15 mars 2011

Extraction d'information conceptuelle de textes, basée sur une annotation interlingue et guidée par une ontologie

David Rouquet, Achille Falaise

Projet ANR OMNIA

LIG-GETALP, LIRIS, XEROX

Introduction

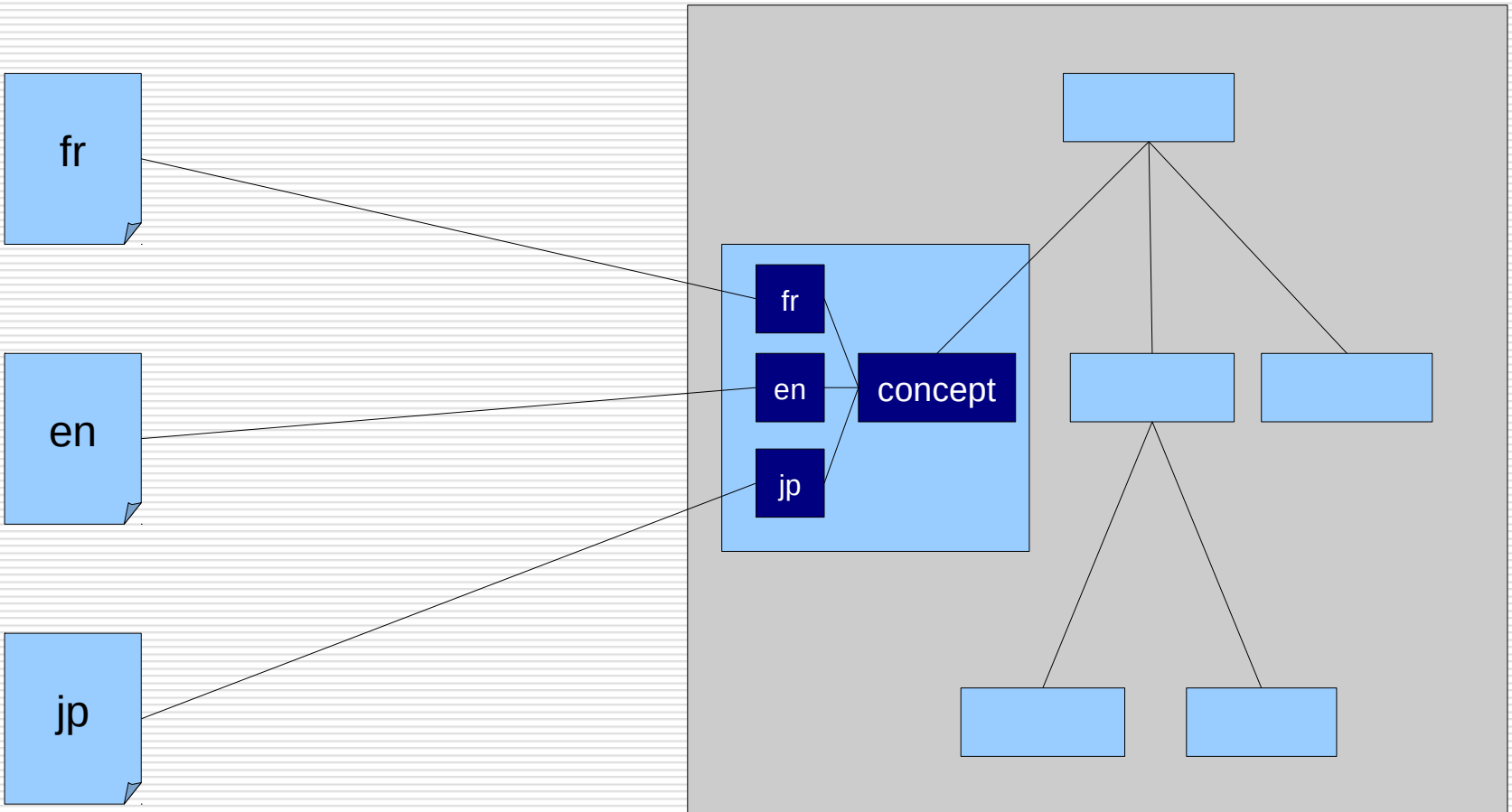
Contexte

- Recherche d'information
  - Annotation
  - Indexation
  - (*Recherche*)
- Approche par ontologie de domaine

Problèmes

- Multilinguisme
- Passage à l'échelle

L'existant en annotation multilingue



Documents

Annotation conceptuelle

Ontologie

Difficultés de l'approche multilingue

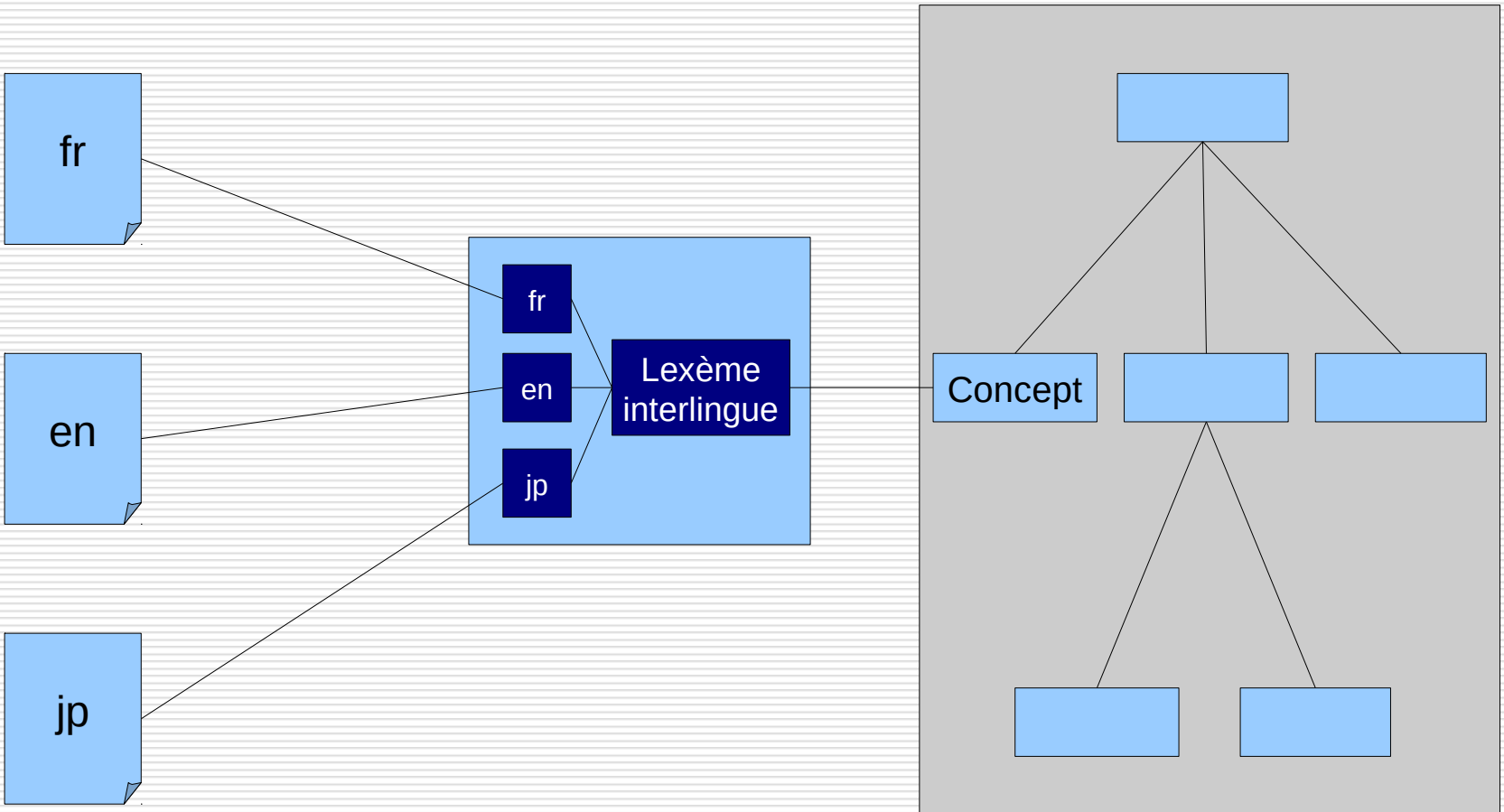
Au niveau de l'ontologie

- Experts multilingues
- Polysémie, ambiguïtés

Au niveau de la recherche d'information

- Couverture pour l'information extra-ontologique

Approche interlingue



Documents

Annotation interlingue

Lexique interlingue

Ontologie

Intérêts de l'approche interlingue

Séparation formelle de l'ontologique (signifié) et du linguistique (signifiant)

- Expertise
  - Du domaine
  - (Multi)linguistique
- Ressources
  - Ontologie de domaine monolingue
  - Dictionnaires bilingues
  - → dictionnaire et ontologie sont indépendants

Recherche d'information multilingue multiniveau

- Au niveau conceptuel
- Au niveau lexical
- → meilleure couverture, mise en avant des notions « métier »

Projet OMNIA

Objectif

- Annotation + indexation
  - multi-facettes (sens, émotion)
  - de documents
    - multimédia (image + texte)
    - multilingues (anglais, français)

Partenaires

- LIG-GETALP : texte
- LIRIS : images/émotions
- XEROX : images/thèmes

Base de textes

- BELGA-news : 500k couples texte (~50 mots, anglais) + image

Base de textes

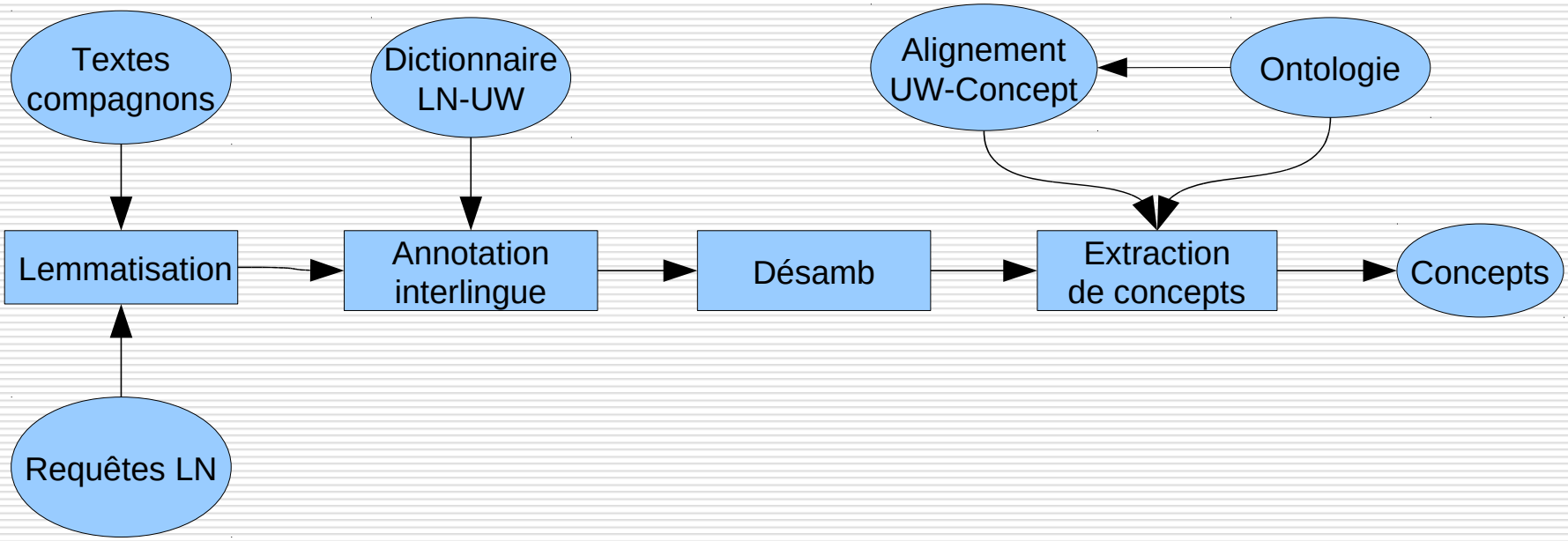
–

1012—Australian Open champion Mary Pierce of France volleys the ball during the second-round match against Kyoko Nagatsuka of Japan in the Toray Pan Pacific Open women's tennis tournament in Tokyo 02 Feburary. Pierce defeated Nagatsuka 6-4, 6-0.

-

1050730—AWA05 - 20020924 - BAGHDAD, IRAQ : Iraqi women sit under a portrait of Iraqi President Saddam Hussein in a waiting room in Baghdad's al-Mansur hospital 24 September 2002. Saddam Hussein is doggedly pursuing the development of weapons of mass destruction and will do his best to hide them from UN inspectors, the British government claimed in a 55-page dossier made public just hours before a special House of Commons debate on Iraq. Iraqi Culture Minister Hamad Yussef Hammadi called the British allegations "baseless." EPA PHOTO AFPI AWAD AWAD -

Architecture



Lemmatisation

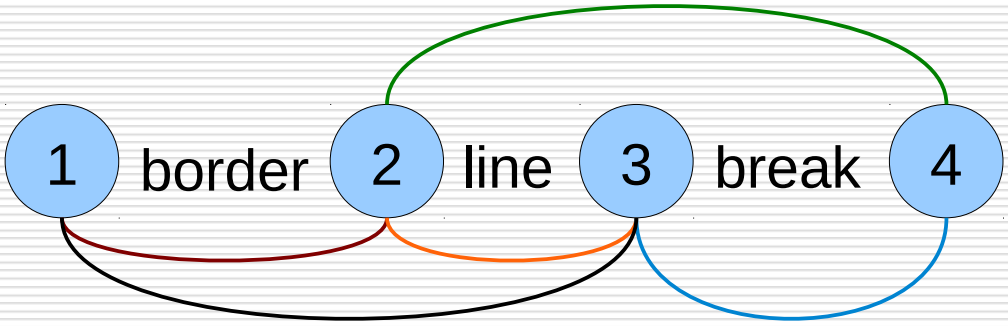
Processus dépendant de la langue

Ressources : dictionnaires forme → lemme

- DELA (anglais, 425k formes → 238k lemmes)
- DELAF (français, 748k formes → 194k lemmes)

Préserver les ambiguïtés

- Lexicales
- Segmentales



-1- LEMMA(border) -2-
-1- LEMMA(border line) -3-
-2- LEMMA(line) -3-
-2- LEMMA(line break) -4-
-3- LEMMA(break) -4-

Lemmatisation

Iraqi women sit under a portrait in hospital.

-0-$OCC($FORME( ))-1-
-1-$OCC($FORME(Iraqi),$LU($LEMMA(Iraqi),$CAT(ADJ)))-2-
-1-$OCC($FORME(Iraqi),$LU($LEMMA(Iraqi),$CAT(NOUN)))-2-
-2-$OCC($FORME( ))-3-
-3-$OCC($FORME(women),$LU($LEMMA(woman),$CAT(NOUN)))-4-
-4-$OCC($FORME( ))-5-
-5-$OCC($FORME(sit),$LU($LEMMA(sit),$CAT(NOUN)))-6-
-5-$OCC($FORME(sit),$LU($LEMMA(sit),$CAT(VERB)))-6-
-6-$OCC($FORME( ))-7-
-7-$OCC($FORME(under),$LU($LEMMA(under),$CAT(ADV)))-8-
-7-$OCC($FORME(under),$LU($LEMMA(under),$CAT(ADJ)))-8-
-7-$OCC($FORME(under),$LU($LEMMA(under),$CAT(UNK)))-8-
-7-$OCC($FORME(under),$LU($LEMMA(under),$CAT(PREP)))-8-
-8-$OCC($FORME( ))-9-
-9-$OCC($FORME(a),$LU($LEMMA(a),$CAT(DET)))-10-
-9-$OCC($FORME(a),$LU($LEMMA(a),$CAT(NOUN)))-10-
-10-$OCC($FORME( ))-11-
-11-$OCC($FORME(portrait),$LU($LEMMA(portrait),$CAT(NOUN)))-12-
-12-$OCC($FORME( ))-13-
-13-$OCC($FORME(in),$LU($LEMMA(in),$CAT(ADJ)))-14-
-13-$OCC($FORME(in),$LU($LEMMA(in),$CAT(NOUN)))-14-
-13-$OCC($FORME(in),$LU($LEMMA(in),$CAT(UNK)))-14-
-13-$OCC($FORME(in),$LU($LEMMA(in),$CAT(PREP)))-14-
-14-$OCC($FORME( ))-15-
-15-$OCC($FORME(hospital),$LU($LEMMA(hospital),$CAT(NOUN)))-16-
-16-$OCC($FORME(.))-17-

Annotation interlingue

Lexique interlingue : les *Universal Words*

- Lexique d'UNL (1996+, langage pivot formel non-ambigu)
- Représente une acception sans ambiguïté
- Exemples :
  - book(icl>do,agt>human,obj>thing)
  - book(icl>thing)
  - ikebana(icl>flower_arrangement)
- 200k UW++ construites à partir de WordNet
  - UW++ (lexique) ≠ Wordnet (réseau lexical)

Ressources : alignements lemme → UW

- 192k lemmes anglais
- 43k lemmes français
- Ressources importantes (100k+ lemme→UW) disponibles pour arabe, allemand, chinois simplifié, espagnol, hindi, indonésien, japonais, coréen (http://www.undl.org/unlexp/)

Annotation interlingue

Iraqi women sit under a portrait in hospital.

-0-$OCC($FORME( ))-1-

-1-$OCC($FORME(Iraqi),$LU($LEMMA(Iraqi),$CAT(ADJ)))-2-

-1-$OCC($FORME(Iraqi),$LU($LEMMA(Iraqi),$CAT(NOUN),$UW($ID(unl.upp.Iraqi.95911),$HW(Iraqi),$REST(icl>Asian>thing))))-2-

-2-$OCC($FORME( ))-3-

-3-$OCC($FORME(women),$LU($LEMMA(woman),$CAT(NOUN),$UW($ID(unl.upp.woman.203992),$HW(woman),$REST(icl>female>thing,ant>man)),
$UW($ID(unl.upp.woman.203991),$HW(woman),$REST(icl>female>thing,ant>man)),$UW($ID(unl.upp.woman.203990),$HW(woman),
$REST(icl>cleaner>thing,equ>charwoman)),$UW($ID(unl.upp.woman.203989),$HW(woman),$REST(icl>class>thing,equ>womanhood))))-4-

-4-$OCC($FORME( ))-5-

-5-$OCC($FORME(sit),$LU($LEMMA(sit),$CAT(NOUN)))-6-

-5-$OCC($FORME(sit),$LU($LEMMA(sit),$CAT(VERB),$UW($ID(unl.upp.sit.159289),$HW(sit),$REST(icl>put>do,equ>seat,agt>thing,obj>thing)),
$UW($ID(unl.upp.sit.159288),$HW(sit),$REST(icl>guard>do,equ>baby_sit,agt>thing)),$UW($ID(unl.upp.sit.159287),$HW(sit),
$REST(icl>travel>do,equ>ride,agt>thing)),$UW($ID(unl.upp.sit.159286),$HW(sit),$REST(icl>expose>do,equ>model,agt>thing)),$UW($ID(unl.upp.sit.159285),
$HW(sit),$REST(icl>convene>occur,obj>thing)),$UW($ID(unl.upp.sit.159284),$HW(sit),$REST(icl>change_posture>do,equ>sit_down,agt>thing)),
$UW($ID(unl.upp.sit.159283),$HW(sit),$REST(icl>be>occur,obj>thing)),$UW($ID(unl.upp.sit.159282),$HW(sit),$REST(icl>occur,obj>thing))))-6-

-6-$OCC($FORME( ))-7-

-7-$OCC($FORME(under),$LU($LEMMA(under),$CAT(ADV)))-8-

-7-$OCC($FORME(under),$LU($LEMMA(under),$CAT(ADJ)))-8-

-7-$OCC($FORME(under),$LU($LEMMA(under),$CAT(UNK)))-8-

-7-$OCC($FORME(under),$LU($LEMMA(under),$CAT(PREP)))-8-

-8-$OCC($FORME( ))-9-

-9-$OCC($FORME(a),$LU($LEMMA(a),$CAT(DET)))-10-

-9-$OCC($FORME(a),$LU($LEMMA(a),$CAT(NOUN)))-10-

-10-$OCC($FORME( ))-11-

-11-$OCC($FORME(portrait),$LU($LEMMA(portrait),$CAT(NOUN),$UW($ID(unl.upp.portrait.140930),$HW(portrait),$REST(icl>likeness>thing)),
$UW($ID(unl.upp.portrait.140929),$HW(portrait),$REST(icl>word_picture>thing,equ>portrayal)),$UW($ID(unl.upp.portrait.140928),$HW(portrait),
$REST(icl>painting>thing))))-12-

-12-$OCC($FORME( ))-13-

-13-$OCC($FORME(in),$LU($LEMMA(in),$CAT(ADJ)))-14-

-13-$OCC($FORME(in),$LU($LEMMA(in),$CAT(NOUN),$UW($ID(unl.upp.in.93729),$HW(in),$REST(icl>linear#5fnit>thing,equ>inch))))-14-

-13-$OCC($FORME(in),$LU($LEMMA(in),$CAT(UNK)))-14-

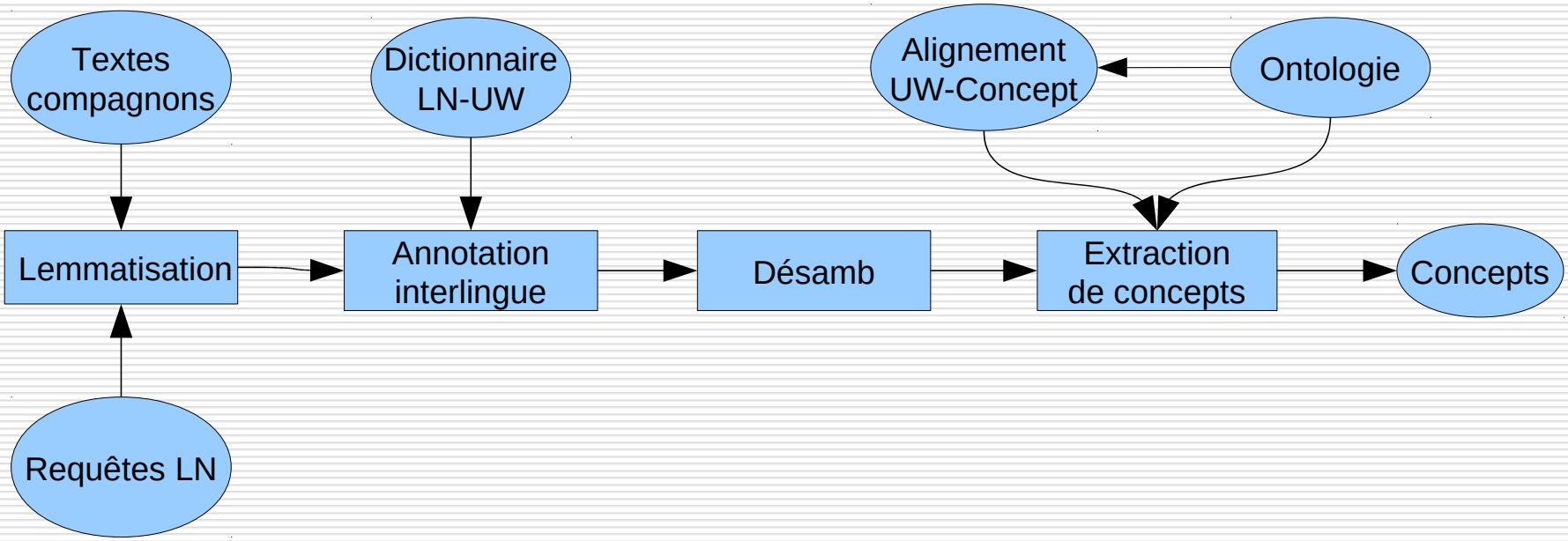
-13-$OCC($FORME(in),$LU($LEMMA(in),$CAT(PREP)))-14-

-14-$OCC($FORME( ))-15-

-15-$OCC($FORME(hospital),$LU($LEMMA(hospital),$CAT(NOUN),$UW($ID(unl.upp.hospital.89342),$HW(hospital),$REST(icl>medical_institution>thing)),
$UW($ID(unl.upp.hospital.89341),$HW(hospital),$REST(icl>medical_building>thing))))-16-

-16-$OCC($FORME(.))-17-

Architecture



Désambiguïsation

Iraqi women sit under a portrait in hospital.

-0- $OCC($FORME( )) -1-

-1- $OCC($FORME(Iraqi),$LU($LEMMA(Iraqi),$CAT(ADJ))) -2-

-1- $OCC($FORME(Iraqi),$LU($LEMMA(Iraqi),$CAT(NOUN),$UW($ID(unl.upp.Iraqi.95911),$HW(Iraqi),$REST(icl>Asian>thing),$SCORE(1.0),$OverallScore(0.003472222222222222)))) -2-

-2- $OCC($FORME( )) -3-

-3- $OCC($FORME(women),$LU($LEMMA(woman),$CAT(NOUN),$UW($ID(unl.upp.woman.203992),$HW(woman),$REST(icl>female>thing,ant>man),$SCORE(0.18309859154929578), $OverallScore(0.003472222222222222)),$UW($ID(unl.upp.woman.203991),$HW(woman),$REST(icl>female>thing,ant>man),$SCORE(0.18309859154929578), $OverallScore(0.003472222222222222)),$UW($ID(unl.upp.woman.203990),$HW(woman),$REST(icl>cleaner>thing,equ>charwoman),$SCORE(0.3098591549295774), $OverallScore(0.034722222222222224)),$UW($ID(unl.upp.woman.203989),$HW(woman),$REST(icl>class>thing,equ>womanhood),$SCORE(0.323943661971831), $OverallScore(0.03819444444444445)))) -4-

-4- $OCC($FORME( )) -5-

-5- $OCC($FORME(sit),$LU($LEMMA(sit),$CAT(NOUN))) -6-

-5- $OCC($FORME(sit),$LU($LEMMA(sit),$CAT(VERB),$UW($ID(unl.upp.sit.159289),$HW(sit),$REST(icl>put>do,equ>seat,agt>thing,obj>thing),$SCORE(0.15), $OverallScore(0.041666666666666664)),$UW($ID(unl.upp.sit.159288),$HW(sit),$REST(icl>guard>do,equ>baby_sit,agt>thing),$SCORE(0.15),$OverallScore(0.041666666666666664)), $UW($ID(unl.upp.sit.159287),$HW(sit),$REST(icl>travel>do,equ>ride,agt>thing),$SCORE(0.058333333333333334),$OverallScore(0.003472222222222222)), $UW($ID(unl.upp.sit.159286),$HW(sit),$REST(icl>expose>do,equ>model,agt>thing),$SCORE(0.06666666666666667),$OverallScore(0.006944444444444444)), $UW($ID(unl.upp.sit.159285),$HW(sit),$REST(icl>convene>occur,obj>thing),$SCORE(0.15833333333333333),$OverallScore(0.04513888888888889)),$UW($ID(unl.upp.sit.159284), $HW(sit),$REST(icl>change_posture>do,equ>sit_down,agt>thing),$SCORE(0.15),$OverallScore(0.041666666666666664)),$UW($ID(unl.upp.sit.159283),$HW(sit), $REST(icl>be>occur,obj>thing),$SCORE(0.1),$OverallScore(0.020833333333333332)),$UW($ID(unl.upp.sit.159282),$HW(sit),$REST(icl>occur,obj>thing), $SCORE(0.16666666666666666),$OverallScore(0.04861111111111111)))) -6-

-6- $OCC($FORME( )) -7-

-7- $OCC($FORME(under),$LU($LEMMA(under),$CAT(ADJ))) -8-

-7- $OCC($FORME(under),$LU($LEMMA(under),$CAT(ADV))) -8-

-7- $OCC($FORME(under),$LU($LEMMA(under),$CAT(PREP))) -8-

-7- $OCC($FORME(under),$LU($LEMMA(under),$CAT(UNK))) -8-

-8- $OCC($FORME( )) -9-

-9- $OCC($FORME(a),$LU($LEMMA(a),$CAT(DET))) -10-

-9- $OCC($FORME(a),$LU($LEMMA(a),$CAT(NOUN))) -10-

-10- $OCC($FORME( )) -11-

-11- $OCC($FORME(portrait),$LU($LEMMA(portrait),$CAT(NOUN),$UW($ID(unl.upp.portrait.140930),$HW(portrait),$REST(icl>likeness>thing),$SCORE(0.27906976744186046), $OverallScore(0.003472222222222222)),$UW($ID(unl.upp.portrait.140929),$HW(portrait),$REST(icl>word_picture>thing,equ>portrayal),$SCORE(0.32558139534883723), $OverallScore(0.017361111111111112)),$UW($ID(unl.upp.portrait.140928),$HW(portrait),$REST(icl>painting>thing),$SCORE(0.3953488372093023), $OverallScore(0.03819444444444445)))) -12-

-12- $OCC($FORME( )) -13-

-13- $OCC($FORME(in),$LU($LEMMA(in),$CAT(ADJ))) -14-

-13- $OCC($FORME(in),$LU($LEMMA(in),$CAT(NOUN),$UW($ID(unl.upp.in.93729),$HW(in),$REST(icl>linear#5fnit>thing,equ>inch),$SCORE(1.0), $OverallScore(0.003472222222222222)))) -14-

-13- $OCC($FORME(in),$LU($LEMMA(in),$CAT(PREP))) -14-

-13- $OCC($FORME(in),$LU($LEMMA(in),$CAT(UNK))) -14-

-14- $OCC($FORME( )) -15-

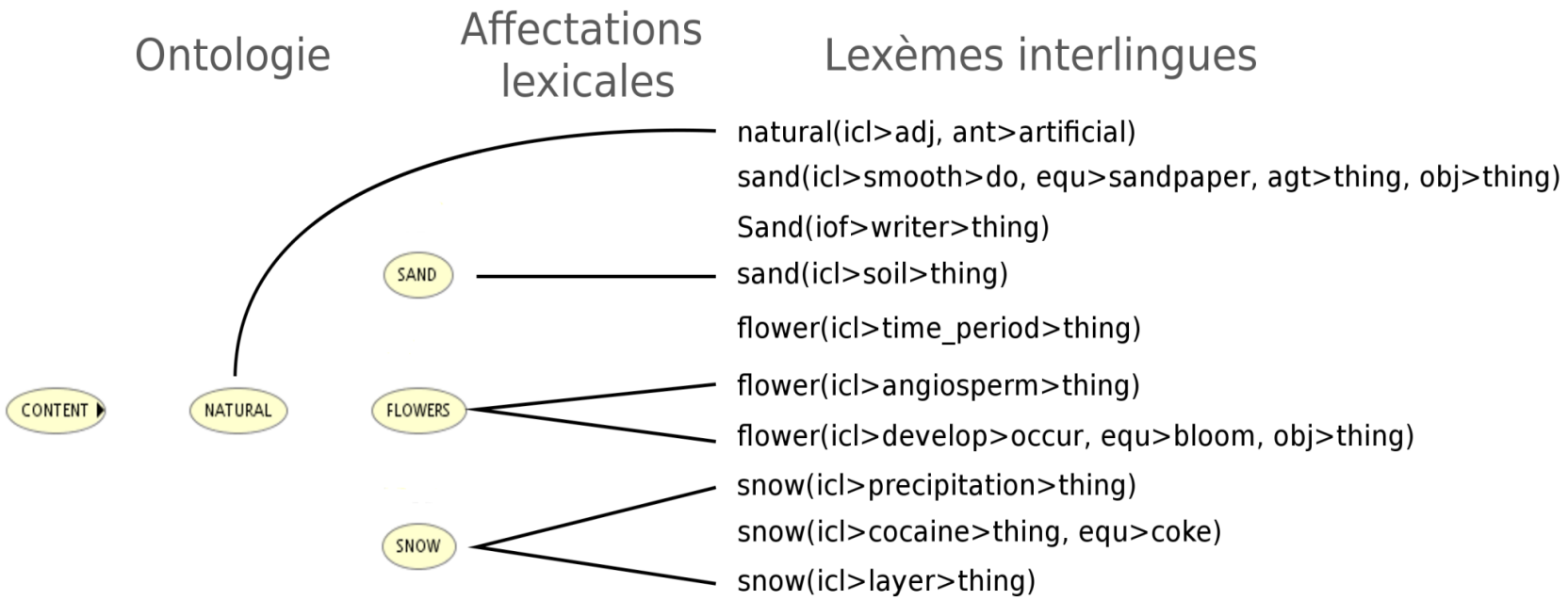
-15- $OCC($FORME(hospital),$LU($LEMMA(hospital),$CAT(NOUN),$UW($ID(unl.upp.hospital.89342),$HW(hospital),$REST(icl>medical_institution>thing), $SCORE(0.5111111111111111),$OverallScore(0.006944444444444444)),$UW($ID(unl.upp.hospital.89341),$HW(hospital),$REST(icl>medical_building>thing), $SCORE(0.4888888888888889),$OverallScore(0.003472222222222222)))) -16-

-16- $OCC($FORME(.)) -17-

Alignement UW → concept

Correspondance automatique

- Pour les ontologies en anglais
- Post-édition possible
- Pondération des liens



Annotation conceptuelle

Iraqi women sit under a portrait in hospital.

-0- $OCC($FORME( )) -1-

-1- $OCC($FORME(Iraqi),$LU($LEMMA(Iraqi),$CAT(ADJ))) -2-

-1- $OCC($FORME(Iraqi),$LU($LEMMA(Iraqi),$CAT(NOUN),$UW($ID(unl.upp.Iraqi.95911),$HW(Iraqi),$REST(icl>Asian>thing),$SCORE(1.0),$OverallScore(0.003472222222222222)))) -2-

-2- $OCC($FORME( )) -3-

-3- $OCC($FORME(women),$LU($LEMMA(woman),$CAT(NOUN),$UW($ID(unl.upp.woman.203992),$CEPT(WOMAN),$Link2CeptAbsScore(1),$HW(woman), $REST(icl>female>thing,ant>man),$SCORE(0.18309859154929578),$OverallScore(0.003472222222222222)),$UW($ID(unl.upp.woman.203991),$CEPT(WOMAN), $Link2CeptAbsScore(1),$HW(woman),$REST(icl>female>thing,ant>man),$SCORE(0.18309859154929578),$OverallScore(0.0034722222222222222)),$UW($ID(unl.upp.woman.203990), $CEPT(WOMAN),$Link2CeptAbsScore(1),$HW(woman),$REST(icl>cleaner>thing,equ>charwoman),$SCORE(0.3098591549295774),$OverallScore(0.034722222222222224)), $UW($ID(unl.upp.woman.203989),$CEPT(WOMAN),$Link2CeptAbsScore(1),$HW(woman),$REST(icl>class>thing,equ>womanhood),$SCORE(0.323943661971831), $OverallScore(0.03819444444444445)))) -4-

-4- $OCC($FORME( )) -5-

-5- $OCC($FORME(sit),$LU($LEMMA(sit),$CAT(NOUN))) -6-

-5- $OCC($FORME(sit),$LU($LEMMA(sit),$CAT(VERB),$UW($ID(unl.upp.sit.159289),$HW(sit),$REST(icl>put>do,equ>seat,agt>thing,obj>thing),$SCORE(0.15), $OverallScore(0.041666666666666664)),$UW($ID(unl.upp.sit.159288),$HW(sit),$REST(icl>guard>do,equ>baby_sit,agt>thing),$SCORE(0.15),$OverallScore(0.041666666666666664)), $UW($ID(unl.upp.sit.159287),$HW(sit),$REST(icl>travel>do,equ>ride,agt>thing),$SCORE(0.058333333333333334),$OverallScore(0.003472222222222222)), $UW($ID(unl.upp.sit.159286),$HW(sit),$REST(icl>expose>do,equ>model,agt>thing),$SCORE(0.06666666666666667),$OverallScore(0.006944444444444444)), $UW($ID(unl.upp.sit.159285),$HW(sit),$REST(icl>convene>occur,obj>thing),$SCORE(0.15833333333333333),$OverallScore(0.04513888888888889)),$UW($ID(unl.upp.sit.159284), $HW(sit),$REST(icl>change_posture>do,equ>sit_down,agt>thing),$SCORE(0.15),$OverallScore(0.041666666666666664)),$UW($ID(unl.upp.sit.159283),$HW(sit), $REST(icl>be>occur,obj>thing),$SCORE(0.1),$OverallScore(0.020833333333333332)),$UW($ID(unl.upp.sit.159282),$HW(sit),$REST(icl>occur,obj>thing), $SCORE(0.16666666666666666),$OverallScore(0.04861111111111111)))) -6-

-6- $OCC($FORME( )) -7-

-7- $OCC($FORME(under),$LU($LEMMA(under),$CAT(ADJ))) -8-

-7- $OCC($FORME(under),$LU($LEMMA(under),$CAT(ADV))) -8-

-7- $OCC($FORME(under),$LU($LEMMA(under),$CAT(PREP))) -8-

-7- $OCC($FORME(under),$LU($LEMMA(under),$CAT(UNK))) -8-

-8- $OCC($FORME( )) -9-

-9- $OCC($FORME(a),$LU($LEMMA(a),$CAT(DET))) -10-

-9- $OCC($FORME(a),$LU($LEMMA(a),$CAT(NOUN))) -10-

-10- $OCC($FORME( )) -11-

-11- $OCC($FORME(portrait),$LU($LEMMA(portrait),$CAT(NOUN),$UW($ID(unl.upp.portrait.140930),$HW(portrait),$REST(icl>likeness>thing),$SCORE(0.27906976744186046), $OverallScore(0.003472222222222222)),$UW($ID(unl.upp.portrait.140929),$HW(portrait),$REST(icl>word_picture>thing,equ>portrayal),$SCORE(0.32558139534883723), $OverallScore(0.017361111111111112)),$UW($ID(unl.upp.portrait.140928),$HW(portrait),$REST(icl>painting>thing),$SCORE(0.3953488372093023), $OverallScore(0.03819444444444445)))) -12-

-12- $OCC($FORME( )) -13-

-13- $OCC($FORME(in),$LU($LEMMA(in),$CAT(ADJ))) -14-

-13- $OCC($FORME(in),$LU($LEMMA(in),$CAT(NOUN),$UW($ID(unl.upp.in.93729),$HW(in),$REST(icl>linear#5fnit>thing,equ>inch),$SCORE(1.0), $OverallScore(0.003472222222222222)))) -14-

-13- $OCC($FORME(in),$LU($LEMMA(in),$CAT(PREP))) -14-

-13- $OCC($FORME(in),$LU($LEMMA(in),$CAT(UNK))) -14-

-14- $OCC($FORME( )) -15-

-15- $OCC($FORME(hospital),$LU($LEMMA(hospital),$CAT(NOUN),$UW($ID(unl.upp.hospital.89342),$CEPT(HOSPITAL),$Link2CeptAbsScore(1),$HW(hospital), $REST(icl>medical_institution>thing),$SCORE(0.5111111111111111),$OverallScore(0.006944444444444444)),$UW($ID(unl.upp.hospital.89341),$CEPT(HOSPITAL), $Link2CeptAbsScore(1),$HW(hospital),$REST(icl>medical_building>thing),$SCORE(0.4888888888888889),$OverallScore(0.003472222222222222)))) -16-

-16- $OCC($FORME(.)) -17-

Extraction conceptuelle

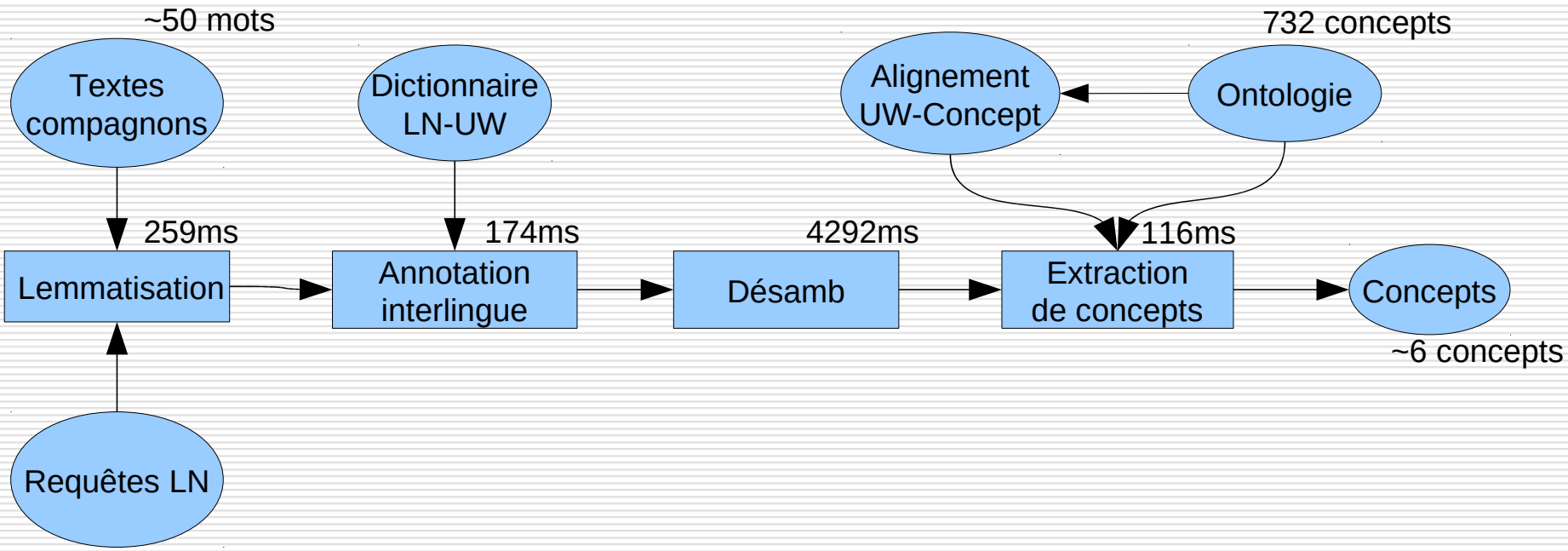
Propagation des concepts

- <contentExtraction>
  <comment>Version 0.4.1</comment>
  <concept name="BUILDING" likelihood="0.00347222222222222"/>
  <concept name="HOSPITAL" likelihood="0.00347222222222222"/>
  <concept name="PERSON" likelihood="0.0381944444444445"/>
  <concept name="WOMAN" likelihood="0.0381944444444445"/>
  </contentExtraction>

Sortie basée uniquement sur l'ontologie

- Finalité
  - Indexation conceptuelle du document
  - Fusion avec les analyses visuelles du LIRIS et de XEROX

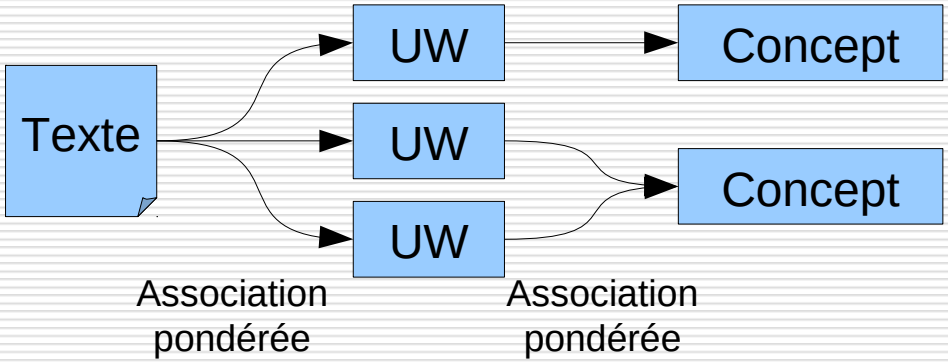
Temps de calcul



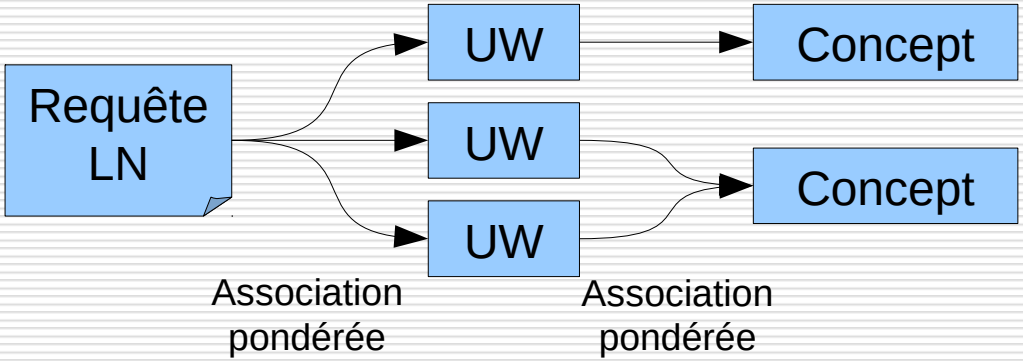
Temps de calcul moyen et nombre moyen de concepts trouvés
pour des textes anglais d'environ 50 mots et une ontologie de 732 concepts

Indexation des textes

Indexation à 2 niveaux



Recherche d'information à 2 niveaux



Calcul d'un score de similarité (UW et concepts)

Démo

- http://aiakide.net/omnia/test2

Conclusion

Approche interlingue

- Séparation formelle entre le signifiant (langue) et le signifié (ontologie)
- Ontologie indépendante de la langue
- Recherche d'information multiniveau (lexicale et conceptuelle) multilingue

Passage à l'échelle

- Système OMNIA : 20k légendes / jour (1M mots/jour)
- Wikimedia Commons : 6k images (+légende) / jour