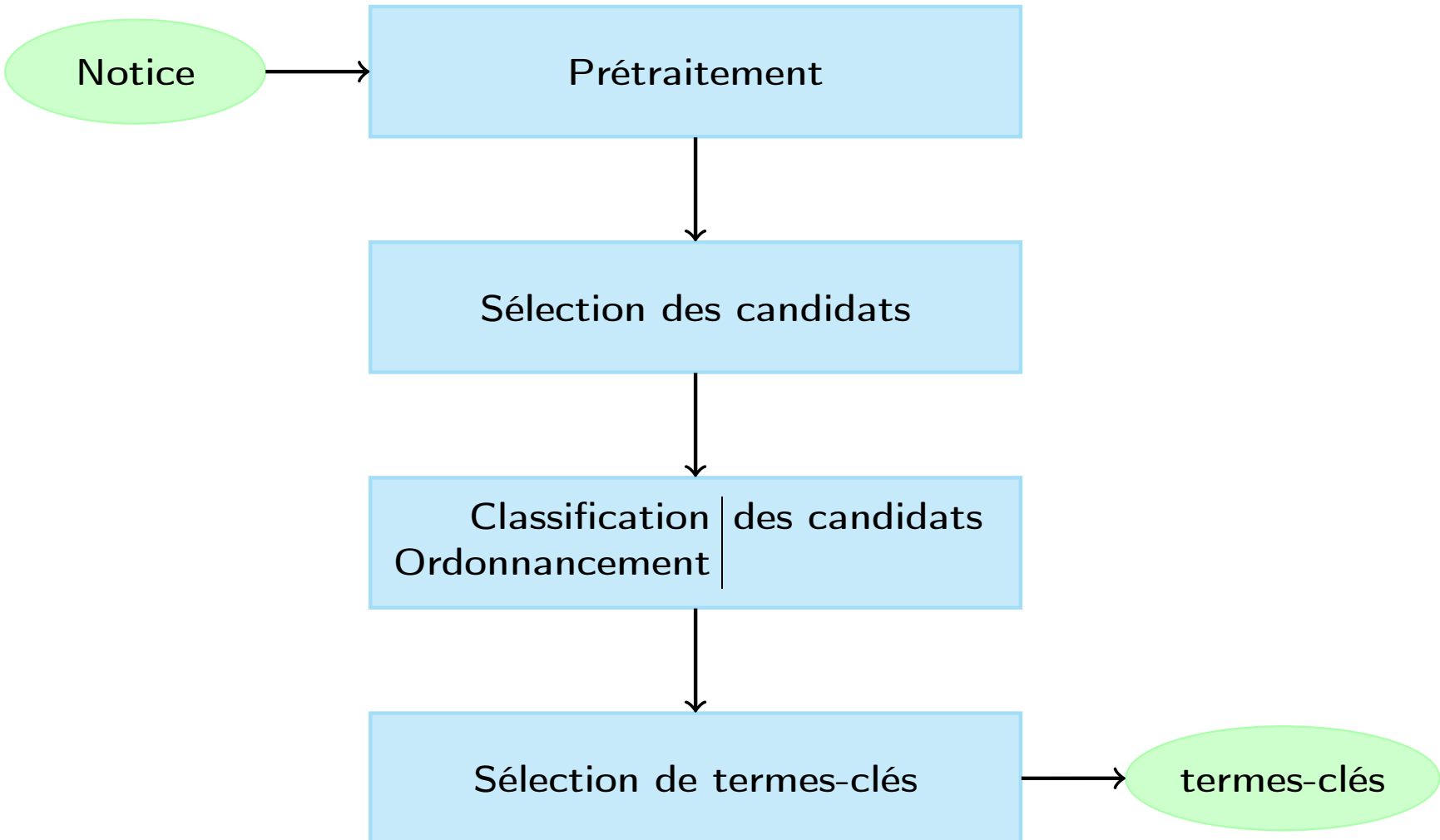
- Peu de contenu
- Différence d'organisation du discours
- Différence de complexité des termes-clés
- Diversité plus importante en chimie
- Difficulté a priori très importante

Plan

1. Données
2. Extraction de termes-clés
3. Expérience
4. Conclusion et perspectives

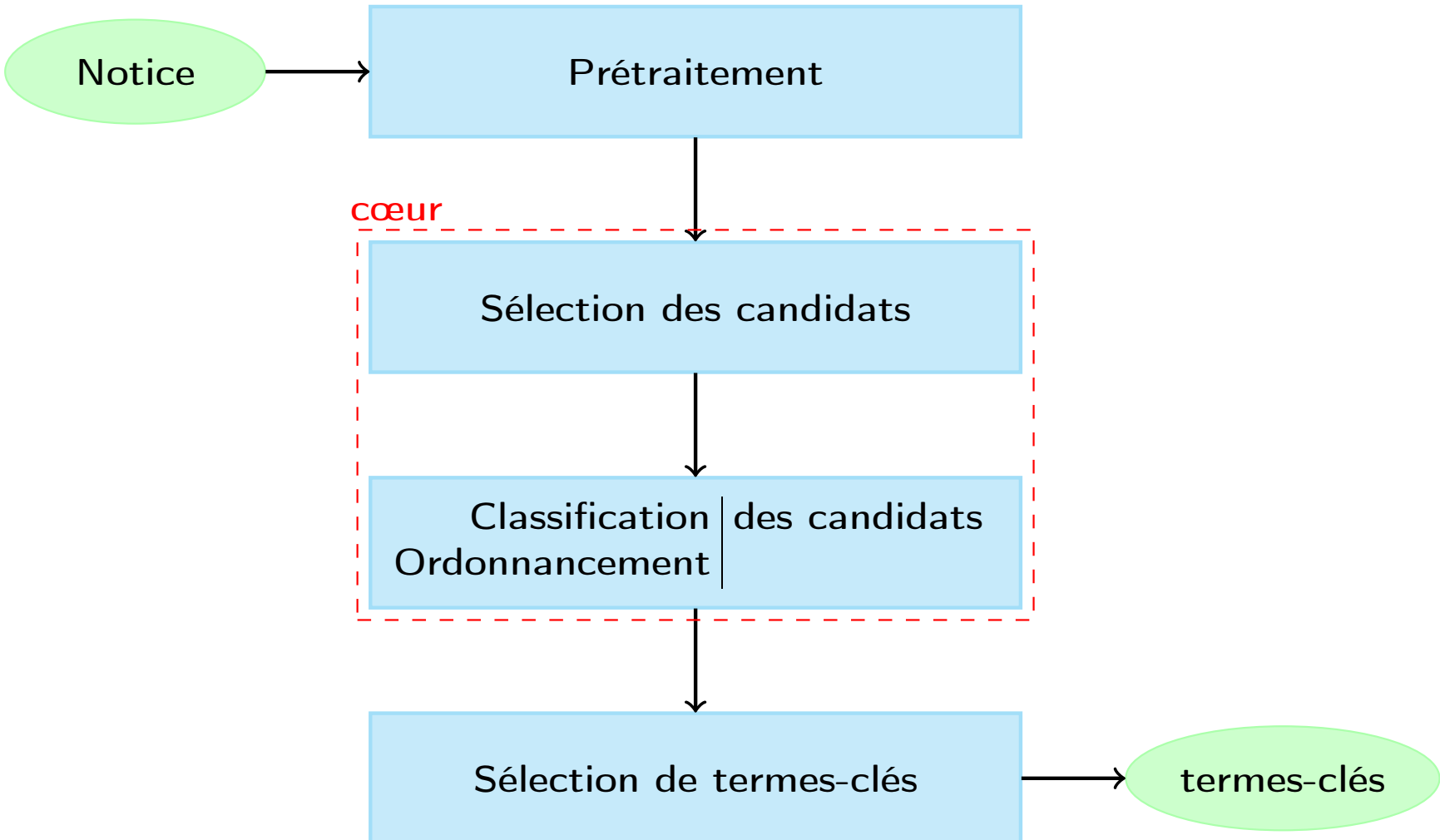
Extraction de termes-clés

Chaîne de traitements



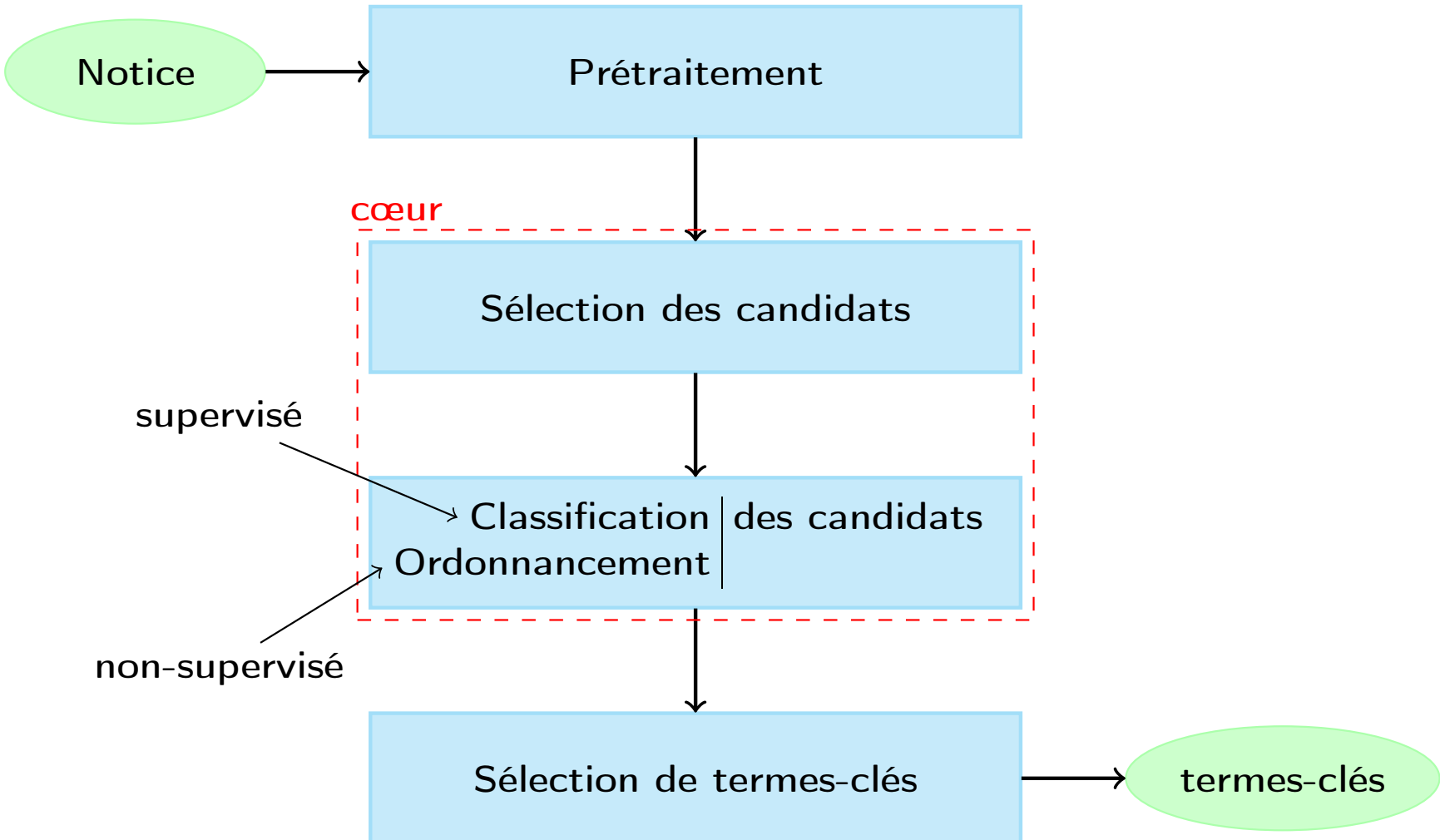
Extraction de termes-clés

Chaîne de traitements



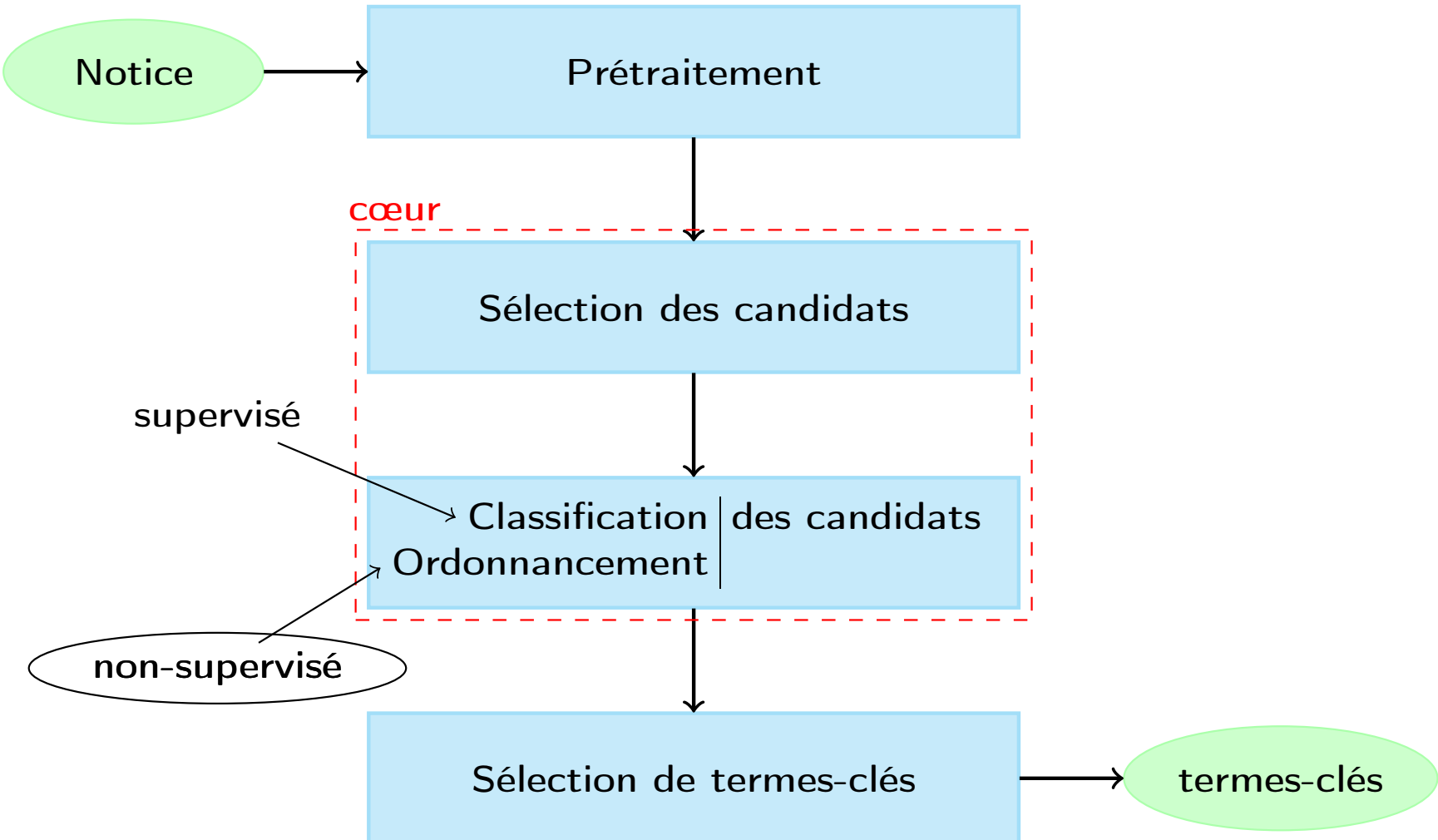
Extraction de termes-clés

Chaîne de traitements

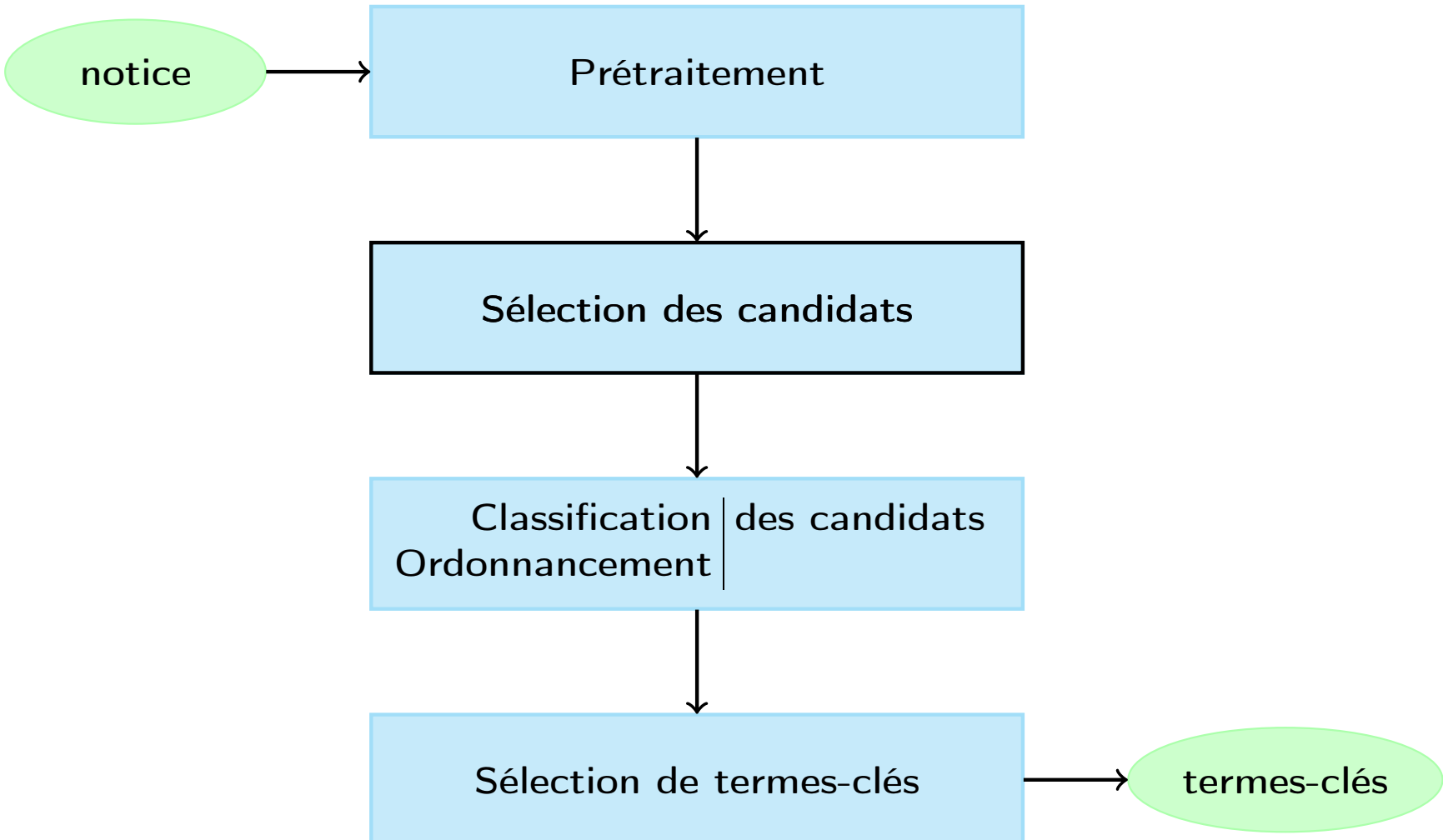


Extraction de termes-clés

Chaîne de traitements



Extraction de termes-clés



Extraction de termes-clés

Sélection des candidats

Deux approches classiques :

- Extraction des n-grammes
  - $n \subseteq \{1..3\}$
  - Filtrage avec un anti-dictionnaire
  - Sursélection des candidats $\Rightarrow$ faible qualité
- Reconnaissance de formes
  - `(NOM | ADJ)+`

Une approche non explorée jusqu'alors :

- Extraction des candidats termes
  - Utilisation de TermSuite
  - Formes très précises :
    - `NOM à NOM`
    - `NOM en NOM`
    - `NOM à NOM ADJ`
    - etc.

Extraction de termes-clés

Sélection des candidats

Deux approches classiques :

- Extraction des n-grammes
  - $n \subseteq \{1..3\}$
  - Filtrage avec un anti-dictionnaire
  - Sursélection des candidats $\Rightarrow$ faible qualité
- Reconnaissance de formes
  - `(NOM | ADJ)+`

Une approche non explorée jusqu'alors :

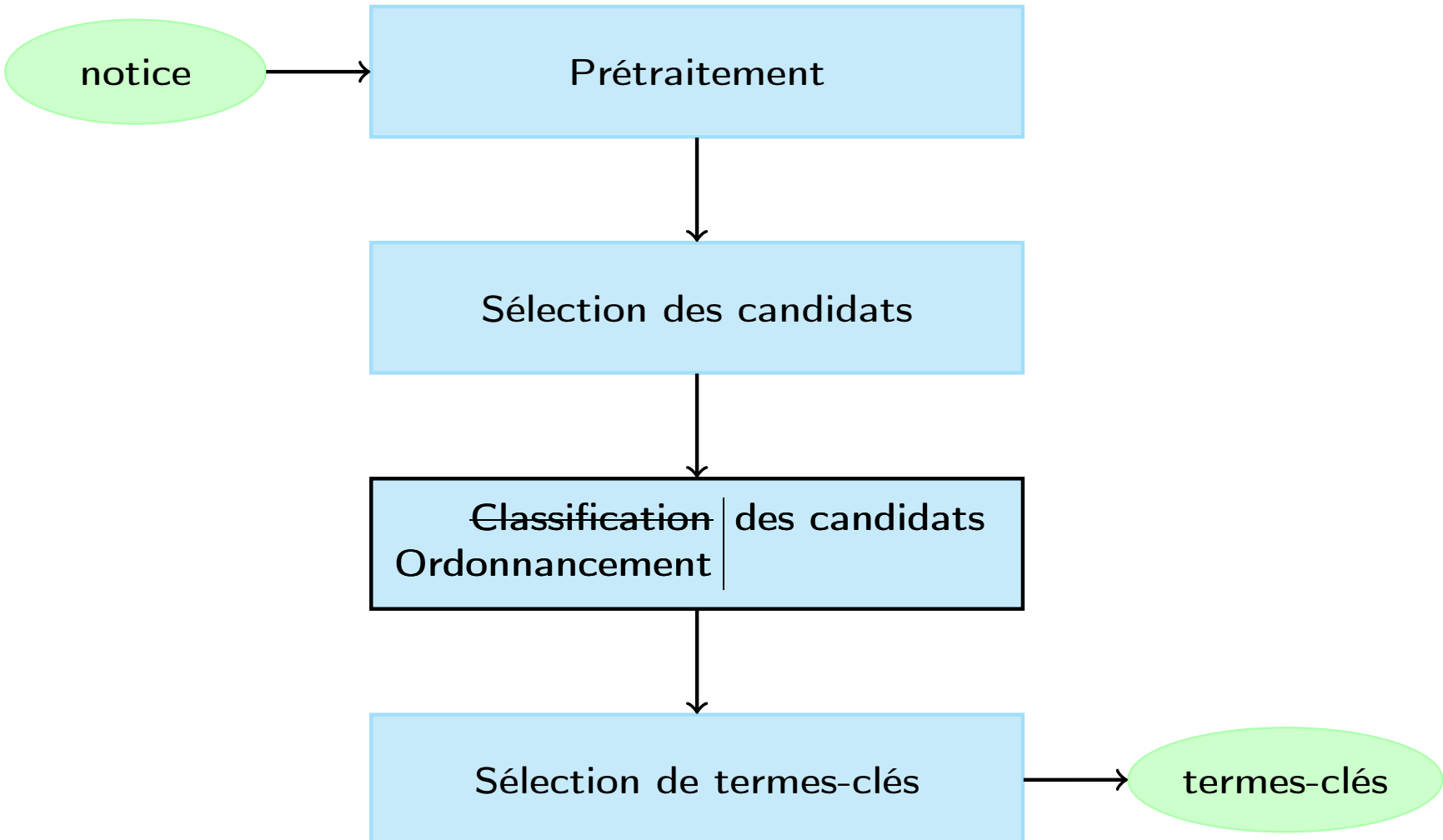
- Extraction des candidats termes
  - Utilisation de TermSuite
  - Formes très précises :
    - `NOM à NOM`
    - `NOM en NOM`
    - `NOM à NOM ADJ`
    - etc.

Extraction de termes-clés

Sélection des candidats — Exemples

« *bassin moyen du Don* »		
{1..3}-grammes	`(NOM | ADJ)+`	Candidats termes
« *bassin* »	« *bassin moyen* »	« *bassin moyen du Don* »
« *moyen* »	« *Don* »	↪ « *bassin* »
« *Don* »		↪ « *moyen* »
« *bassin moyen* »		↪ « *Don* »
« *moyen du Don* »		↪ « *bassin moyen* »

Extraction de termes-clés



Extraction de termes-clés

TF×IDF

Hypothèse

Dans une notice, un mot est d'autant plus important qu'il y est fréquent (TF) et spécifique (IDF).

$$\text{importance}(\text{candidat}) = \sum_{\text{mot} \in \text{candidat}} \text{TF} \times \text{IDF}(\text{mot})$$

Extraction de termes-clés

TopicRank

Hypothèses

1. Plusieurs candidats désignent le même sujet (concept)
2. Seul le candidat le plus représentatif du sujet doit être extrait
3. Les sujets qui cooccurrent se recommandent mutuellement :
   - Plus un sujet cooccurre avec d'autres sujets, plus il est important
   - Plus un sujet est important, plus les sujets avec lesquels il cooccurre sont important

Plan

1 Données

2 Extraction de termes-clés

3 Expérience

4 Conclusion et perspectives

Expérience

Configuration

- Cinq disciplines :
  - Archéologie
  - Linguistique
  - Sciences de l'information
  - Psychologie
  - Chimie
- Six systèmes d'extraction de termes-clés
  - {1..3}-grammes $\longrightarrow$ TF×IDF
  - (NOM | ADJ)+ $\longrightarrow$ TF×IDF
  - Candidats termes $\longrightarrow$ TF×IDF
  - {1..3}-grammes $\longrightarrow$ TopicRank
  - (NOM | ADJ)+ $\longrightarrow$ TopicRank
  - Candidats termes $\longrightarrow$ TopicRank

Expérience

Configuration

- Cinq disciplines :
  - Archéologie
  - Linguistique
  - Sciences de l'information
  - Psychologie
  - Chimie
- Six systèmes d'extraction de termes-clés
  - $\{1..3\}$-grammes $\longrightarrow$ TF$\times$IDF
  - `(NOM | ADJ)+` $\longrightarrow$ TF$\times$IDF
  - Candidats termes $\longrightarrow$ TF$\times$IDF
  - $\{1..3\}$-grammes $\longrightarrow$ TopicRank
  - `(NOM | ADJ)+` $\longrightarrow$ TopicRank
  - Candidats termes $\longrightarrow$ TopicRank

Expérience

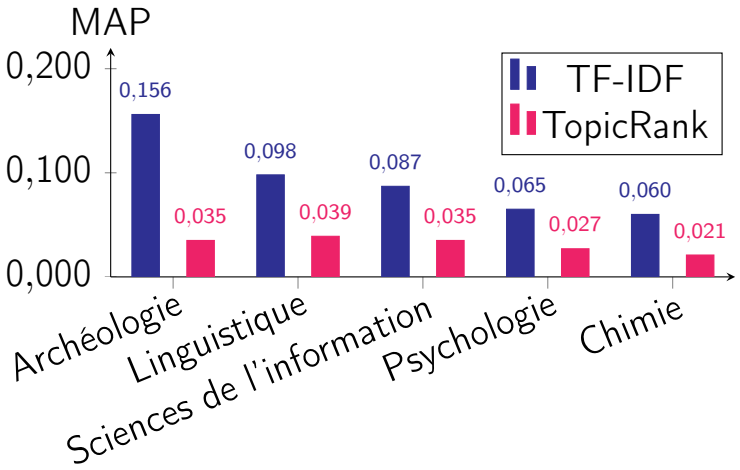
Mesure d'évaluation

- Évaluation de l'ordonnancement des candidats
- MAP (*Mean Average Precision*) :

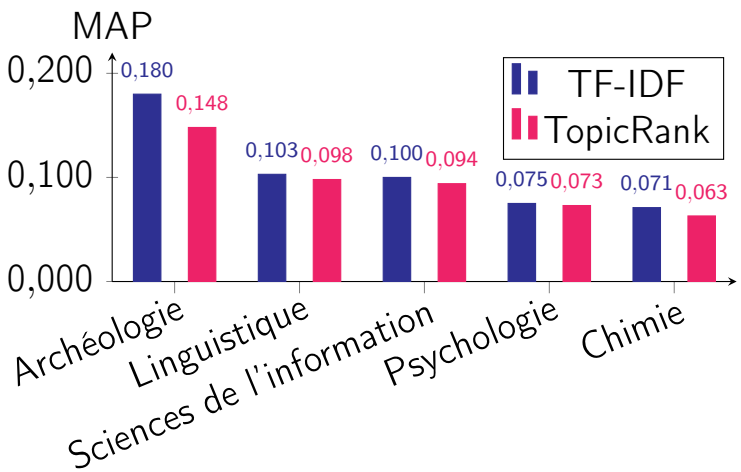
$$\text{MAP} = \frac{1}{\|\text{DOCUMENTS}\|} \sum_{d \in \text{DOCUMENTS}} \frac{\sum_{t \in \text{CORRECTS}_d} \text{précision@rang}_d(t)}{\|\text{REFERENCE}_d\|}$$

Expérience

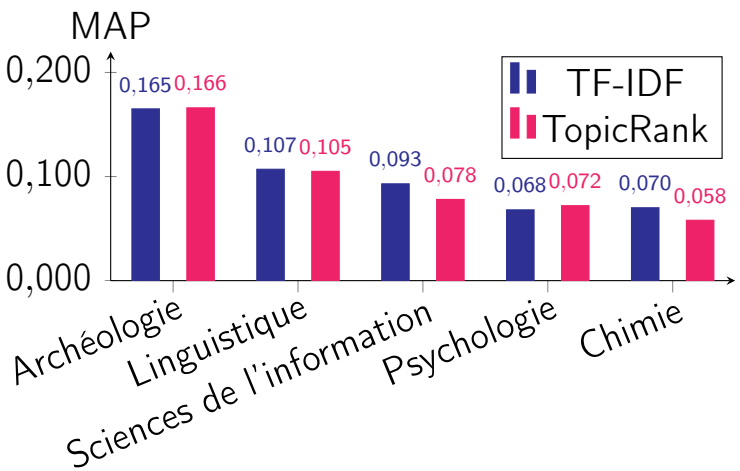
Résultats et observations



{1..3}-grammes



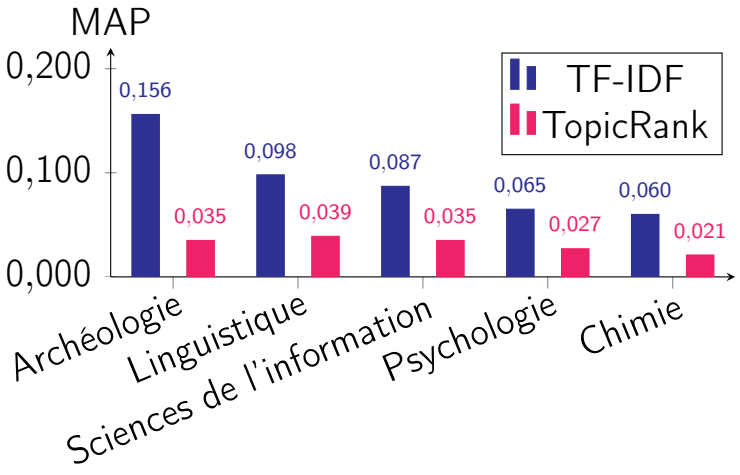
`(NOM | ADJ)+`



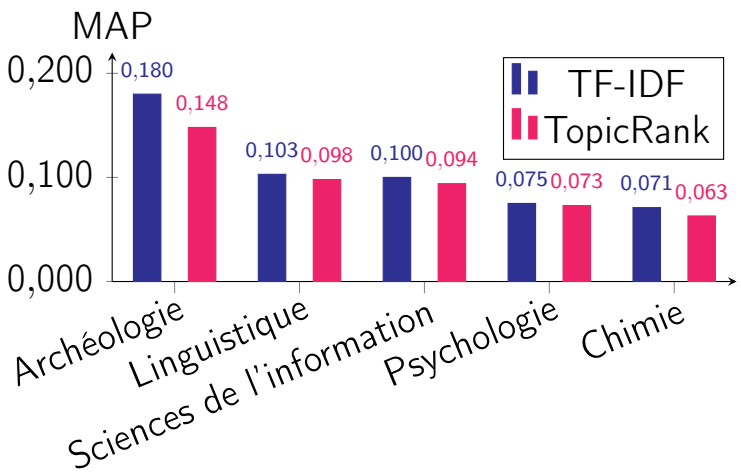
Candidats termes

Expérience

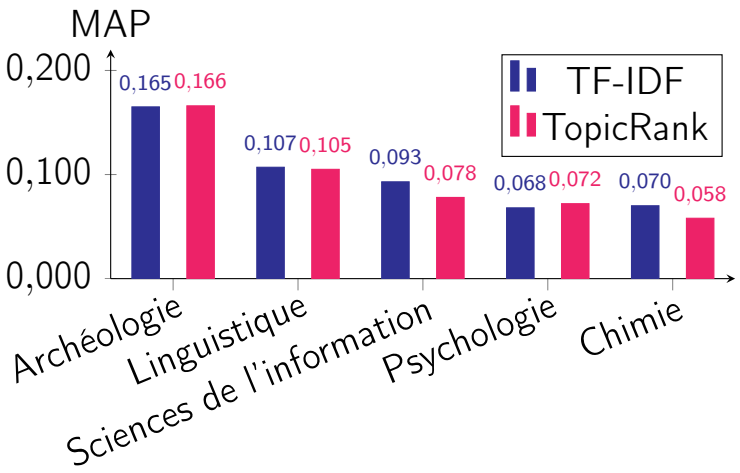
Résultats et observations



{1..3}-grammes



`(NOM | ADJ)+`

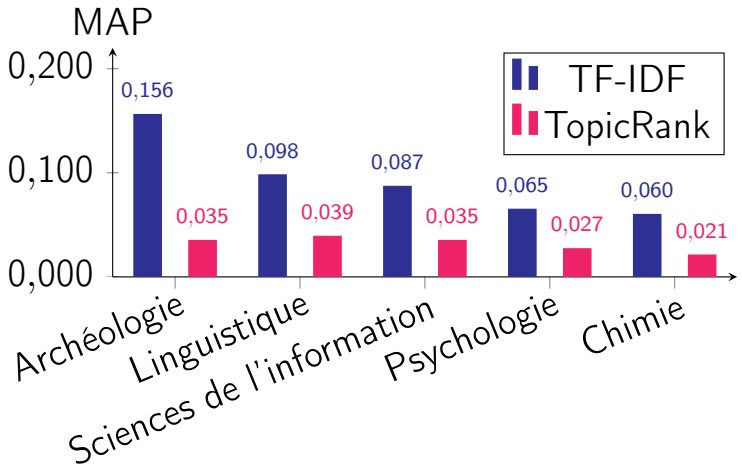


Candidats termes

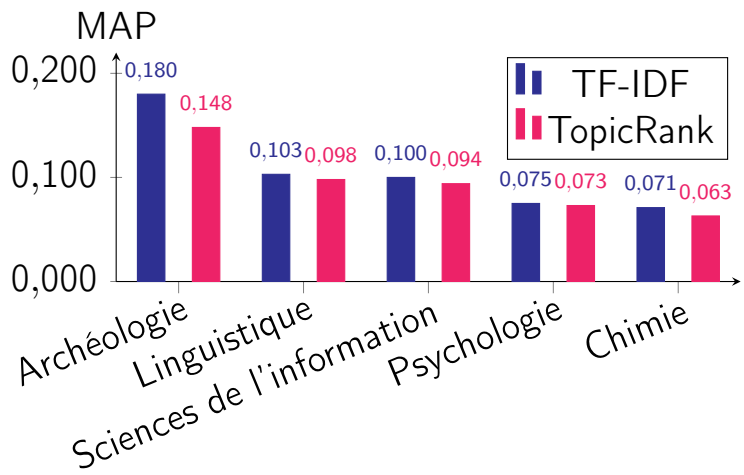
- Même échelle de difficulté pour TF-IDF et TopicRank
- Meilleur stabilité du TF-IDF
  - $\Rightarrow$ Spécificité IDF
- TopicRank dépend de la qualité des candidats

Expérience

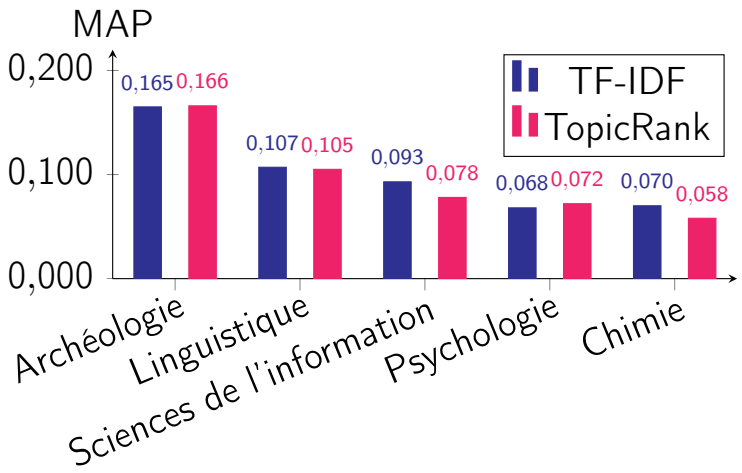
Résultats et observations



{1..3}-grammes



`(NOM | ADJ)+`



Candidats termes

- Même échelle de difficulté pour TF-IDF et TopicRank
- Meilleur stabilité du TF-IDF
  - $\Rightarrow$ Spécificité IDF
- TopicRank dépend de la qualité des candidats

Expérience

Observations :

1. Même échelle de difficulté pour les deux méthodes
2. Meilleure stabilité du TF-IDF en fonction de la qualité des candidats

   $\Rightarrow$ Les candidats retournés sont les candidats spécifiques
3. Plus les candidats sont de bonne qualité, plus TopicRank est compétitif avec TF-IDF

Conclusions :

1. Il y a bien une influence de la discipline sur la difficulté de l'extraction de termes-clés
2. La présence ou non, dans les termes-clés, de mots à usage courant dans le language de la discipline est un facteur influent

   $\rightarrow$ *« réaction sonochimique », « réaction électrochimique », etc.*
3. La cohésion au sein du texte est un facteur influent

Expérience

Observations :

1. Même échelle de difficulté pour les deux méthodes
2. Meilleure stabilité du TF-IDF en fonction de la qualité des candidats

   $\Rightarrow$ Les candidats retournés sont les candidats spécifiques
3. Plus les candidats sont de bonne qualité, plus TopicRank est compétitif avec TF-IDF

Conclusions :

1. Il y a bien une influence de la discipline sur la difficulté de l'extraction de termes-clés
2. La présence ou non, dans les termes-clés, de mots à usage courant dans le language de la discipline est un facteur influent

   $\rightarrow$ *« réaction sonochimique », « réaction électrochimique », etc.*
3. La cohésion au sein du texte est un facteur influent

Expérience

Observations :

1. Même échelle de difficulté pour les deux méthodes
2. Meilleure stabilité du TF-IDF en fonction de la qualité des candidats
   - $\Rightarrow$ Les candidats retournés sont les candidats spécifiques
3. Plus les candidats sont de bonne qualité, plus TopicRank est compétitif avec TF-IDF

Conclusions :

1. Il y a bien une influence de la discipline sur la difficulté de l'extraction de termes-clés
2. La présence ou non, dans les termes-clés, de mots à usage courant dans le language de la discipline est un facteur influent
   - $\rightarrow$ *« réaction sonochimique », « réaction électrochimique », etc.*
3. La cohésion au sein du texte est un facteur influent

Expérience

Observations :

1. Même échelle de difficulté pour les deux méthodes
2. Meilleure stabilité du TF-IDF en fonction de la qualité des candidats

   $\Rightarrow$ Les candidats retournés sont les candidats spécifiques

3. Plus les candidats sont de bonne qualité, plus TopicRank est compétitif avec TF-IDF

Conclusions :

1. Il y a bien une influence de la discipline sur la difficulté de l'extraction de termes-clés
2. La présence ou non, dans les termes-clés, de mots à usage courant dans le language de la discipline est un facteur influent

   $\rightarrow$ *« réaction sonochimique », « réaction électrochimique », etc.*

3. La cohésion au sein du texte est un facteur influent

Expérience

Variabilité du Gravettien de Kostienki (bassin moyen du Don) et des territoires associés

Dans la région de Kostienki-Borschevo, on observe l'expression, à ce jour, la plus orientale du modèle européen de l'évolution du Paléolithique supérieur. Elle est différente à la fois du modèle Sibérien et du modèle de l'Asie centrale. Comme ailleurs en Europe, le Gravettien apparaît à Kostienki vers 28 ka (Kostienki 8 /II/). Par la suite, entre 24-20 ka, les techno-complexes gravettiens sont représentés au moins par quatre faciès dont deux, ceux de Kostienki 21/III/ et Kostienki 4 /II/, ressemblent au Gravettien occidental et deux autres, Kostienki-Avdeevo et Kostienki 11/II/, sont des faciès propres à l'Europe de l'Est, sans analogie à l'Ouest.

Descripteurs (termes-clés) : Europe, Kostienko, Borschevo, variation, typologie, industrie osseuse, industrie lithique, Europe centrale, Avdeevo, Paléolithique supérieur, Gravettien.

Archéologie

1 Il y a bien une influence de la discipline sur la difficulté de

Etude d'un condensat acide isocyanurique-urée-formaldéhyde

La synthèse d'un condensat acide isocyanurique-urée-formaldéhyde utilisant la pyridine en tant que solvant a été effectuée par réaction sonochimique.

Descripteurs (termes-clés) : Réaction sonochimique, hétérocycle azote, cycle 6 chaînons, ether.

Chimie

3 La cohésion au sein du texte est un facteur influent

Expérience

Observations :

1. Même échelle de difficulté pour les deux méthodes
2. Meilleure stabilité du TF-IDF en fonction de la qualité des candidats

   $\Rightarrow$ Les candidats retournés sont les candidats spécifiques

3. Plus les candidats sont de bonne qualité, plus TopicRank est compétitif avec TF-IDF

Conclusions :

1. Il y a bien une influence de la discipline sur la difficulté de l'extraction de termes-clés
2. La présence ou non, dans les termes-clés, de mots à usage courant dans le language de la discipline est un facteur influent

   $\rightarrow$ « *réaction sonochimique* », « *réaction électrochimique* », *etc.*

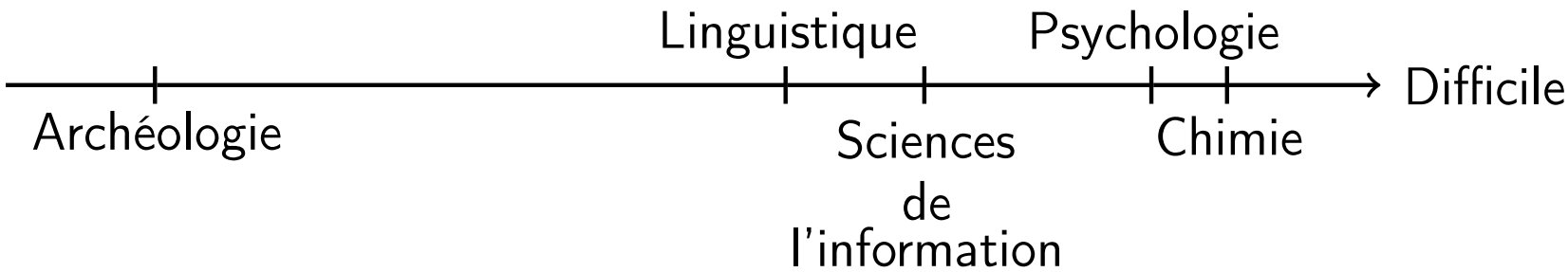
3. La cohésion au sein du texte est un facteur influent

Plan

1. Données
2. Extraction de termes-clés
3. Expérience
4. Conclusion et perspectives

Conclusion

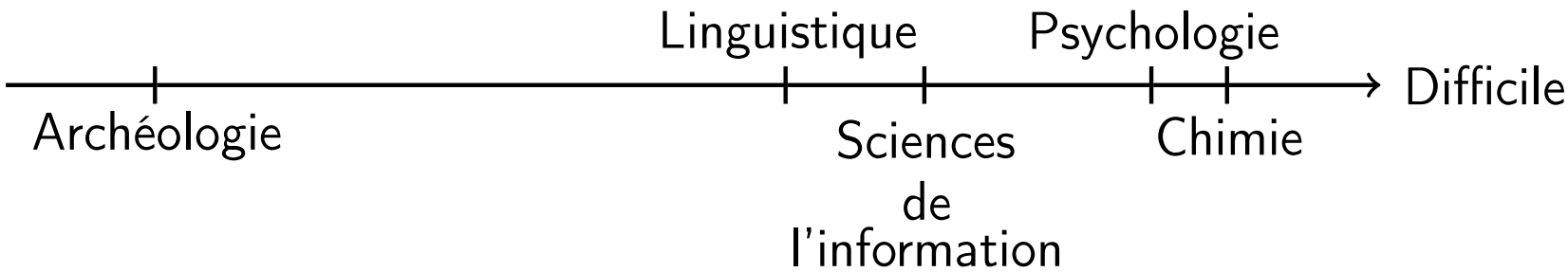
- La difficulté de l'extraction de termes-clés est liée à la discipline :



- Deux facteurs observés :
  - Vocabulaire de la discipline
  - Organisation du discours dans le résumé

Conclusion

- La difficulté de l'extraction de termes-clés est liée à la discipline :



- Deux facteurs observés :
  - Vocabulaire de la discipline
  - Organisation du discours dans le résumé

Perspectives

- Élargir notre étude aux articles complets
- Vérifier l'hypothèse de départ avec plus de disciplines
- Mesurer automatiquement la difficulté a priori
- Adapter automatiquement les méthodes selon la difficulté

MERCI

Références

Rada Mihalcea et Paul Tarau : TextRank : Bringing Order Into Texts. In Dekang Lin et Dekai Wu, éditeurs : Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.