Détection automatique de sentiments dans les textes

Cours ECDA – 2014-2015

Mathieu Roche

cirad
LA RECHERCHE AGRONOMIQUE POUR LE DÉVELOPPEMENT

TETIS

Le projet Senterritoire

TETIS

Géographes - Informaticiens

Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier

LIRMM

Informaticiens

→ proposer un environnement décisionnel fondé sur une analyse automatique des textes liés à l'aménagement du territoire

Démarche générale

Documents d'actualités

1ère Phase

Extraction des entités spatiales

2ème Phase

07/07/2014

Identification des opinions des acteurs

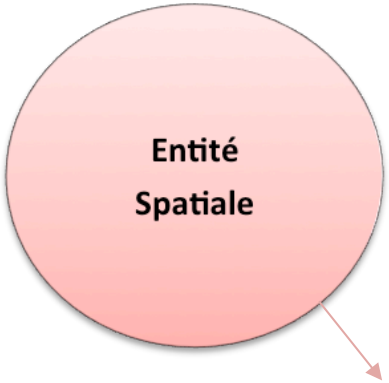
La perception de l'aménagement d'un territoire

Plan

1) Extraction des entités spatiales (ES) – Text2Geo
2) Extraction des relations spatiales (RS)
3) Fouille d'opinion
4) Imagerie satellitaire – Animitex
5) Conclusion et Perspectives

Qu'entend-on par entité spatiale?

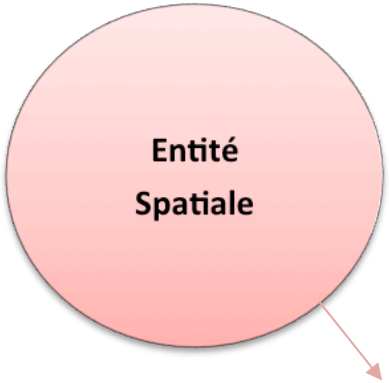
ENTITÉ GÉOGRAPHIQUE VS ENTITÉ SPATIALE *(Usery, 2000)*



Les instruments de musique dans les environs de Montpellier au XIXe siècle

Qu'entend-on par entité spatiale?

ENTITÉ GÉOGRAPHIQUE VS ENTITÉ SPATIALE *(Usery, 2000)*



Les instruments de musique dans les environs de Montpellier au XIXe siècle

PROBLÉMATIQUE

Quelle est la forme de l'entité spatiale dans les textes et comment l'extraire?

Extraction d'Entités Spatiales

Objectif

- Identifier les ES finement (précision) et de façon exhaustive (rappel) et désambiguiser les ENs de type spatial des ENs de type Organisation

Etat de l'art

- Approches TAL (Ehrmann et al., 2006; Bilhaut, 2006): bonne précision
- Approches supervisées de Fouilles de données (Tjong Kim Sang et al., 2003) : bon rappel

Généralement, les méthodes de détection d'ENs permettent uniquement le marquage des toponymes pour les Lieux.

Contribution

Combiner une approche de TAL fondée sur des patrons avec une approche de classification supervisée permettant d'explorer le contexte

Approche TAL à base de patrons (1/3)

Choix du modèle Pivot (Lesbegueries 2007) : définit deux types d'entités spatiales

- Entité spatiale absolue (**ESA**) : <(Indicateur spatiale)*, Entité Nommée>
- Entité spatiale relative (**ESR**) : <(Relation)+, ESA> ; <(Relation) +, ESR>;

Exemples

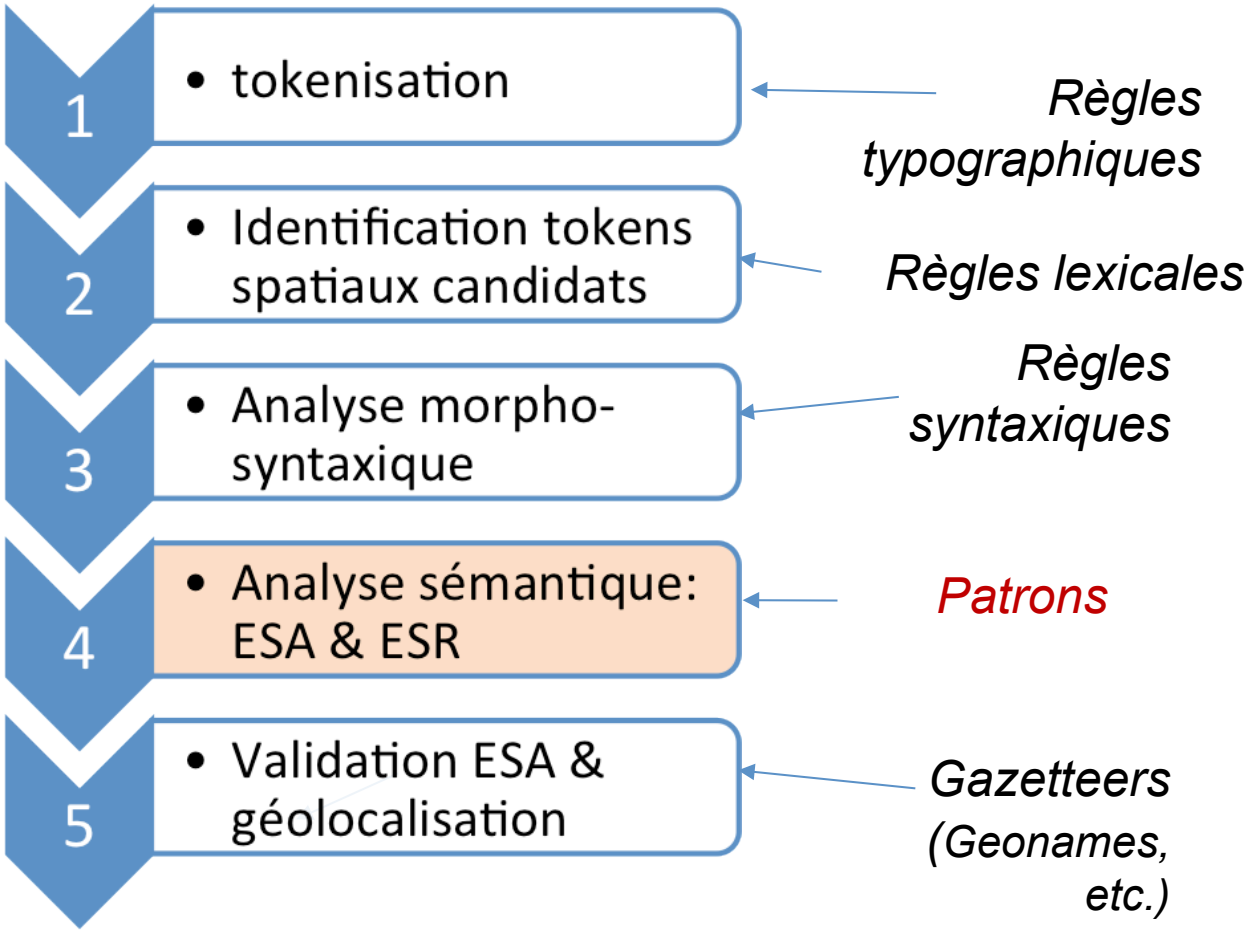
« le sud de la ville de Montpellier »

Relation: **orientation** — **ESA**

«au nord des environs de Montpellier »

Relation: **orientation** — **ESR**

Approche TAL à base de patrons (2/3)



Approche TAL à base de patrons (3/3)

Contribution : Ajout de patrons linguistiques pour l'extraction d'ES définis sur la base des travaux de (Lesbegueries, 2007) et le marquage des Organisations

1. Relation topologique et Indicateurs spatiaux placés après EN de Lieu – *Ex : "Montpellier Nord"*

2. Distribution des relations spatiales et coréférence

Ex : Les environs de Montpellier, Marseille….= les environs de Montpellier & les environs de Marseille.

→ 2 ESRs

3. Entité spatiale vs. Entité d'organisation

<Verbe Action, EN > ; <Préposition, EN>

Ex : "Le projet défendu par Montpellier Agglomération ..."

"La France a autorisé un quota de ..."

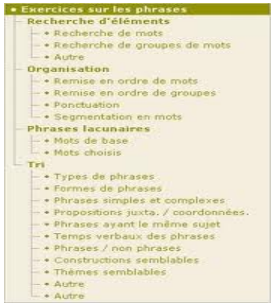
Limites : Les patrons présentés exploitent un contexte local assez réduit

→ Prendre en compte un contexte plus vaste pour distinguer ES et Organisations

Approche hybride RI + EI (1/3) : classification supervisée des phrases

Approche « sac de mots »

a. Classification supervisée pour construire un modèle de prédiction



Liste de phrases (vérité terrain)

$$A := \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

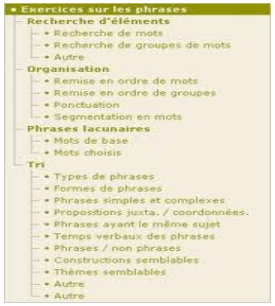
1. Matrice très creuse
- Mots non vides et non rares (1 occurrence) **+ Descripteurs TEXT2GEO**

2. Classes par type d'EN
La phrase décrit une ES ou une Organisation

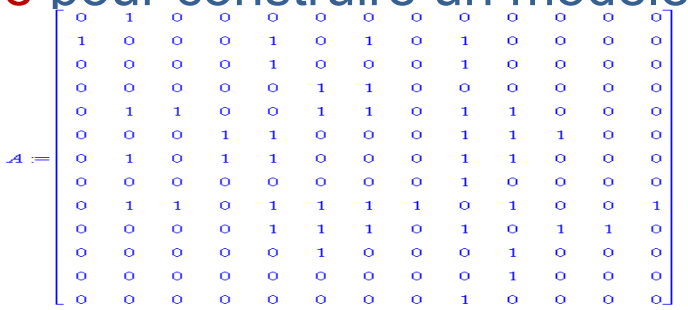
Approche hybride RI + EI (2/3) : classification supervisée des phrases

Approche « sac de mots »

a. Classification supervisée pour construire un modèle de prédiction



Liste de phrases (vérité terrain)



1. Matrice très creuse
- Mots non vides et non rares (1 occurrence) **+ Descripteurs TEXT2GEO**

2. Classes par type d'EN
La phrase décrit une ES ou une Organisation

Avantages de cette représentation

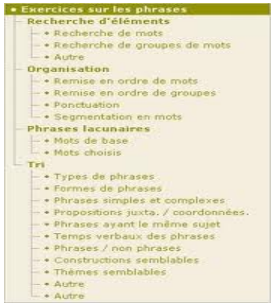
- Donne plus de poids aux mots propres au domaine de la Recherche d'Information Géographique (prépositions spatiales et d'organisation, indicateurs spatiaux, etc.);

- Prise en compte d'un ordre partiel des mots – approche sac de mots classique.

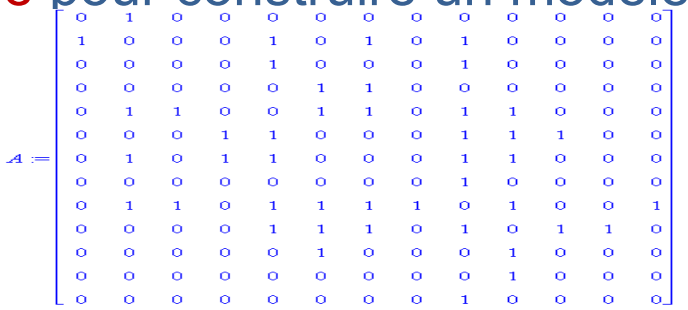
Approche hybride RI + EI (3/3) : classification supervisée des phrases

Approche « sac de mots »

a. Classification supervisée pour construire un modèle de prédiction



Liste de phrases (vérité terrain)



1. Matrice très creuse
- Mots non vides et non rares (1 occurrence) + Descripteurs TEXT2GEO

2. Classes par type d'EN
La phrase décrit une ES ou une Organisation

b. Calcul proximité vecteurs signatures (SVM) (Joachims (1998)

Indicateurs	Indicateurs actuels[40]	Scénario tendanciel 2020[41]	Scénario cible 2020
Croissance potentielle	2 %	1,3 %	2,5% à 3%
Espérance de vie[43]	80,9 ans	82 ans	83 ans[44]
Taux de décrochage scolaire[45][46]	11,8%	> 11,8%	9,5 %
Taux de chômage	9,6 %	9,0 %	4,5%
Dette publique (% PIB)	83,3 %	Plus de 100 %	Moins de 70% et vers 60 %
Solde public (% PIB)	- 8,2 %	- 8 %	Equilibré

Expérimentations

exhaustivité → $\text{Rappel} = \frac{\text{Nombre de réponses correctes extraites}}{\text{Nombre de réponses correctes existantes}}$

qualité → $\text{Précision} = \frac{\text{Nombre de réponses correctes extraites}}{\text{Nombre de toutes les réponses extraites}}$

$$\text{F-mesure} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Corpus
300 phrases

Patrons Pivot

Patrons Text2Geo

Mesure	ESA	ESR	ORG
Précision	53%	84%	92%

	ESA	ESR
Rappel	63%	27%
Précision	20%	48%
F-mesure	30%	34%

Expérimentations

- Ensemble d'apprentissage :
  - 138 phrases contenant des entités spatiales
  - 134 phrases contenant des ENs Organisation.

- Les évaluations données utilisent le principe de validation croisée.

CLASSIFICATION DES PHRASES

SVM

	sp	org
sp	103	35
org	44	90

Taux d'exactitude : 70%

CLASSIFICATION DES PHRASES AVEC CONTRAINTES

Contrainte Préposition org (avec, par...)

	sp	org
sp	108	30
org	47	87

Taux d'exactitude : 71,69%

Contrainte Préposition sp et indicateurs spatiaux (à, en, sud...)

	sp	org
sp	112	26
org	19	115

Taux d'exactitude : 83,45%

Les deux contraintes à la fois

	sp	org
sp	113	25
org	19	115

Taux d'exactitude : 83,82%

Plan

1) Extraction des entités spatiales (ES) – Text2Geo
2) **Extraction des relations spatiales (RS)**
3) Fouille d'opinion
4) Imagerie satellitaire – Animitex
5) Conclusion et Perspectives

Extraction de Relation Spatiale

Objectif

A partir des couples d'ES, identifier les RS associées (directionnelles, topologiques et de distance) sans l'aide d'un corpus annoté.

Etat de l'art

- Modèle d'apprentissage supervisé : avec relations marquées manuellement (Zhang et al 2011), à base de patrons définis manuellement (KordJamshidi et al, 2011) ou découvert par apprentissage (Suchanek et al, 2006)

Contribution

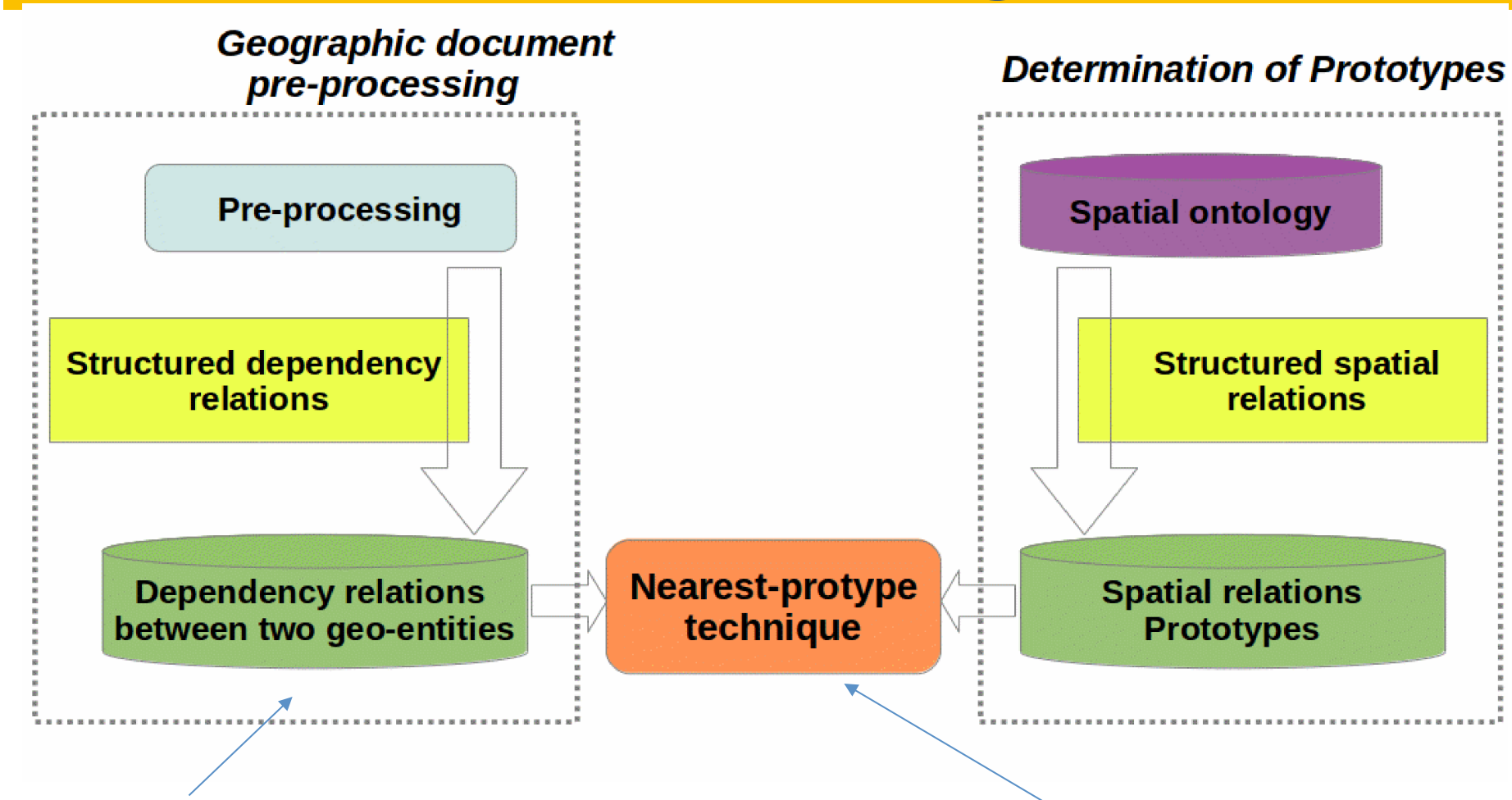
Utiliser une ontologie spatiale (SUMO) pour une approche non supervisée permettant d'identifier les relations spatiales entre 2 entités spatiales (Loglisci et al, 2012)

Approche non supervisée

Avantages

- Détection automatique des instances des relations candidates
- L'ontologie définit les classes de relations spatiales
- Les classes sélectionnées correspondent à des relations spatiales bien identifiées dans la littérature (e.g., topologique)
- Pas de nécessité d'avoir des documents annotés

Le processus global



Les relations de dépendances (RD) entre deux ES peuvent exprimer une relation spatiale

***Similarité** entre les RDs et les prototypes*

Identification des prototypes de relation

spatial relation

+ – – ...

+ – – connected

| + – – meets spatially

- Dans SUMO (Niles et al, 2001) propose une description textuelle des relations spatiales

- Les relations spatiales sont décrites en termes de glosses, d'exemples et d'ensemble de synonymes

meets spatially

⌊ **synset**: adjacent; next; side by side

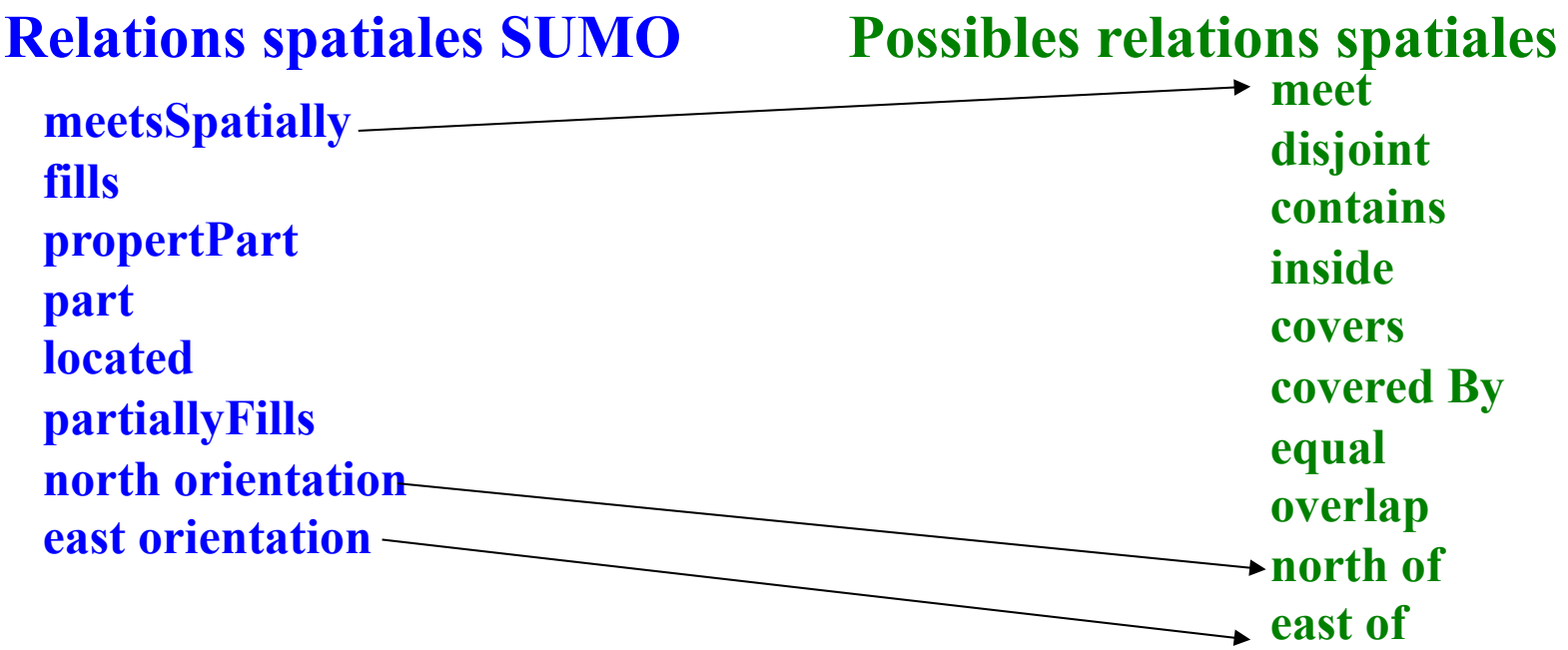
⌊ **glosses**: nearest in space or position;

immediately adjoining without intervening space

⌊ **examples**: "had adjacent rooms"; "in the next room";

Identification des prototypes de relation

- Mise en correspondance des relations de SUMO avec les relations spatiales



Expérimentations

- Documents de Wikipedia de la catégorie “Geography and places”
- La relation de dépendance est affectée à la **classe** la plus “proche”
- Mesures de similarité comme la mesures de Lin 1998 pour calculer la similarité sémantique

$$Tri(Ch1,Ch2) = 1/[\ 1+ |tr(Ch1)| + |tr(Ch2)| - 2\ X\ |tr(Ch1)\ Intersect\ tr(Ch2)|\ ]$$

Exemple : Ch1 = part / Ch2 = part of :
tr(Ch1) = {”par”, ”art”} / tr(Ch2) = {”par”, ”art”, “rt “, “t o”, “ of”}
Tri(part, part of) = 1/[1+2+5-2X2]=0.25

F-score : 0,6 (Relations topologiques) et 0.4 (Relations directionnelles)

Plan

1) Extraction des entités spatiales (ES) – Text2Geo
2) Extraction des relations spatiales (RS)
3) **Fouille d'opinion**
4) Imagerie satellitaire – Animitex
5) Conclusion et Perspectives

Motivations

Objectif

Identification des opinions par des méthodes de fouille de textes

L'orientation sémantique d'une opinion est exprimée par l'intermédiaire des adjectifs (Turney, 2002, Taboada et al., 2006)

Contribution

Construction d'un vocabulaire d'opinion spécialisé et contextualisé (Dray et al, 2009)

Exemple :

"The picture quality of this camera is high"
"The ceilings of the building are high"

Qu'entend-on par Opinion?

- Opinion : négative, positive, neutre ≠ Sentiments : joie, tristesse…

PROBLEMATIQUE

Quelles sont les spécificités des opinions relatives aux informations géographiques et comment les extraire ?

"**J'aime beaucoup le centre ville de Montpellier** et **je déteste la plage de Carnon**"

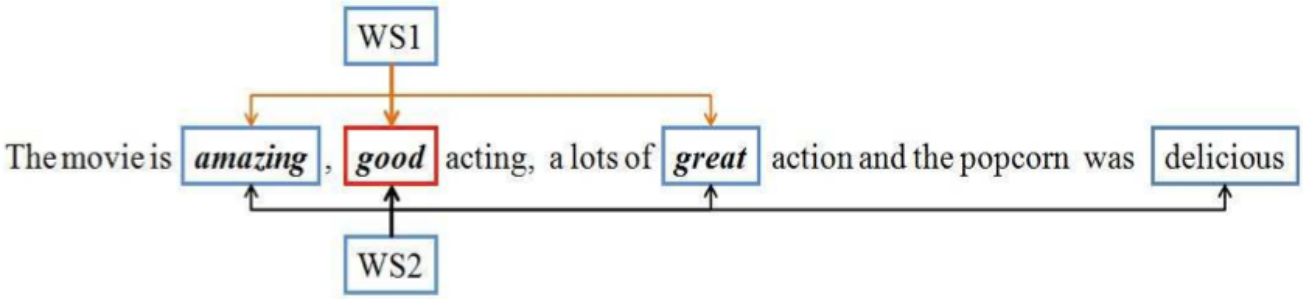
Extraction des adjectifs porteurs d'opinion

- **Choix d' « éléments graines » porteurs d'opinion**

$$P = \{good, nice, excellent, positive, fortunate, correct, superior\}$$
$$Q = \{bad, nasty, poor, negative, unfortunate, wrong, inferior\}$$

Constitution de 14 corpus sur un domaine donné

- ***Premiers traitements :*** *étiquetage, recherche de règles d'association*



Extraction des adjectifs porteurs d'opinion

- **Filtrage par des mesures web**

Principe : calcul d'une dépendance entre deux adjectifs *x*, *y*

$$Dice(x,y) = 2 \times P(x,y) / (P(x)+P(y))$$

[Positif] Funny
Adjective [17.374968071888]
good [118.48338420044]
nice [3.0036930590468]
excellent [0.13462592320458]
positive [0.0030219335932606]
correct [4.9693050945934E-005]
superior [1.6938799522831E-006]
fortunate [6.4929162283948E-018]

Adaptation de la mesure pour :

- prendre en compte les *hits* du web
- prendre en compte le contexte
- calculer une proximité stricte (recherche exacte avec symétrie)

Motivations

Questions :

- pourquoi se limiter aux adjectifs ?
- pourquoi ne pas exploiter des dictionnaires d'opinion ?
- comment les combiner et les contextualiser ?

LIRMM Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier

TETIS

Labex NUMEV

Ressources

Data

SENT_100

99 Articles du *Midi libre* depuis 2006, organisés en deux classes par les experts géographes (positifs et négatifs)

SENT_150

150 articles du *Midi libre* organisés en deux classes par notre équipe (positifs et négatifs)

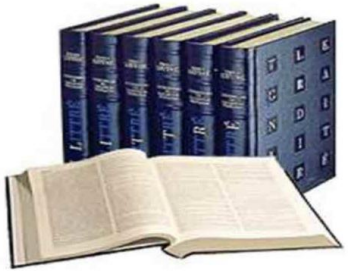
3 corpus DEFT

- 300 extraits d'un débat parlementaire de l'Assemblée nationale;
- 1000 Critiques de jeux vidéos;
- 1500 critiques de films.

Lexicons

General Inquirer

Version française (Bestgen 2011), 1246 mots positifs et 1527 mots négatifs



Jeux de mots (JDM)

9653 mots positifs et 6700 mots négatifs en français

LIWC

Version française (Piolat and al. 2011), sélection des termes positifs et négatifs

Extraction des mots polarisés relatifs au domaine

GeneralInquirer inter ∩ LIWC ∩ JeuxDeMots = S_1 ;

GeneralInquirer ∩ LIWC = S_2 ;

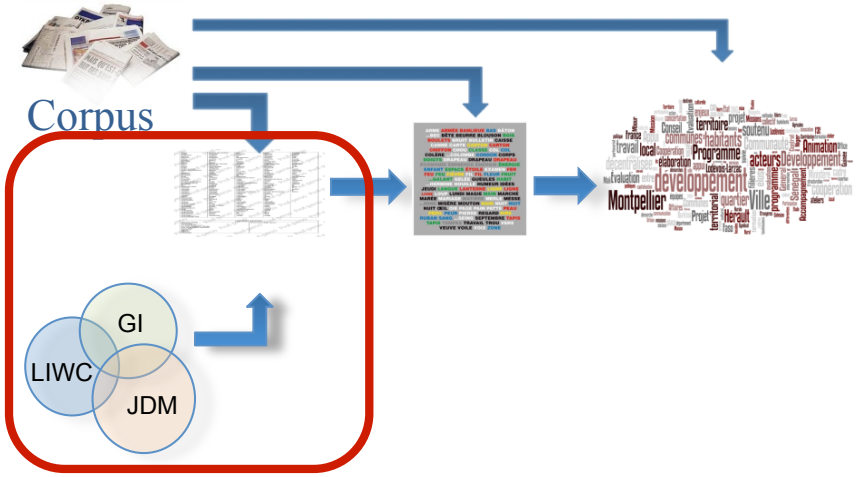
GeneralInquirer ∩ JeuxDeMots = S_3 ;

LIWC ∩ JeuxDeMots = S_4 ;

GeneralInquirer = S_5 : Mots du GeneralInquirer distincts des autres lexiques ;

LIWC = S_6 : Mots du LIWC distincts des autres lexiques ;

JeuxDeMots = S_7 : Mots du JeuxDeMots distincts des autres lexiques.



Evaluation du vocabulaire d'opinions pivots généraliste pour la classification de textes

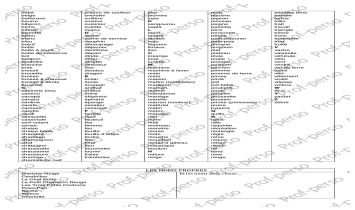
Tests	GI-LIWC-JDM (S_1)	GI-LIWC (S_2)	GI-JDM (S_3)	LIWC-JDM (S_4)	GI (S_5)	LIWC (S_6)	JDM (S_7)	Scores
1	1	1	1	1	1	1	1	**52,5%**
2	1	1	1	1	0	0	0	59,6%
3	3	0	2	0	1	0	1	54,5%
4	3	0	0	2	0	1	1	54,5%
5	3	2	0	0	1	1	0	63,6%
6	3	2	0	0	1	0	0	64,6%

Démarche générale

3. Identification des descripteurs représentatifs de polarité du domaine

Corpus Senterritoire

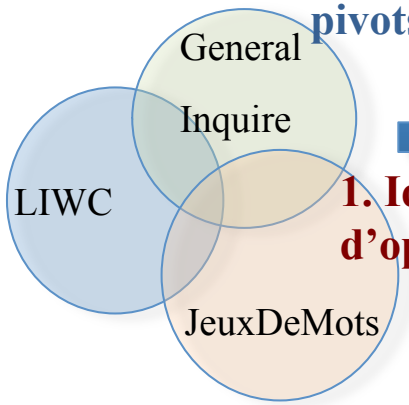
2. Prise en compte du contexte des opinions pivots



Vocabulaire d'opinions pivots généraliste VPOG

Vocabulaire d'opinions contextualisé VOC

Vocabulaire d'opinion spécialisé VOS



Intersection de lexiques d'opinion

1. Identification du lexique d'opinions pivots

Exemples :

magnifique, belle → **ES**

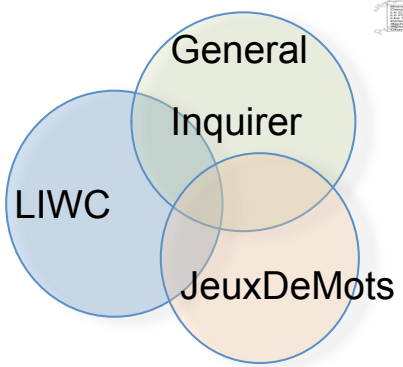
craint, critiqué, solution → **Organisation**

Extraction des mots polarisés relatifs au domaine

Corpus

3. Identification des descripteurs représentatifs de polarité du domaine

2. Analyse contextuelle du vocabulaire d'opinion



1. Sélection des mots polarisés du vocabulaire général

Corpus	Correct score			
	General Inquirer vocabulary	General vocabulary	Contextual vocabulary	Specialized vocabulary
SENT_100	57.5%	64.6%	65.6%	91.9%

Plan

1) Extraction des entités spatiales (ES) – Text2Geo
2) Extraction des relations spatiales (RS)
3) Fouille d'opinion
4) Imagerie satellitaire – Animitex
5) Conclusion et Perspectives

Contexte et Objectif

Contexte

- Mise à disposition de données satellitaires à haute et très haute résolution
- Limite des algorithmes de classification pour une analyse fine des images (par exemple, distinguer les types de cultures)
- Investissement humain conséquent

Objectif

Exploiter des données textuelles massives et hétérogènes (blogs, rapports, articles de presse, etc.) permettant de compléter l'analyse des images satellites (cercle vertueux)

Scénario retenu : *Aménagement du territoire, exemple Hinterland*

Animitex
Analyse d'images fondée sur des informations textuelles

ANalyse d'IMages fondée sur des Informations TEXtuelles

Projet CNRS MASTODONS - Masses de Données Scientifiques et problématique du *big data*

TETIS

iCUBE

LIRMM

GREYC

LIUPPA

http://www.lirmm.fr/~mroche/ANIMITEX

Méthodologie

- Indexation spatiale des images et des documents textuels
- Identification des segments de textes du corpus spatialement pertinents pour une image donnée (exploitation de coordonnées géospatiales via l'utilisation de ressources disponibles)
- Visualisation des résultats
- Désambiguisation de certaines situations (type de végétation par exemple)

Perspectives

- Utilisation d'ontologie des images et des documents textuels
- Utilisation des relations spatiales découvertes dans les textes et existantes dans les images
- Raisonnement, classification et annotation automatique

Plan

1) Extraction des entités spatiales (ES) – Text2Geo
2) Extraction des relations spatiales (RS)
3) Fouille d'opinion
4) Imagerie satellitaire – Animitex
5) **Conclusion et Perspectives**

Perspectives

→ **Viewer Senterritoire** pour les décideurs

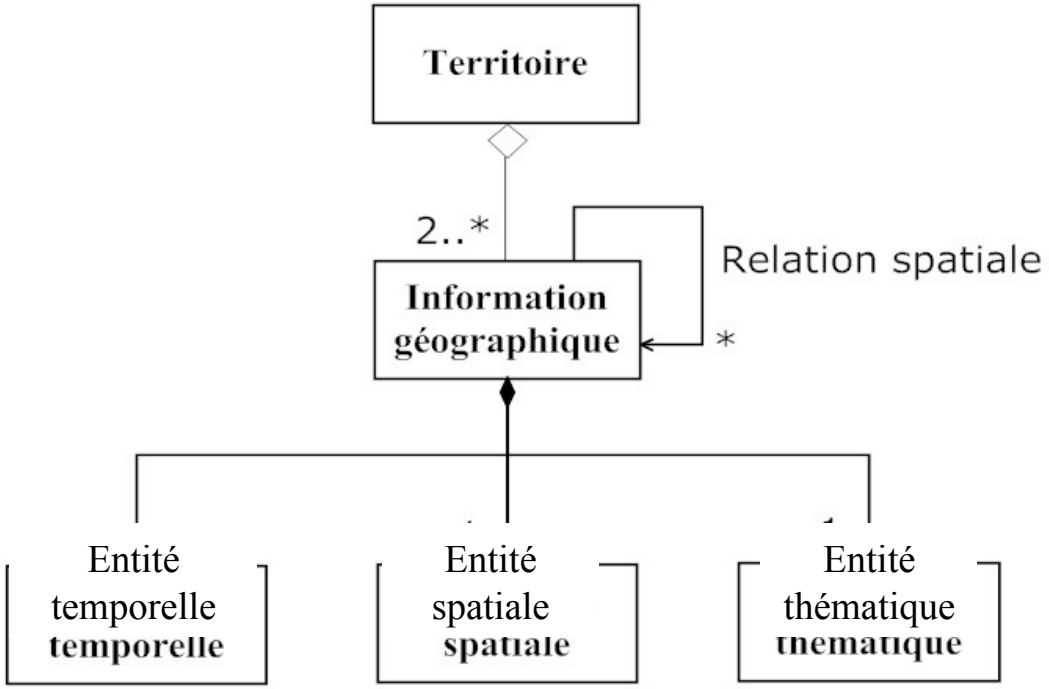
→ Définition de la notion de **Territoire** et extraction des informations correspondantes

→ Prise en compte des composantes **temporelle et thématique**

Perspectives

- Continuer le travail initié avec les géographes

Territoire = un ensemble de lieux, de relations spatiales et temporelles mis en évidence par un ensemble de faits



Et les acteurs dans tout ça ?

- Méthodes d'extraction des entités temporelles et thématiques

Remerciements

Eric Kergosien (TETIS-LIRMM)

Maguelonne Teisseire (Istea, TETIS)

Pascal Poncelet (Univ. Montpellier 2, LIRMM)