Deux approches pour la détection d'itinéraires dans du texte

Pierre LOUSTAU, Julien LESBEGUERRIES

Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour
FRANCE

Atelier RTE 2007 - Grenoble
2 juillet 2007

Plan

1. Problématique
2. Extraction d'Entités Spatiales
   - Modèle d'extraction des ES
   - Exemples
3. Approche géographique
   - Motifs spatiaux
   - Caractéristiques
   - Classification
4. Approche linguistico-géographique
   - Extraction des déplacements de l'auteur
   - Exemples
   - Reconstruction de l'itinéraire
5. Conclusion : complémentarité des deux approches
6. Bibliographie

1. Problématique

Problématique 1/3

Un objectif commun : revitaliser un corpus territorialisé

- Corpus composé de documents anciens (XVIIIème, XIXème et XXème siècle)
- Récits de voyages dans les Pyrénées
- Exploitation du contenu nécessite modélisation du contenu
- Modélisation du monde impossible, restriction sur l'aspect spatial et temporel des documents

Des usages différents

- Navigation, recherche, découverte : Recherche d'Information
- Ressources pour la construction d'activités pédagogiques : conception d'applications

Problématique 1/3

Un objectif commun : revitaliser un corpus territorialisé

- Corpus composé de documents anciens (XVIIIème, XIXème et XXème siècle)
- Récits de voyages dans les Pyrénées
- Exploitation du contenu nécessite modélisation du contenu
- Modélisation du monde impossible, restriction sur l'aspect spatial et temporel des documents

Des usages différents

- Navigation, recherche, découverte : Recherche d'Information
- Ressources pour la construction d'activités pédagogiques : conception d'applications

Problématique 2/3

Recherche d'Information (RI)

- Construire des indexes pour permettre un accès fin et rapide aux contenus d'une base documentaire.
- Approche sur notre corpus : constitution d'indexes spatiaux basés sur le géo-référencement des évocations de lieux extraites du corpus.
- Appariement en comparant le géoréférencement de la requête et des indexes.
- S'abastraire de la manière dont est évoqué un territoire pour le retrouver

Problématique 3/3

Conception d'applications pédagogiques

- Nombreuses applications pédagogiques basées sur des documents "territorialisés" (ex. des itinéraires)
- Fort intérêt de la communauté éducative
- Faire intervenir les contenus dans la phase de conception [NLG+06]

2. Extraction d'Entités Spatiales (ES)

Extraction d'Entités Spatiales

Modèle d'entités PIV (MSU) [LL06]

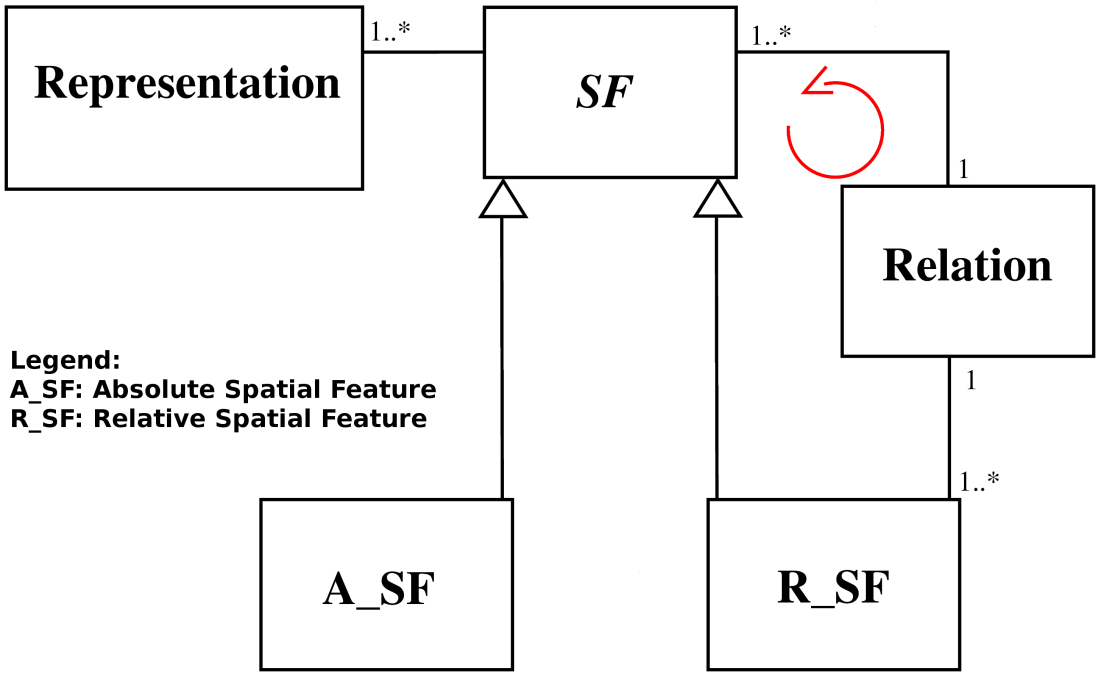


FIG.: Définition d'une entité spatiale dans PIV (MSU).

Extraction d'Entités Spatiales

Modèle d'entités PIV (MSU) [LL06]

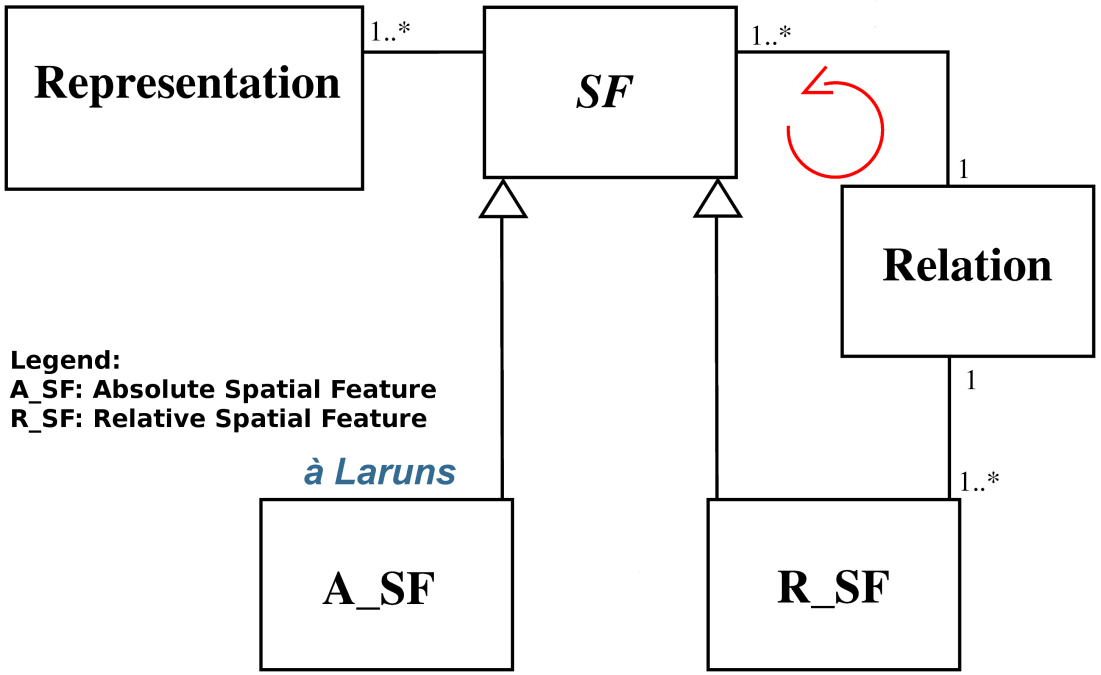


FIG.: Définition d'une entité spatiale dans PIV (MSU).

Extraction d'Entités Spatiales

Modèle d'entités PIV (MSU) [LL06]

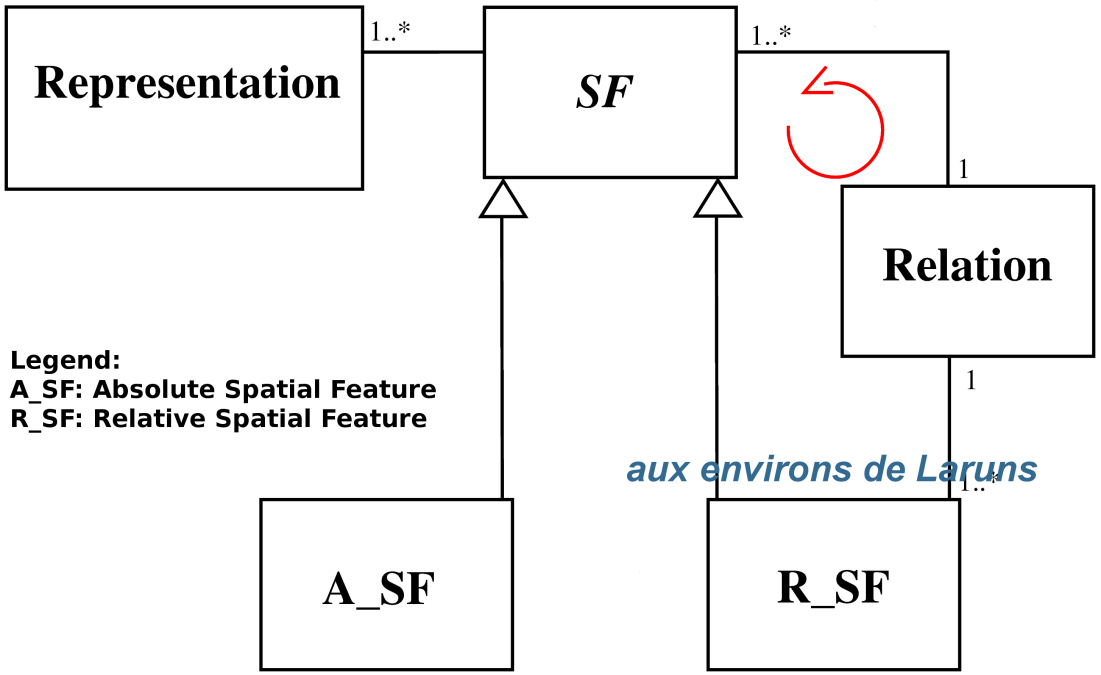


FIG.: Définition d'une entité spatiale dans PIV (MSU).

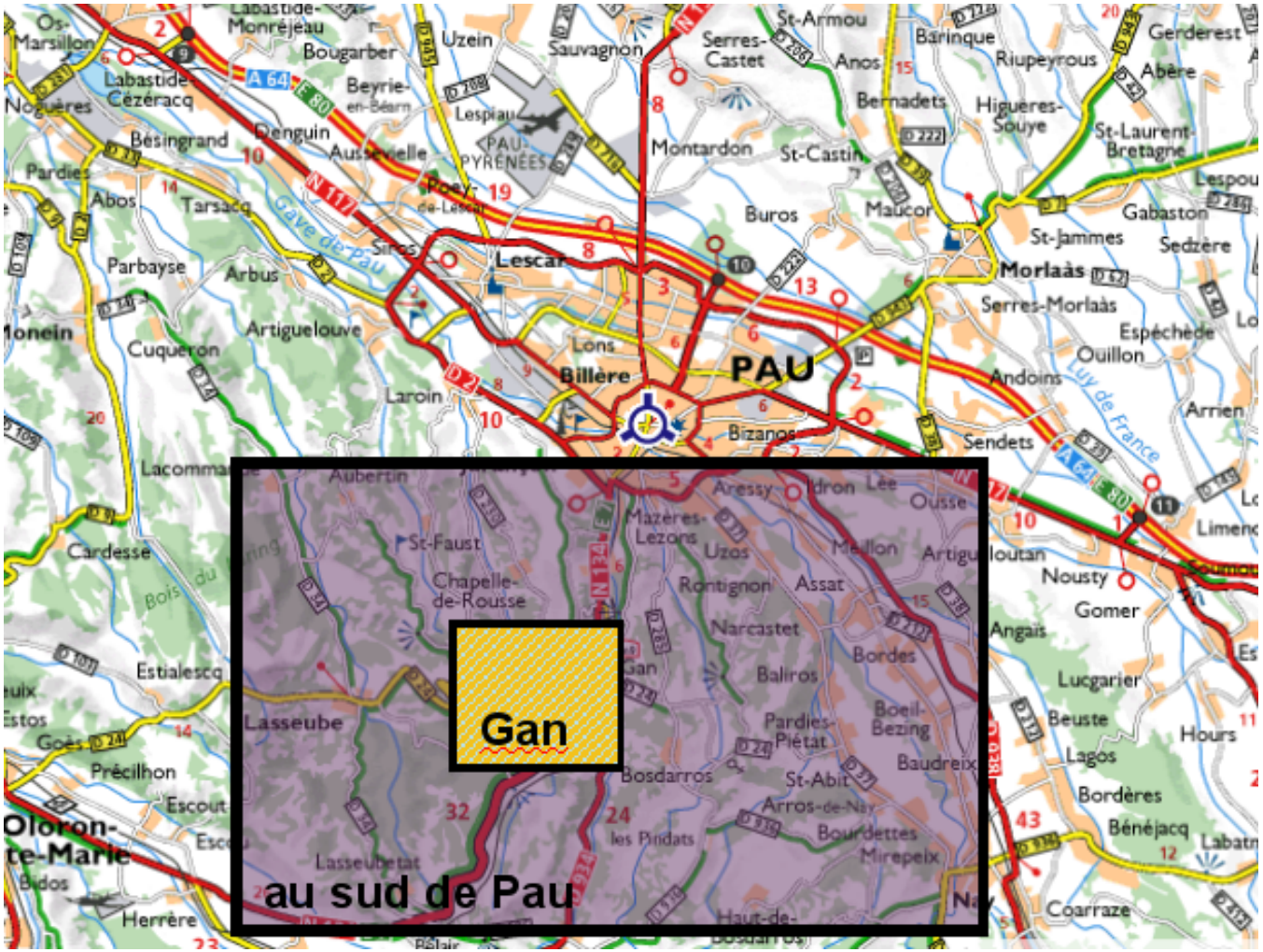
Extraction des ES

[§1]JOURNAL DE JAMES DAVID FORBES [/§]

[§2]5 juillet. Dimanche. Je me suis baigné. Je suis parti à 11 hl /4 avec un groupe de gens pour le phare de Cordouan dans une pinasse. Vent contraire. En louvoyant, en ramant et avec l'aide de la marée, nous arrivâmes à Cordouan vers trois heures; cependant la journée fut belle et la compagnie plutôt agréable. J'ai vu toutes les parties du phare; la base fut construite au temps de Louis XIV, la partie supérieure en 1789. Le dispositif lenticulaire actuel sert depuis douze ans. Le gardien se plaint seulement de la lampe qui lui cause des ennuis. Nous avons quitté Cordouan vers 5 heures. Presque calme; nous avons mis deux heures pour arriver à Royan. J'y ai passé la nuit. 6 juillet. Je suis parti par le bateau à vapeur à 6 heures du matin et j'ai atteint Bordeaux à midi après une brève traversée. Après-midi pluvieuse. J'ai dîné avec M. Guestierl. J'ai écrit à Elisa.[/§]

[§3]7 juillet. J'ai quitté Bordeaux à 7 heures en diligence pour Pau. J'ai été agréablement surpris par la beauté de la campagne. Jusqu'à Langon, nous avons longé la Garonne, traversant un véritable jardin rempli de vignes et d'arbres fruitiers. La campagne est par la suite toujours digne d'intérêt et il n'y a qu'une ou deux étapes assez tristes au moment où on traverse une partie des Landes, ce qui permet de se faire une idée de ce qu'est cette région; elle n'est pas aussi déserte que je le pensais. Elle est en grande partie couverte d'arbres dont on enlève l'écorce à certains endroits pour obtenir de la résine et de la térébenthine. Le sol est marécageux en hiver et en conséquence Napoléon construisit une chaussée de bois, qui constitue aujourd'hui la plus mauvaise partie de la route et est vraiment en très mauvais état, car elle est réduite en une poudre impalpable. A Roquefort, la campagne redevient belle et le reste jusqu'à Pau, du moins je le crois, car j'ai dormi pendant une partie du trajet. Mont-de-Marsan, où nous avons soupé, est une ville importante dont les alentours semblent agréables. Lorsqu'on approche de Pau, la campagne, riche et boisée, est des plus

Extraction des ES, exemple proto PIV



Extraction des ES, bilan

Hypothèses, ressources nécessaires

- Le concept cible/site [Van86, Bor98]
- Implémentation sous forme de grammaire DCG des relations syntaxiques modifiant la localisation (ex. "Au sud de")
- Mise en correspondance avec des fonctions géographiques des relations syntaxiques extraites
- Ressources géographiques pour géo-référencement

Résultats obtenus

- Prototype PIV [LL06]

Extraction des ES, bilan

Hypothèses, ressources nécessaires

- Le concept cible/site [Van86, Bor98]
- Implémentation sous forme de grammaire DCG des relations syntaxiques modifiant la localisation (ex. "Au sud de")
- Mise en correspondance avec des fonctions géographiques des relations syntaxiques extraites
- Ressources géographiques pour géo-référencement

Résultats obtenus

- Prototype PIV [LL06]

3. Approche géographique

Approche géographique

Principes généraux

- Extraction des ES
- Analyse de la disposition de sacs d'ES, indépendamment de toutes autres considérations linguistiques
- Classification

Trois motifs retenus

- Itinéraire
- Description locale
- Comparaison de lieux

Approche géographique

Principes généraux

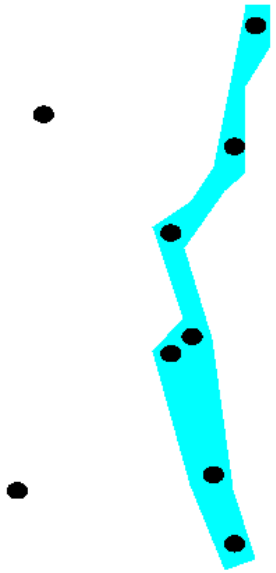
- Extraction des ES
- Analyse de la disposition de sacs d'ES, indépendamment de toutes autres considérations linguistiques
- Classification

Trois motifs retenus

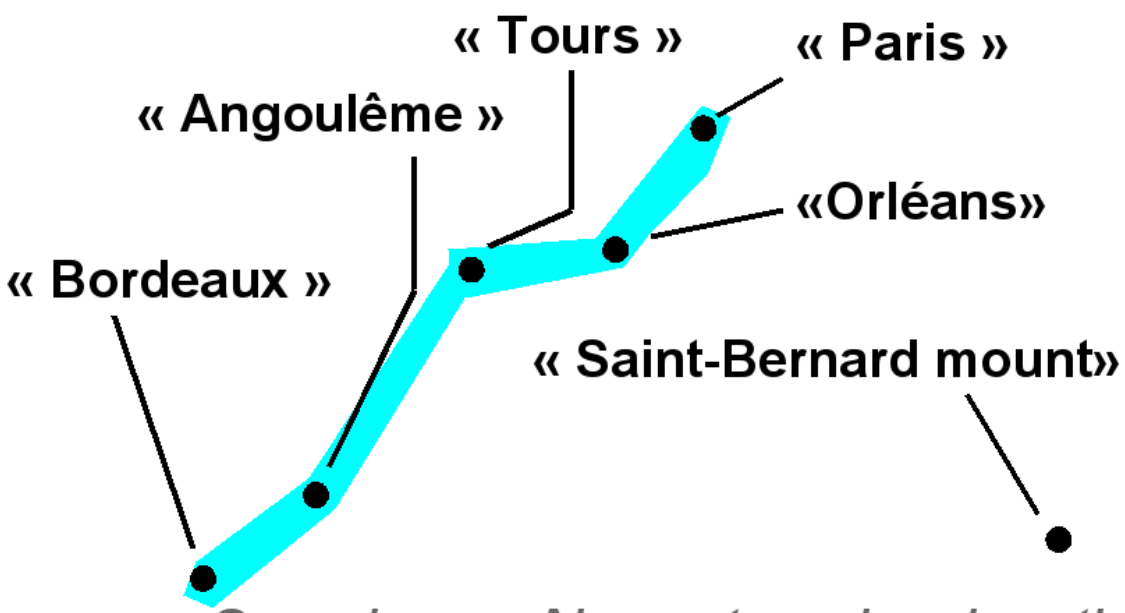
- Itinéraire
- Description locale
- Comparaison de lieux

Approche géographique

Motif « itinéraire »



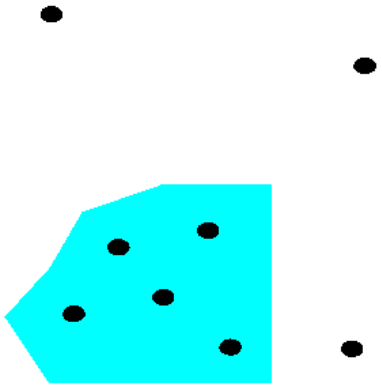
Itinerary pattern



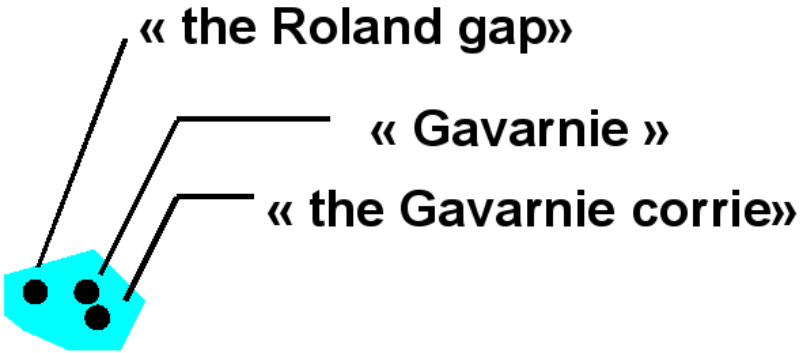
Sample: « New travels in the Pyrenees » p 34 - PyréGraph

Approche géographique

Motif « description locale »



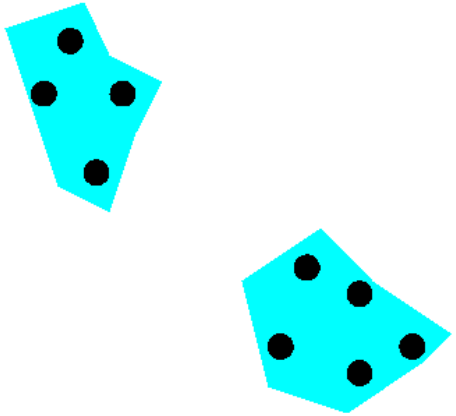
Point of view pattern



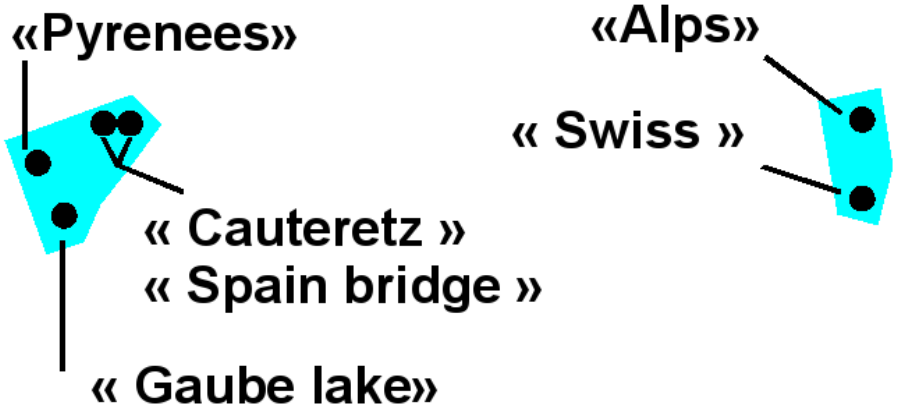
Sample: «Travel in Pyrenees» p 43 - Pimientos.

Approche géographique

motif « Comparaison de lieux »



Area comparison pattern



Sample: « New travels in the Pyrenees » p 46 - PyréGraph

Approche géographique

Caractéristiques

- la **dispersion** ; les ESs peuvent être contenues les unes dans les autres, être proches ou dispersées dans un espace géographique,
- l'**ordonnancement** ; les ESs connexes dans le texte dessinent un chemin ou non dans un espace géographique,
- la **saillance** ; les ESs connexes dans le texte forment des angles plats, obtus ou aigus dans un espace géographique.
- Plus des propriétés intrinsèques aux ESs comme leur échelle par exemple, sont utilisables pour le calcul des caractéristiques.

Illustrations

Dispersion

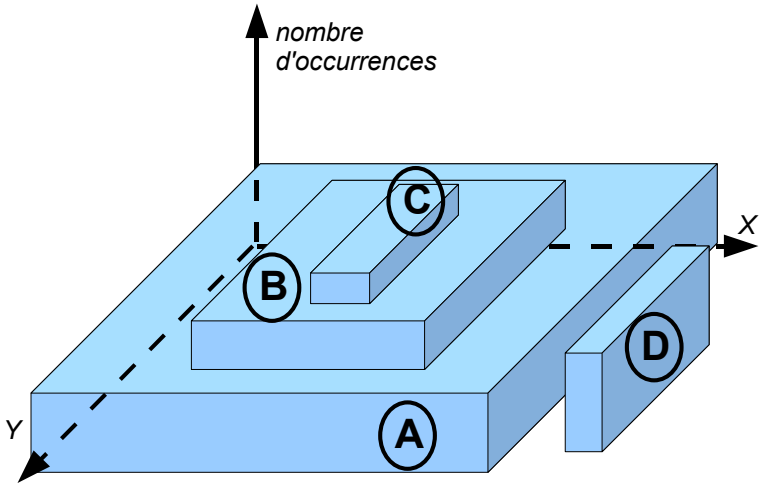


FIG.: A_{prev} : exemple de calcul. Dans une unité de texte où A, B, C et D sont des ESs, (A) intersecte avec (B) et (C), mais pas avec (D).

Illustrations

Ordonnancement

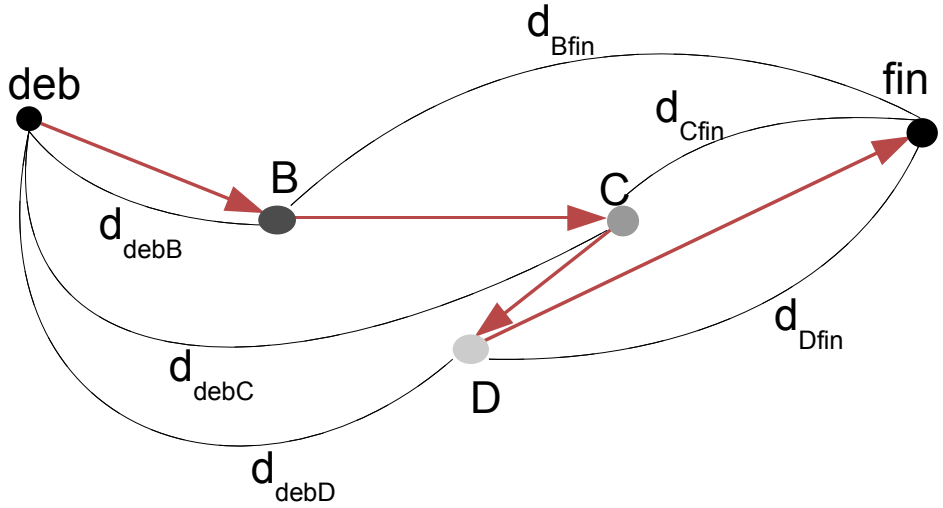


FIG.: O_{SFs} : exemple de calcul d'ordonnancement des points intermédiaires B, C, D.

Illustrations

Saillance

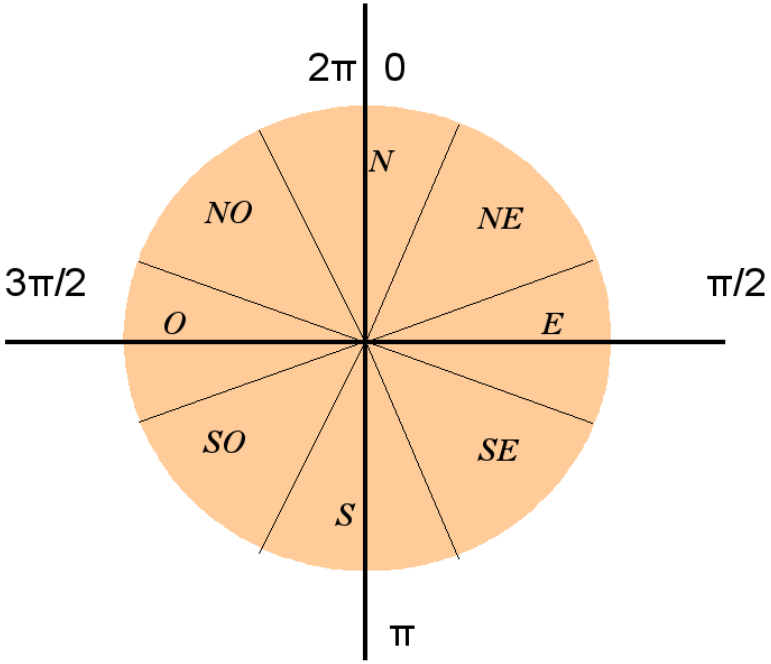


FIG.: Rose des vents, orientation donnée en radians par la fonction SIG "azimuth()"

Classification

Basée sur SVM[Has06, Joa98]

- méthode consistant à associer à un élément $\vec{x}$ d'un ensemble χ, une classe notée y
- Ce couple est nommé « observation ». Une série de n observations S notée : $S = \{(\vec{x}_1, y_1), ..., (\vec{x}_n, y_n)\}$, est utilisée comme base d'apprentissage.
- Construction de règles qui peuvent être utilisées pour classifier un nouvel objet $\vec{x} \in \chi$.
- Pour nous :
  - $\vec{x}$: unité de texte
  - χ : corpus
  - y : {itinéraire, description locale, comparaison de lieux}

4. Approche linguistico-géographique

La modélisation des déplacements 1/3

Principes généraux

- Extraction des ES
- Extraction des déplacements
- Reconstruction de l'itinéraire

Expériences sur le corpus

		Expérience 1			Expérience 2		
Texte	Mots	PP	PP ES	%	PP	PP V	%
J.-D. Forbes	3662	75	63	80%	75	63	80%
A. Lister	3225	40	20	50%	40	35	87,5%
J.-R. Bals	19015	213	170	80%	213	190	89%

Figure 2. *Expériences 1 et 2 : le poids des propositions utilisant les ES (PPES) et des verbes (PPV) dans les propositions évoquant la position (PP)*

La modélisation des déplacements 1/3

Principes généraux

- Extraction des ES
- Extraction des déplacements
- Reconstruction de l'itinéraire

Expériences sur le corpus

		Expérience 1			Expérience 2		
Texte	Mots	PP	PP ES	%	PP	PP V	%
J.-D. Forbes	3662	75	63	80%	75	63	80%
A. Lister	3225	40	20	50%	40	35	87,5%
J.-R. Bals	19015	213	170	80%	213	190	89%

Figure 2. *Expériences 1 et 2 : le poids des propositions utilisant les ES (PPES) et des verbes (PPV) dans les propositions évoquant la position (PP)*

La modélisation des déplacements 2/3

Le triplé (V,P,E)

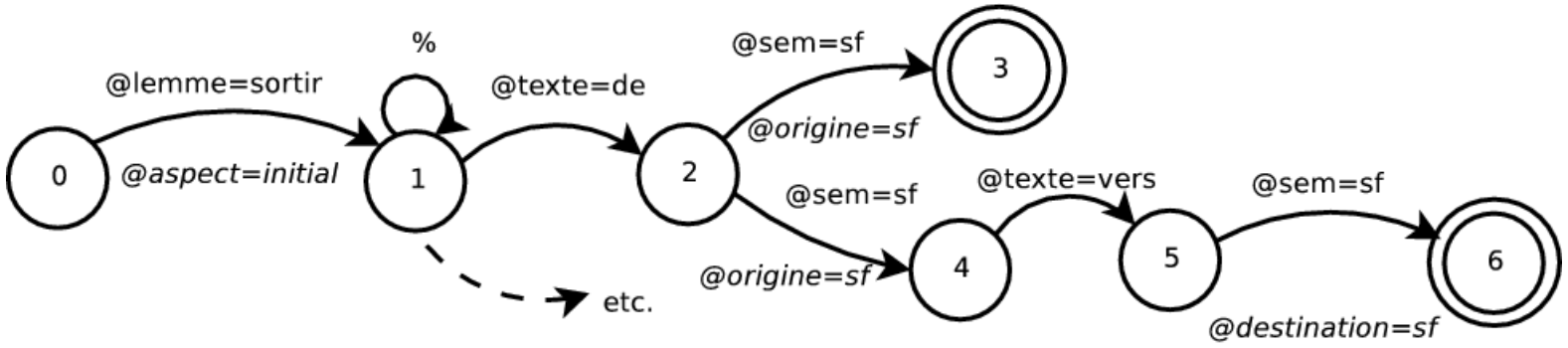
- Verbe + Préposition + Entité : discriminant pour le sens spatial
- verbes de déplacement (reprise des notions de polarité aspectuelle [Boo87, Lau91])
- notions d'origine de destination et de position médiane
- modalité du transport

La modélisation des déplacements 3/3

Transducteurs spatiaux

- On considère que l'évocation d'un déplacement sous forme verbale s'apparente à un langage régulier
- Modélisation du sens spatial des verbes sous forme de transducteurs

Exemple : le transducteur spatial du verbe sortir

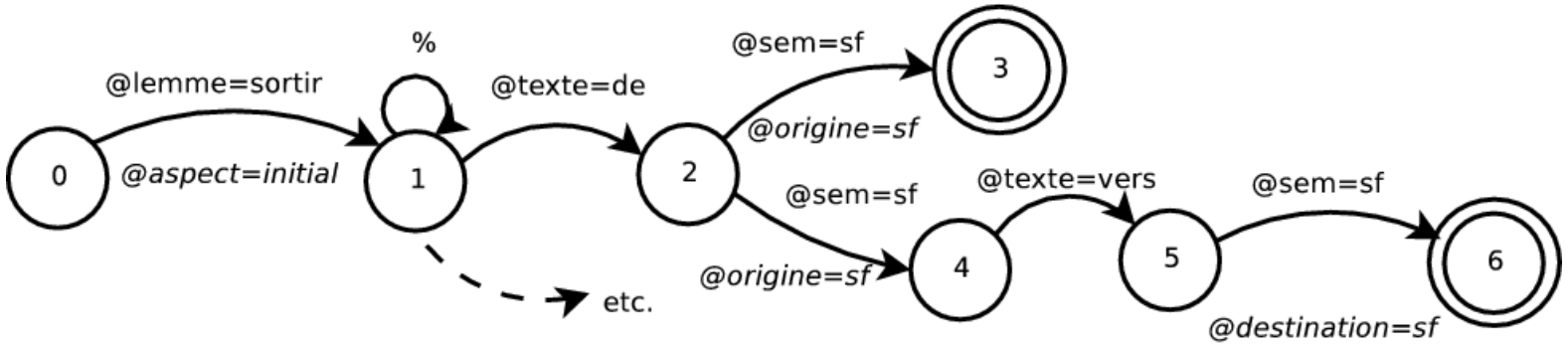


La modélisation des déplacements 3/3

Transducteurs spatiaux

- On considère que l'évocation d'un déplacement sous forme verbale s'apparente à un langage régulier
- Modélisation du sens spatial des verbes sous forme de transducteurs

Exemple : le transducteur spatial du verbe sortir



Extraction des déplacements

[§1]JOURNAL DE JAMES DAVID FORBES [/§]

[§2]5 juillet. Dimanche. Je me suis baigné. Je suis parti à 11 hl /4 avec un groupe de gens pour le phare de Cordouan dans une pinasse. Vent contraire. En louvoyant, en ramant et avec l'aide de la marée, nous arrivâmes à Cordouan vers trois heures; cependant la journée fut belle et la compagnie plutôt agréable. J'ai vu toutes les parties du phare; la base fut construite au temps de Louis XIV, la partie supérieure en 1789. Le dispositif lenticulaire actuel sert depuis douze ans. Le gardien se plaint seulement de la lampe qui lui cause des ennuis. Nous avons quitté Cordouan vers 5 heures. Presque calme; nous avons mis deux heures pour arriver à Royan. J'y ai passé la nuit. 6 juillet. Je suis parti par le bateau à vapeur à 6 heures du matin et j'ai atteint Bordeaux à midi après une brève traversée. Après-midi pluvieuse. J'ai dîné avec M. Guestierl. J'ai écrit à Elisa.[/§]

[§3]7 juillet. J'ai quitté Bordeaux à 7 heures en diligence pour Pau. J'ai été agréablement surpris par la beauté de la campagne. Jusqu'à Langon, nous avons longé la Garonne, traversant un véritable jardin rempli de vignes et d'arbres fruitiers. La campagne est par la suite toujours digne d'intérêt et il n'y a qu'une ou deux étapes assez tristes au moment où on traverse une partie des Landes, ce qui permet de se faire une idée de ce qu'est cette région; elle n'est pas aussi déserte que je le pensais. Elle est en grande partie couverte d'arbres dont on enlève l'écorce à certains endroits pour obtenir de la résine et de la térébenthine. Le sol est marécageux en hiver et en conséquence Napoléon construisit une chaussée de bois, qui constitue aujourd'hui la plus mauvaise partie de la route et est vraiment en très mauvais état, car elle est réduite en une poudre impalpable. A Roquefort, la campagne redevient belle et le reste jusqu'à Pau, du moins je le crois, car j'ai dormi pendant une partie du trajet. Mont-de-Marsan, où nous avons soupé, est une ville importante dont les alentours

Extraction des déplacements

[§1]JOURNAL DE JAMES DAVID FORBES [/§]

[§2]5 juillet. Dimanche. Je me suis baigné. Je suis parti à 11 hl /4 avec un groupe de gens pour le phare de Cordouan dans une pinasse. Vent contraire. En louvoyant, en ramant et avec l'aide de la marée, nous arrivâmes à Cordouan vers trois heures; cependant la journée fut belle et la compagnie plutôt agréable. J'ai vu toutes les parties du phare; la base fut construite au temps de Louis XIV, la partie supérieure en 1789. Le dispositif lenticulaire actuel sert depuis douze ans. Le gardien se plaint seulement de la lampe qui lui cause des ennuis. Nous avons quitté Cordouan vers 5 heures. Presque calme; nous avons mis deux heures pour arriver à Royan. J'y ai passé la nuit. 6 juillet. Je suis parti par le bateau à vapeur à 6 heures du matin et j'ai atteint Bordeaux à midi après une brève traversée. Après-midi pluvieuse. J'ai dîné avec M. Guestierl. J'ai écrit à Elisa.[/§]

[§3]7 juillet. J'ai quitté Bordeaux à 7 heures en diligence pour Pau. J'ai été agréablement surpris par la beauté de la campagne. Jusqu'à Langon, nous avons longé la Garonne, traversant un véritable jardin rempli de vignes et d'arbres fruitiers. La campagne est par la suite toujours digne d'intérêt et il n'y a qu'une ou deux étapes assez tristes au moment où on traverse une partie des Landes, ce qui permet de se faire une idée de ce qu'est cette région; elle n'est pas aussi déserte que je le pensais. Elle est en grande partie couverte d'arbres dont on enlève l'écorce à certains endroits pour obtenir de la résine et de la térébenthine. Le sol est marécageux en hiver et en conséquence Napoléon construisit une chaussée de bois, qui constitue aujourd'hui la plus mauvaise partie de la route et est vraiment en très mauvais état, car elle est réduite en une poudre impalpable. A Roquefort, la campagne redevient belle et le reste jusqu'à Pau, du moins je le crois, car j'ai dormi pendant une partie du trajet. Mont-de-Marsan, où nous avons soupé, est une ville importante dont les alentours

Extraction des déplacements

ite au temps de Louis XIV, la partie supérieure en 17
sert depuis douze ans. Le gardien se plaint seulement
Nous avons quitté Cordouan vers 5 heures. Presque
pour arriver à
ures du matin
se. J'ai dîné ave

llet. J'ai quitté
par la beaut
ant un véritabl
ujours digne d
e une partie de
st pas aussi d
. enlève l'écorce
marécageux er
stitue aujourd'

deplacement_dcg/2 [X]

- *deplacement:*
  - *polarite:* *initial*
  - *origine:*
    - *es_a:*
      - *type:* *inconnu*
      - *label:* *Cordouan*
  - *destination:* *nc*
  - *actuel:* *nc*

Extraction des déplacements

[§1]JOURNAL DE JAMES DAVID FORBES [/§]

[§2]5 juillet. Dimanche. Je me suis baigné. Je suis parti à 11 hl /4 avec un groupe de gens pour le phare de Cordouan dans une pinasse. Vent contraire. En louvoyant, en ramant et avec l'aide de la marée, nous arrivâmes à Cordouan vers trois heures; cependant la journée fut belle et la compagnie plutôt agréable. J'ai vu toutes les parties du phare; la base fut construite au temps de Louis XIV, la partie supérieure en 1789. Le dispositif lenticulaire actuel sert depuis douze ans. Le gardien se plaint seulement de la lampe qui lui cause des ennuis. Nous avons quitté Cordouan vers 5 heures. Presque calme; nous avons mis deux heures pour arriver à Royan. J'y ai passé la nuit. 6 juillet. Je suis parti par le bateau à vapeur à 6 heures du matin et j'ai atteint Bordeaux à midi après une brève traversée. Après-midi pluvieuse. J'ai dîné avec M. Guestierl. J'ai écrit à Elisa.[/§]

[§3]7 juillet. J'ai quitté Bordeaux à 7 heures en diligence pour Pau. J'ai été agréablement surpris par la beauté de la campagne. Jusqu'à Langon, nous avons longé la Garonne, traversant un véritable jardin rempli de vignes et d'arbres fruitiers. La campagne est par la suite toujours digne d'intérêt et il n'y a qu'une ou deux étapes assez tristes au moment où on traverse une partie des Landes, ce qui permet de se faire une idée de ce qu'est cette région; elle n'est pas aussi déserte que je le pensais. Elle est en grande partie couverte d'arbres dont on enlève l'écorce à certains endroits pour obtenir de la résine et de la térébenthine. Le sol est marécageux en hiver et en conséquence Napoléon construisit une chaussée de bois, qui constitue aujourd'hui la plus mauvaise partie de la route et est vraiment en très mauvais état, car elle est réduite en une poudre impalpable. A Roquefort, la campagne redevient belle et le reste jusqu'à Pau, du moins je le crois, car j'ai dormi pendant une partie du trajet. Mont-de-Marsan, où nous avons soupé, est une ville importante dont les alentours

Extraction des déplacements

. J'ai quitté Bordeaux à 7 heures en diligence pour Pau.

deplacement_dcg/4 [X]

- *deplacement:*
  - *polarite:* *initial*
  - *origine:*
    - *es_a:*
      - *type:* *inconnu*
      - *label:* *Bordeaux*
  - *destination:*
    - *es_a:*
      - *type:* *inconnu*
      - *label:* *Pau*
  - *actuel:* *nc*

Extraction des déplacements

Résultats obtenus

		Evaluation 1		Evaluation 2	
Texte	Auto.	Manu. quelconque	%	Manu. avec (V,P,E)	%
J.-D. Forbes	44	75	58,6%	64	68,7%
A. Lister	22	40	55%	25	88%
J.-R. Bals	46	91	50,5%	58	79%

Figure 7. *Evaluation 1 par rapport au marquage manuel de tous les déplacements, évaluation 2 par rapport aux déplacements évoqués avec le triplé (V,P,E)*

Des déplacements à l'itinéraire

Méthode proposée

- une hypothèse : celle de la chronologie de l'évocation des déplacements
- des données factuelles : couches SIG, Gazeteers, base toponymiques, etc.
- raisonnements
  - *bas niveau* : offerts par les outils traitant l'info géo
  - *haut niveau* : contraintes (règles de sens commun)

Des déplacements à l'itinéraire

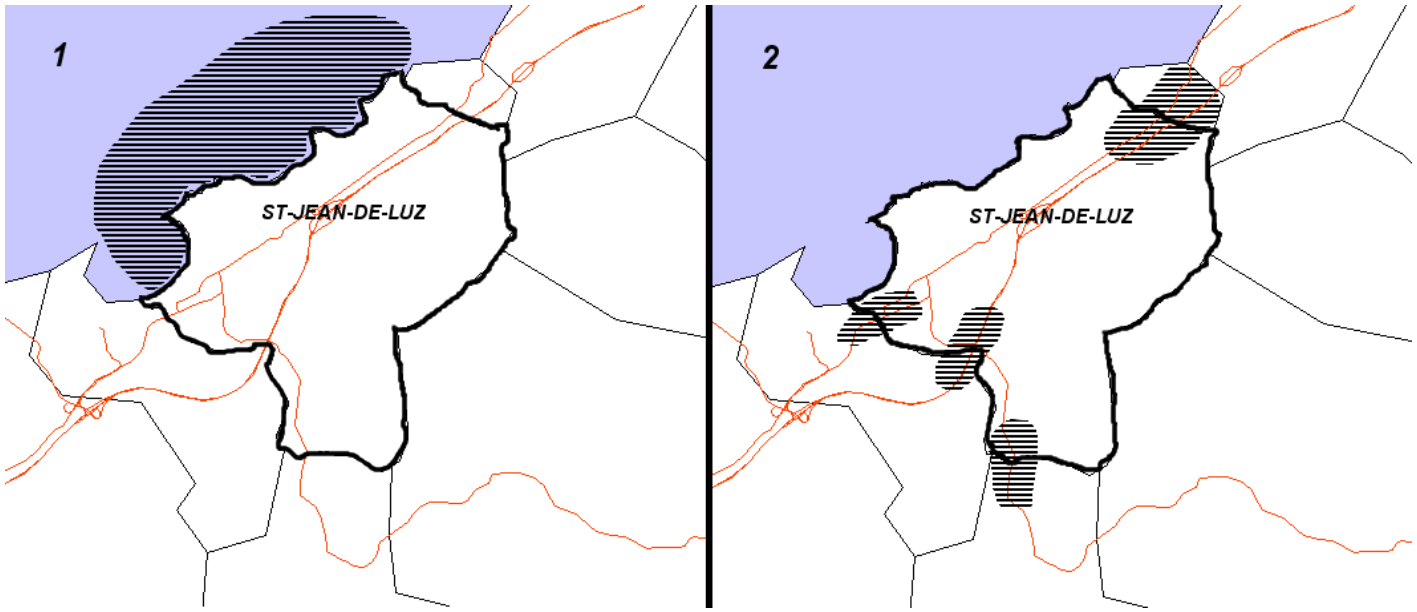
Les contraintes

- règles concernant la polarité aspectuelle et les notions d'origine, de destination (ex : *quitter Bordeaux en diligeance pour Pau*)
- règles concernant les modalités de transport et les zones probables de localisation (voiture sur route, bateau sur eau, etc.)
- étendue des zones selon vitesse du déplacement

Illustration

Illustration

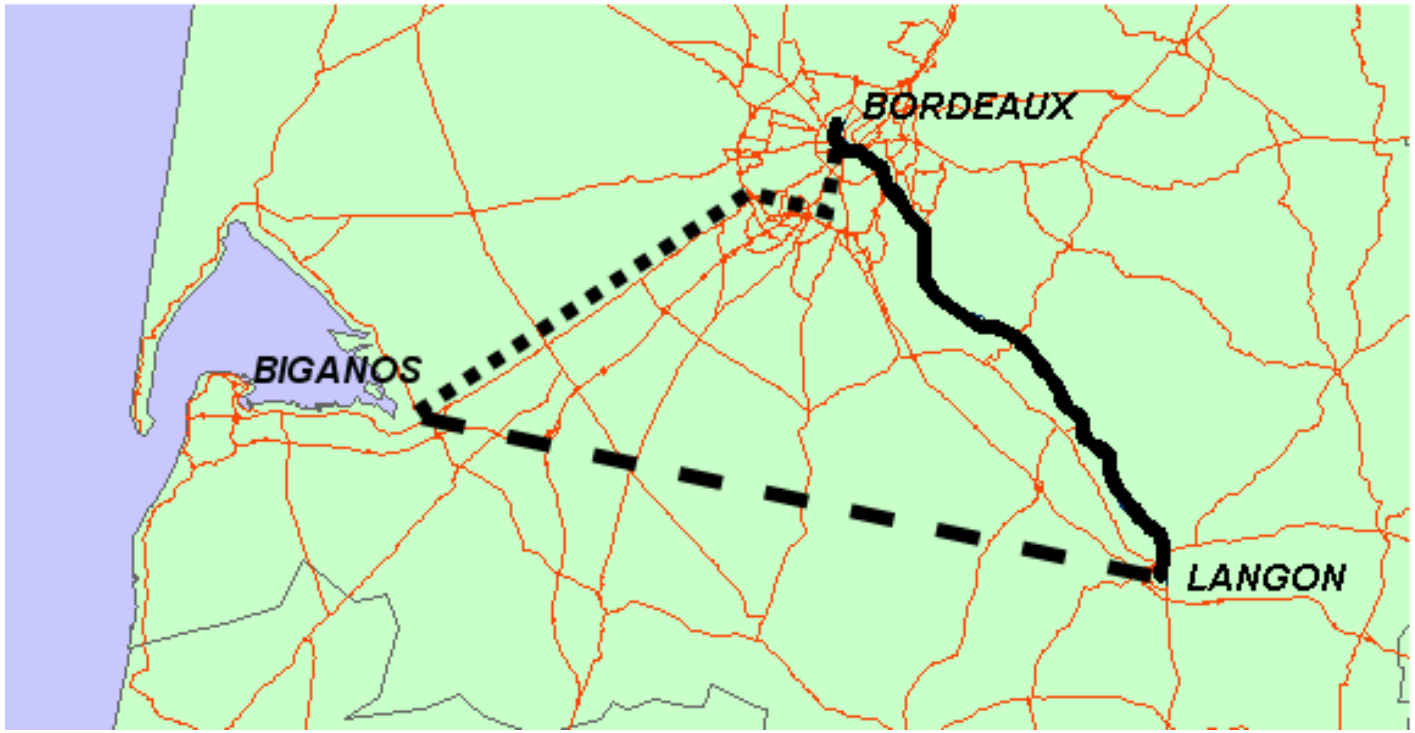
Quitter Saint-Jean-de-Luz en bateau / en voiture



Illustration

Illustration

Bordeau-Langon par la route, Langon-Biganos par une modalité inconnue, Langon-Bordeaux par le train.



5. Conclusion : complémentarité des deux approches

Complémentarité des deux approches

Quand l'approche géographique est en échec...

- Certains « motifs » difficilement différenciables
- Ex : *Description d'un point de vue* vs *itinéraire* ?
- Une analyse des déplacements pourraient résoudre l'ambiguïté

Complémentarité des deux approches

Quand l'approche linguistico-géographique est en échec...

- Lors d'une allusion par l'auteur à un autre voyage
- Le « saut géographique » n'est pas « capté » par cette approche
- Une analyse de la *dispersion* et/ou de l'*ordonnancement* pourrait lever l'erreur.

Complémentarité des deux approches

Exemple (*Excursions Autour du Vignemale*, § 435)

Fatigué et prudent, Ramond, le 25 juillet 1792, abandonnant l'Assemblée Législative et Paris, est venu à Barèges se reposer et attendre les événements. Sa sœur Rosalie l'accompagne. Le 8 août (jour où Junker stationne la Pahule), il la conduit au Pic du Midi ; le 12, au lac d'Escoubous ; le 17, avec Rondo, c'est le passage de Héas à Gavarnie par le Coumélie. Le 21, Cauterets. Le 23, ayant couché dans une cabane de la vallée d'Aspe, passage du col d'Aspe et descente au Saoussa débat. Là, deux choses admirables : le fort déjeuner, et le Vignemale vu dans toute sa majesté.

Complémentarité des deux approches

Exemple (*Excursions Autour du Vignemale*, § 435)

Fatigué et prudent, Ramond, le 25 juillet 1792, abandonnant l'Assemblée Législative et Paris, est venu à Barèges se reposer et attendre les événements. Sa sœur Rosalie l'accompagne. Le 8 août (jour où Junker stationne la Pahule), il la conduit au Pic du Midi ; le 12, au lac d'Escoubous ; le 17, avec Rondo, c'est le passage de Héas à Gavarnie par le Coumélie. Le 21, Cauterets. Le 23, ayant couché dans une cabane de la vallée d'Aspe, passage du col d'Aspe et descente au Saoussa débat. Là, deux choses admirables : le fort déjeuner, et le Vignemale vu dans toute sa majesté.

Conclusion

Bilan

- Approche géographique a pour objectif la RI avancée
- Approche linguistico-géographique a pour objectif la conception d'application
- Les deux approches proposent des solutions pour la détection d'itinéraires
- Ces solutions semblent être complémentaires

Conclusion

Perspectives

- Mixer les deux approches pour lever les ambiguïtés
- Faire intervenir les notions temporelles
- Evaluer les résultats des deux approches sur des corpus plus conséquents
- Valider la complémentarité des deux approches par des expériences plus nombreuses

Merci !

Contacts

- pierre.loustau@univ-pau.fr
- julien.lesbeguerries@univ-pau.fr

Jean-Paul Boons.
La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs.
Langue Française, 76 :5–40, 1987.

A. Borillo.
L'espace et son expression en français.
L'essentiel. Ophrys, 1998.

M. Hasan.
Svm : Machines à vecteurs de support ou séparateurs à vastes marges.
Technical report, Versailles St Quentin, France, 2006.

T. Joachims.
Text categorization with support vector machines : learning with many relevant features.

Lecture Notes in Computer Scienc, Proceedings of ECML-98, 10th European Conference on Machine Learning, series(1398) :137–142, 1998.

David Laur.
Sémantique du déplacement et de la localisation en français : une étude des verbes, des prépositions et de leur relation dans la phrase simple.
PhD thesis, Université de Toulouse II, 1991.

J. Lesbegueries, M. Gaio, P. Loustau, and C. Sallaberry.
Geographical information access for non-structured data.
In *21st ACM Symposium on Applied Computing - Advances in Spatial and Image based Information Systems track, SAC'06*, pages 83–89, Dijon, 2006.

Julien Lesbeguerries and Pierre Loustau.

Extraction et interprétation d'information géographique dans des données non-structurées.
In *Actes de la 3ème Conférence en Recherche d'Information et Applications (CORIA'06)*, Mars 2006.

J. Lesbegueries, C. Sallaberry, and M. Gaio.
Associating spatial patterns to text-units for summarizing geographic information.
In *29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval - GIR (Geographic Information Retrieval) Workshop*, pages 40–43, Seattle (US), 2006. ACM SIGIR.

Thierry Nodenot, Pierre Loustau, Mauro Gaio, Christian Sallaberry, and Philippe Lopisteguy.

From electronic documents to problem-based learning environments : an ongoing challenge for educational modeling languages.
In *7th International Conference on Information Technology Based Higher Education and Training*, pages 75–86. IEEE, juillet 2006.

Claude Vandeloise.
L'espace en français.
Seuil, Paris, France, 1986.