Identification automatique des types de relations spatiales dans les textes

Atelier GAST'2016

(EGC 2016, REIMS)

Sarah ZENASNI

Maguelonne TEISSEIRE, Mathieu ROCHE

19/01/2016

UNIVERSITÉ
DE MONTPELLIER

Plan

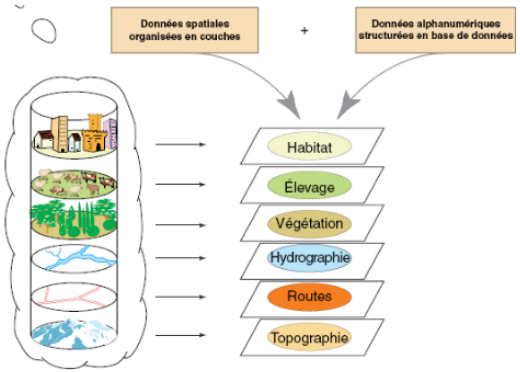
- Brève introduction
  - Information spatiale et relation spatiale
- Méthode
  - Identification de trois types de relations de relations spatiales
- Expérimentations
  - Travail sur le corpus SpRL (campagne SEM-EVAL 2013)
- Conclusion
- Perspectives
  - Généricité de l'approche sur des corpus français (presse, sms)

Brève introduction

- Une requête sur cinq effectuée est liée à la géographie (Sanderson et al., 2004).
- 13-15% des requêtes Web transmises aux moteurs de recherche contiennent un nom de lieu (Jones et al., 2008).

Brève introduction

Système d'information géographique



Brève introduction

Motivations

- L'extraction et l'identification d'informations spatiales à partir des données textuelles.

Motivations

- Identification d'Entités Spatiales :
  - (Nadeau et Sekine, 2007).
  - (Sallaberry et al., 2007).

- Exemple : « La ville de Montpellier a investi dans le club de football »

- Identification de Relations Spatiales :
  - (Grefenstette, 1994).
  - (Weissenbacher et Nazarenko, 2007).

Motivations

- **Objectif global de la thèse** : Identifier des relations sémantiques de façon plus fines (spatiales et non spatiales) entre Entités nommées (Organisation, Personne et Entité spatiale).

- **Objectif du travail présenté** : Identification des types de relations spatiales de façon automatique.

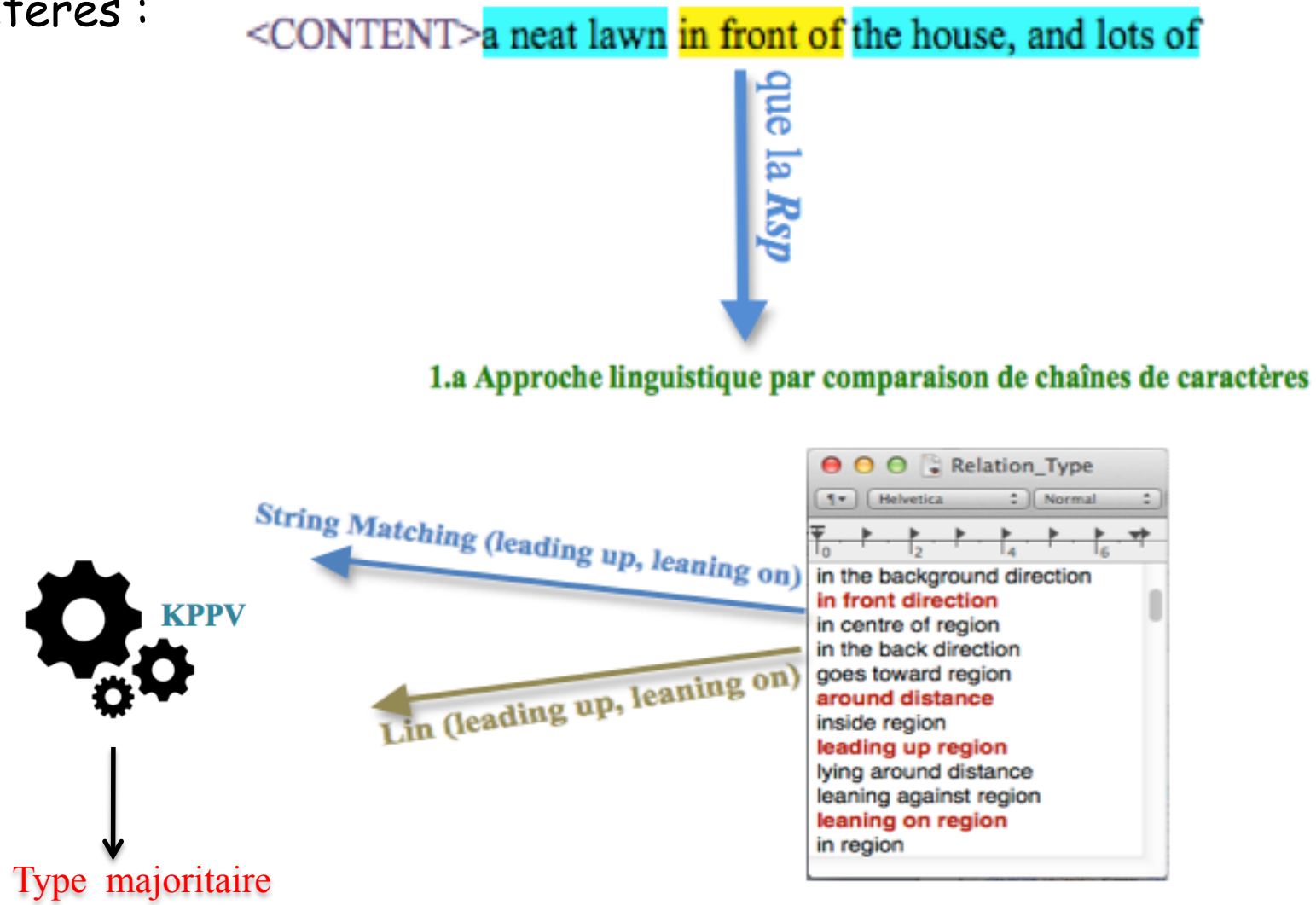
Méthode

- Identification des trois types de relations spatiales de façon automatique.

- Exemple :
  - Street leading up the hill ⟶ Région
  - Four locals are sitting on a bench in a canteen kitchen, leaning on a red brick wall ⟶ Région
  - There are pictures of trees on the wall at the back ⟶ Direction
  - There are three people close to the geysers ⟶ Distance

Par comparaison de chaînes de caractères

- Prédiction du type de relation par comparaison de chaînes de caractères :



Par comparaison de chaînes de caractères

- String Matching (SM) :
  - Distance de Levenshtein (L)

Ch1 :	l	e	a	d	i	n	g	u	p
Opération :				Remplacement				Remplacement	Remplacement
Ch2 :	l	e	a	n	i	n	g	o	n

  - Opérations : Remplacement, Insertion et suppression.

SM(Ch1, Ch2) = max[0; (min(|Ch1| , |Ch2|) - L(Ch1, Ch2)) / min(|Ch1| , |Ch2|)]

  - Exemple
    SM (leading up, leaning on) = max [0, (10 - 3) / 10] = 0,70

Par comparaison de chaînes de caractères

- Lin :
  - N-grammes de caractères (tr)
    - tr (Str1) = {lea,ead,adi,din,ing,ng ,g u, up} = 8
    - tr (Str2) = {lea,ean,ani,nin,ing,ng ,g o, on} = 8
    - tr (Str1) \ tr (Str2) = 3

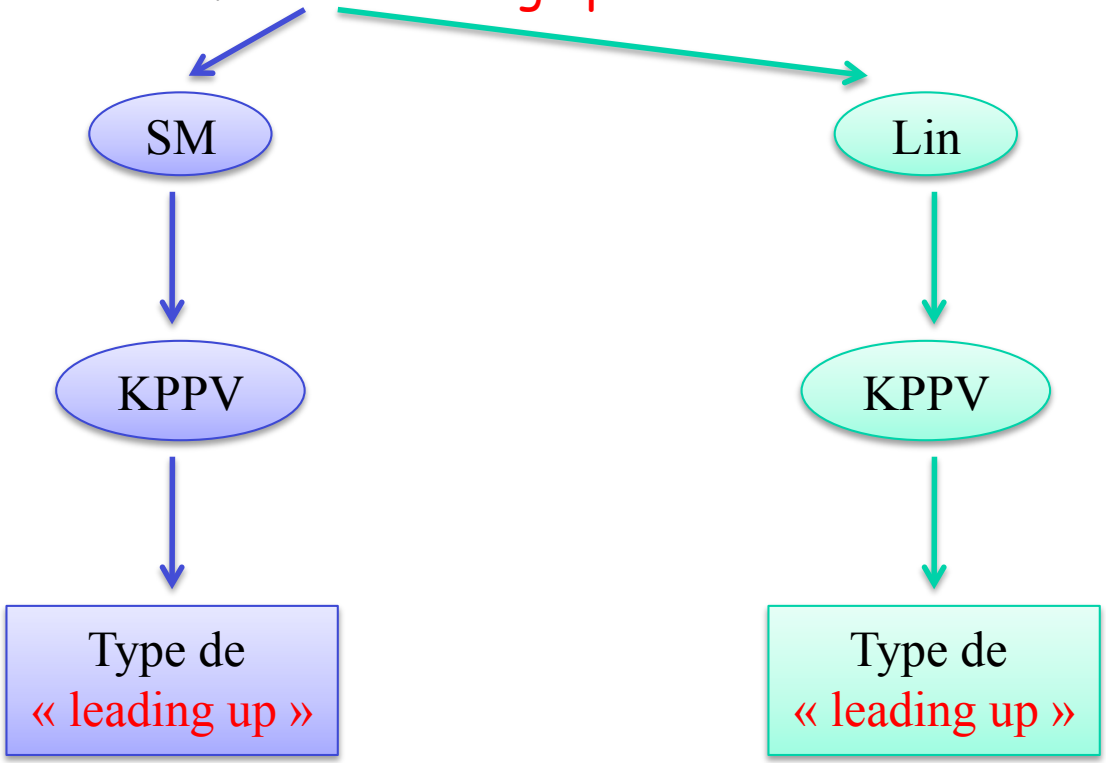
$$Lin(Ch1, Ch2) = \frac{1}{[1 + |tr(Ch1)| + |tr(Ch2)| - 2 \times |tr(Ch1) \cap tr(Ch2)|]}$$

  - Exemple

$$Lin(leading\ up, leaning\ on) = \frac{1}{[(1+8+8)-(2\times3)]} = 0,09$$

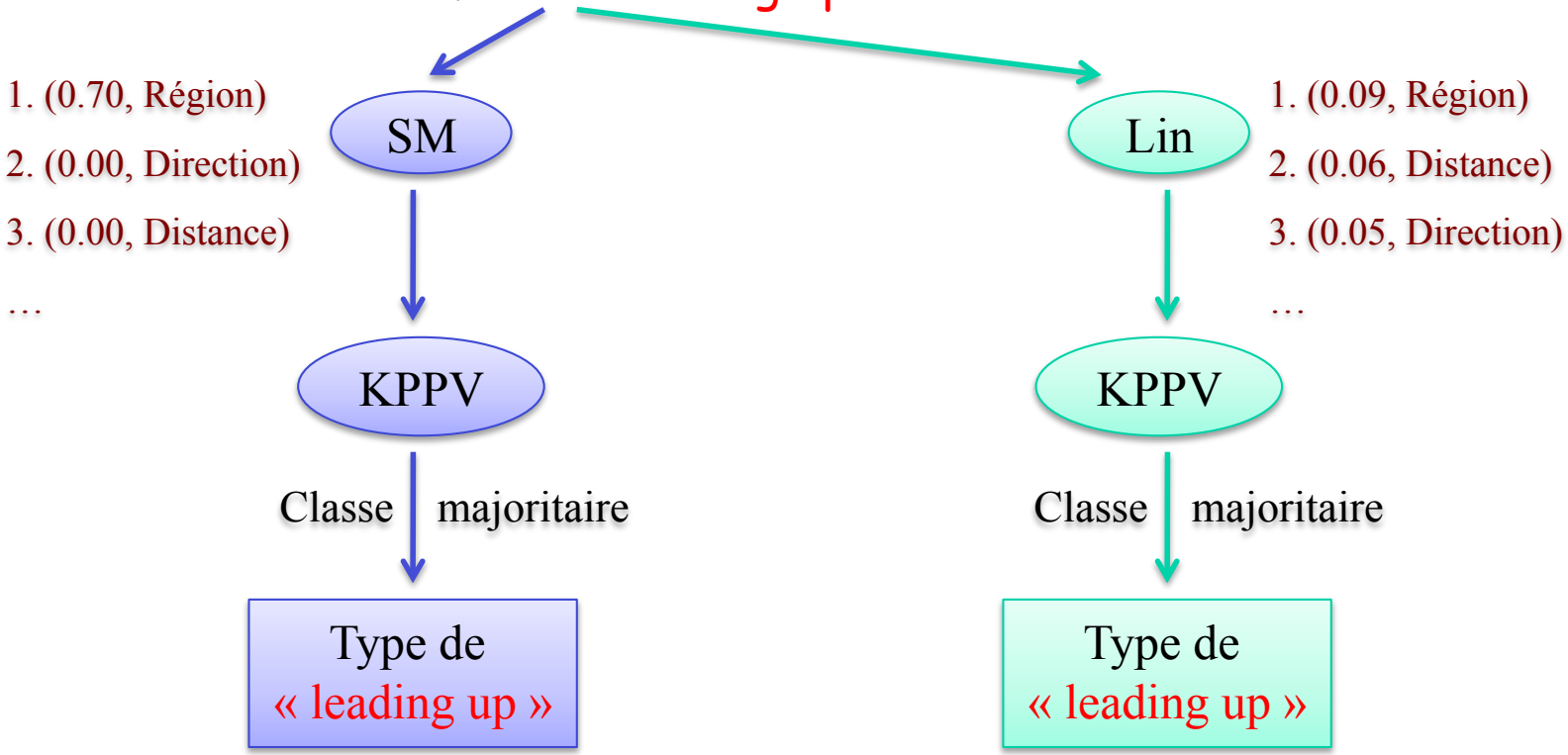
Par comparaison de chaînes de caractères

- Exemple calcule la similarité de relation candidate « leading up » :
  1. « leaning on » - « leading up »
  2. « at the back » - « leading up »
  3. « close to » - « leading up »
  4. … - « leading up »



Par comparaison de chaînes de caractères

- Exemple calcule la similarité de relation candidate « leading up » :
  1. « leaning on » - « leading up »
  2. « at the back » - « leading up »
  3. « close to » - « leading up »
  4. … - « leading up »



Par comparaison de chaînes de caractères

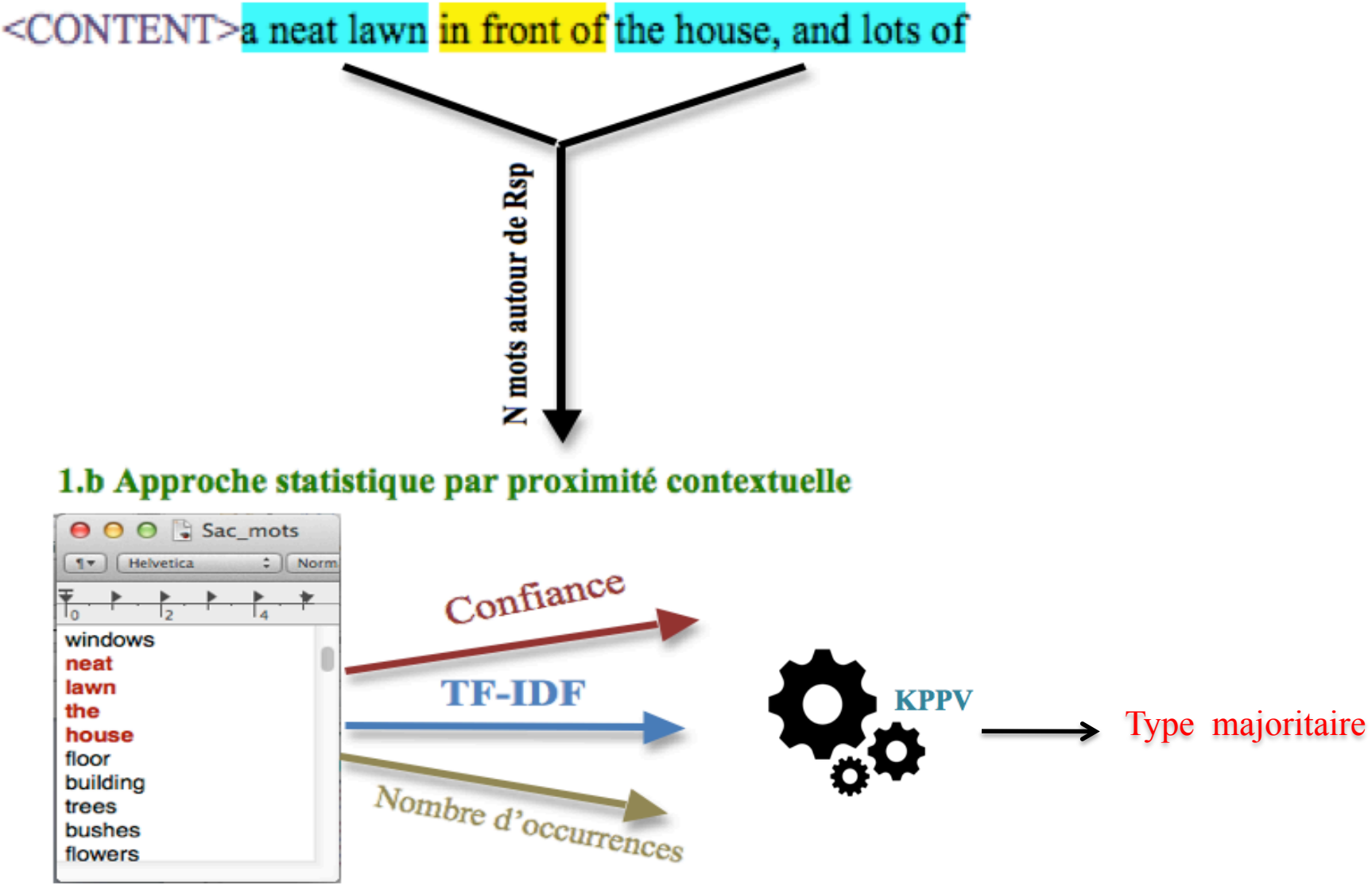
- Pour les relations composées de deux mots dont le deuxième mot est une relation spatiale **next to, standing in, sitting at**…
- L'hypothèse :

  ces relations sont du même type que celui des relations **to, in, at**...

Relation	Type prédit
Leading up	region
In front	direction
Leaning on	region
On the left	direction
Along the left side of	direction
Near	direction
At each side	region
To the left and the right	region
At the back	direction

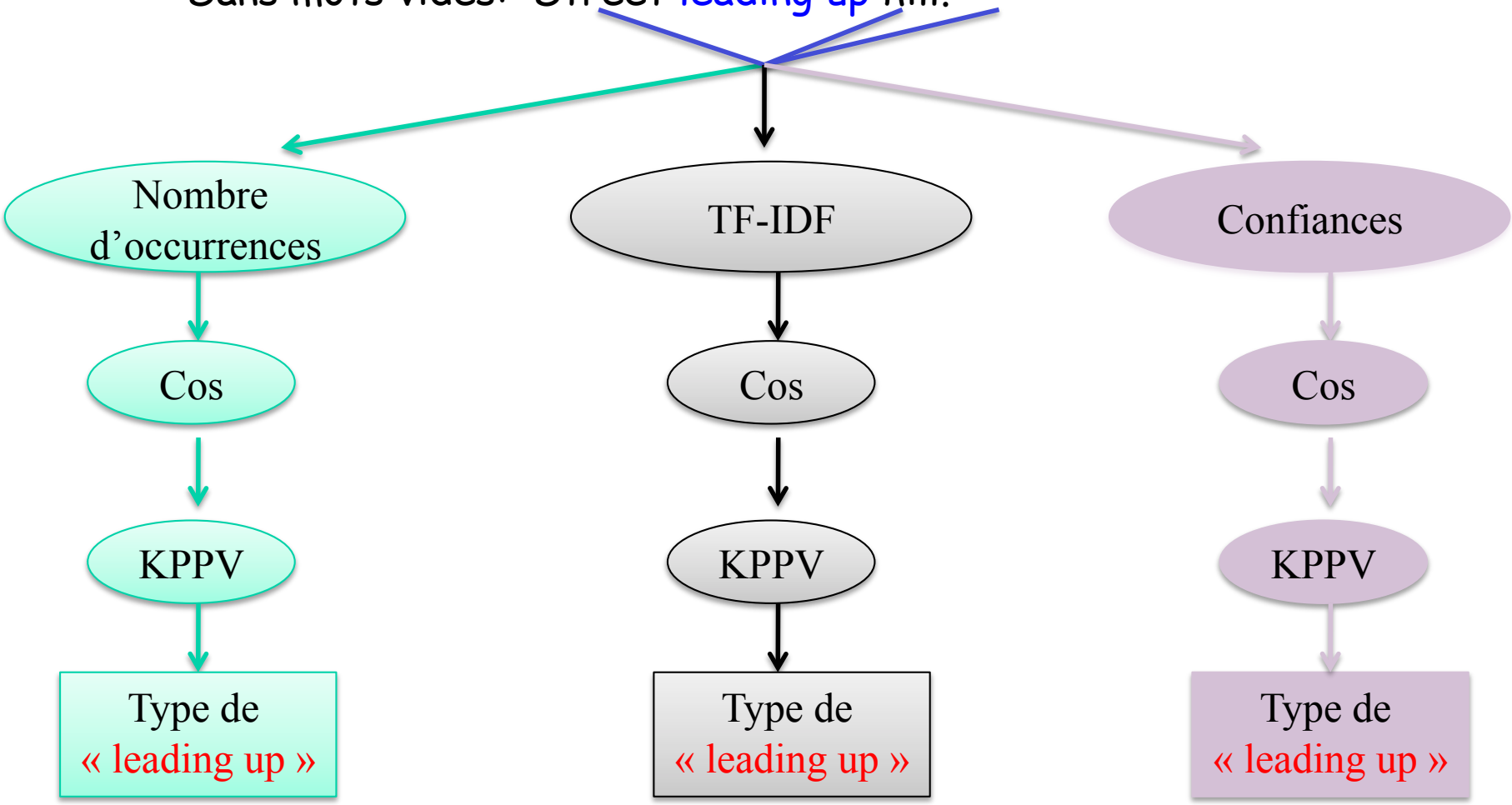
Par proximité contextuelle

- Prédiction du type de relation par proximité contextuelle :



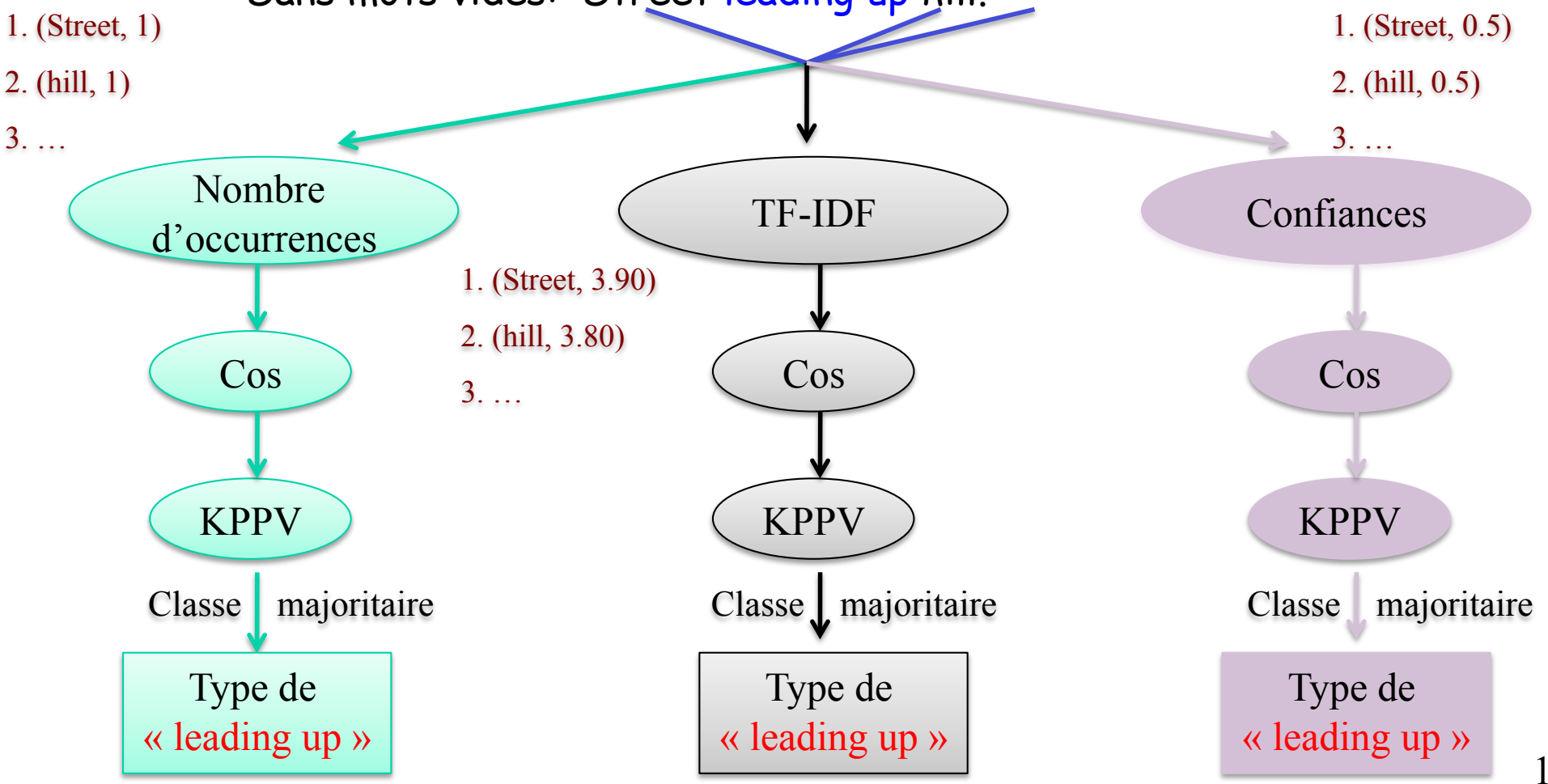
Par proximité contextuelle

- Exemple :
  - Street leading up the hill.
  - Sans mots vides: Street leading up hill.



Par proximité contextuelle

- Exemple :
  - Street leading up the hill.
  - Sans mots vides: Street leading up hill.



Par proximité contextuelle

Relation	Type prédit
Leading up	region
In front	region
Leaning on	region
On the left	direction
Along the left side of	region
Near	region
At each side	region
To the left and the right	direction
At the back	direction

Combinaison

- Exemple

Tableau 1

Relation	Type predicted
In front	Direction
Along the left side of	Direction
Sitting in	Region
To the left and the right	Region
Near	Region

Tableau 2

Relation	Type predicted
In front	Region
Along the left side of	Region
Sitting in	Direction
To the left and the right	Direction
Near	Region

Combinaison

Algorithme 1 Combinaison

ENTRÉES: $taille_{terme}, type_{SM}, type_{Cos}$
SORTIES: $type_{Predit}$
si $(taille_{terme} > 4)$ **alors**
$type_{Predit} \leftarrow type_{Cos}$
sinon
$type_{Predit} \leftarrow type_{SM}$
finsi

Combinaison

- Exemple :

Tableau 1

Relation	Type prédit
In front	direction
Along the left side of	direction
Sitting in	region
To the left and the right	region
Near	region

Tableau 2

Relation	Type prédit
In front	region
Along the left side of	region
Sitting in	direction
To the left and the right	direction
Near	region

Relation	Type prédit
In front	direction
Along the left side of	region
Sitting in	region
To the left and the right	direction
Near	region

Expérimentations

- Corpus SPRL (Spatial Role Labeling, SemEval 2012) : composé de 1213 phrases annotées en anglais.
- Corpus contient 93 relations spatiales.
- Variation les paramètres : ***K*** pour ***KPPV*** et ***N*** pour ***N termes autour de la relation***.
- Mesures de performance ***Précision***, ***Rappel***, ***F-mesure*** et ***l'Exactitude (Accuracy)***.

Par comparaison de chaînes de caractères

- Application d'un processus de validation croisée.
- Corpus est divisé en 3 partitions et chaque partition contient 31 relations (18 régions, 10 directions, 3 distances).

En terme d'Exactitude (Accuracy) :

K	String Matching	Lin
1	**0.82**	0.75
2	0.81	0.75
3	0.79	0.73
4	0.78	0.75
5	0.76	0.69

Par proximité contextuelle

- En terme d'Exactitude, nous avons obtenu les meilleurs résultats avec le monde lexical ***2 termes autour de la relation*** fondé sur ***TF-IDf*** :

K'	N = 1	N = 2	N = 3
1	0.60	0.51	0.53
2	0.58	0.49	0.50
3	0.63	0.66	0.63
4	0.62	0.64	0.61
5	0.64	**0.67**	0.66

Combinaison

En terme d'Exactitude :

Par comparaison de chaines de caractères	Par proximité contextuelle	Exactitude
K = 1		0.82
	K' = 5, N = 2	0.67
K = 1	K' = 5, N = 2	0.84

Conclusion – première contribution

- Une analyse comparative de deux approches et de leur combinaison.
  - adaptation des mesures lexicales selon la spécificité des relations spatiales
  - identification du contexte le plus pertinent à prendre en compte
  - nouvelle approche de combinaison des approches lexicales et contextuelles
- Nos résultats montrent que la combinaison améliore la qualité de l'identification du type des relations spatiales.
- Exploration de nouveaux modes d'hybridation afin de tirer le meilleur parti des deux premières approches.

Perspectives 1/2

1. Etude de la généricité de l'approche selon les langues (en particulier, le français)
   - exploitation d'un corpus contenant 7000 documents de Midi Libre (Roche et al. 2014) - projet Animitex2.

Perspectives 2/2

1. Adaptation de l'approche selon le type de textes traités (presse vs. réseaux sociaux/SMS)
   - exploitation d'un corpus contenant plus de 88.000 SMS (Panckhurst et al. 2014) - projet sud4sciences.
   - Identification des indicateurs spatiaux sans prendre en considération le langage de SMS.
   - Identification et la détection des entités spatiales qui sont exprimées de manière différente (par exemple pour désigner Montpellier : montpellier, montpeul, Montp, monpel, montpelier, etc.).
   - Identification des relations spatiales qui sont écrit de manière différents (par exemple (dans -> ds, vers -> vR, à -> a, etc.).

Bibliographie

- Grefenstette G. (1994). Explorations in automatic thesaurus discovery. Norwell, MA, USA, Kluwer Academic Publishers.
- JONES, C.B., PURVES, R.S., CLOUGH, P.D. and JOHO, H., in press, Modelling Vague Places with Knowledge from the Web. International Journal of Geographical Information Science.
- Nadeau, D. et S. Sekine (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes 30 (1), 3–26.
- Sallaberry, C., M. Baziz, J. Lesbegueries, et M. Gaio (2007). Une approche d'extraction et de recherche d'information spatiale dans les documents textuels - évaluation. In COnférence en Recherche d'Infomations et Applications - CORIA 2007, pp. 53–64.
- Sanderson, M. and Kohler, J. (2004). Analyzing geographic queries. In Proceedings of the Workshop on Geographic Information Retrieval, SIGIR.

Bibliographie

- Weissenbacher D., Nazarenko A. (2007). Identifier les pronoms anaphoriques et trouver leurs antécédents: l'intérêt de la classification bayésienne. In Proceedings of taln, p. 145-155.
- Roche M., al ANalyse d'IMages fondée sur des Informations TEXtuelles Projet CNRS MASTODONS Masses de Données Scientifiques.
- Panckhurst R, Détrie C, Lopez C, Moïse C, Roche M, Verine B (Praxiling, Lirmm, Lidilem, Tetis, Viseo) 88milSMS. A corpus of authentic text messages in French.
- Kordjamshidi P., Steven B., Moens M-F. Semeval-2012 task 3: Spatial role labeling. In Sem 2012: ([semeval] 2012), montreal- canada, p. 365–373. ACL.

Merci de votre attention