RECONSTRUCTION AUTOMATIQUE D'ITINÉRAIRES À PARTIR DE TEXTES DESCRIPTIFS

par Ludovic Moncla

Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour (LIUPPA)
avenue de l'Université, BP 1155 64013 Pau Cedex
ludovic.moncla@univ-pau.fr

Ces travaux s'inscrivent dans le cadre du projet PERDIDO dont les objectifs sont l'extraction et la reconstruction d'itinéraires à partir de documents textuels. Ma thèse a été réalisée en cotutelle entre l'Université de Pau et des Pays de l'Adour (France) et l'Université de Saragosse (Espagne). Elle a été dirigée par le professeur Mauro Gaio du laboratoire d'informatique de l'Université de Pau et des Pays de l'Adour et par le maître de conférences Javier Nogueras-Iso de l'équipe IAAA de l'Université de Saragosse et encadrée par Sébastien Mustière du laboratoire COGIT de l'IGN. Les objectifs de cette thèse ont été de concevoir un système automatique permettant d'extraire et d'interpréter des déplacements décrits en langages naturels dans des récits de voyage ou des descriptions d'itinéraires, puis de les représenter sur une carte. Nous avons implémenté et évalué les différentes étapes de notre approche sur un corpus multilingue de descriptions de randonnées (français, espagnol et italien).

Introduction

Avec l'émergence ces dernières années de nouveaux besoins, liés notamment aux nouvelles technologies et à de nouveaux comportements, les méthodes de fouille de textes et de traitement automatique du langage naturel sont de plus en plus utilisées afin d'extraire et de structurer l'information provenant d'une masse de données toujours plus importante. Dans le cadre de nos travaux, nous avons proposé une solution automatique pour la représentation cartographique d'un itinéraire à partir de sa description en langage naturel. Notre approche est composée de deux tâches principales. La première vise à identifier et extraire les informations qui décrivent l'itinéraire dans le texte, comme par exemple les entités nommées de lieux (toponymes), les relations spatiales et les expressions de déplacement ou de perception. La seconde tâche a pour objectif la reconstruction de l'itinéraire en combinant les informations extraites du texte et des données géographiques provenant de ressources externes.

Contribution

Détection d'entités nommées et annotation d'informations spatiales

Nous avons proposé une chaîne de traitement pour l'annotation et l'interprétation d'informations spatiales à partir de textes décrivant des déplacements en langage naturel. Notre méthodologie s'est appuyée sur une analyse de corpus afin d'identifier les différents éléments de la langue permettant de décrire un déplacement ou un itinéraire. Nous avons proposé une méthode pour l'extraction automatique de ces différents éléments (noms de lieux, verbes de déplacement, expressions de perception, relations spatiales, etc.). Notre approche repose sur une méthode hybride combinant une étape de prétraitement (étiquetage morpho-syntaxique) associant à chaque mot du texte sa catégorie grammaticale et son lemme puis une étape d'annotation réalisée grâce à l'exécution d'une cascade de transducteurs (patrons lexico-syntaxiques) multilingue (français, espagnol, italien). Nous avons défini la notion d'entité nommée étendue, permettant d'associer une partie descriptive à un nom propre (ex : ville de Paris, refuge du lac des Barmettes). Cette extension de la notion d'entité nommée nous permet une meilleure interprétation des entités grâce à une prise en compte de leur contexte mais constitue aussi une aide pour la classification et la désambiguïsation de ces entités.

De plus, à partir de la définition du concept d'itinéraire et des informations utilisées dans la langue pour décrire un itinéraire, nous avons modélisé un langage d'annotation d'informations spatiales adapté à la description de déplacements, élaboré à partir d'une spécialisation de la TEI (Text Encoding and Interchange) en respectant les recommandations. Ce langage d'annotation est multicouche (Moncla et Gaio, 2015) et permet l'encodage des entités nommées étendues et de différentes informations de manière générique mais aussi de manière plus spécifique, par exemple adapté à l'annotation d'informations spatiales.

Désambiguïsation des toponymes

La désambiguïsation des toponymes est une sous-tâche de la résolution des toponymes (Leidner, 2007) et reste un problème mal résolu en Reconnaissance d'Entités Nommées. On peut distinguer différents types d'ambiguïtés définis dans la littérature (Smith and Mann, 2003) : le même nom peut être utilisé dans un contexte non géographique (referent class ambiguity), le même nom peut être utilisé pour plusieurs lieux (referent ambiguity) et un même lieu peut avoir plusieurs noms (reference ambiguity). La figure 1 illustre la referent ambiguity dans le cadre d'une description de randonnée. Il existe également une ambiguïté au niveau des mots constituant le nom (structual ambiguity) par exemple dans le cas d'entités nommées étendues. Nous avons par ailleurs proposé de considérer l'incomplétude des ressources géographiques comme une forme d'ambiguïté (unreferenced ambiguity). Nous avons proposé des méthodes de désambiguïsation adaptées à trois types d'ambiguïtés : *referent ambiguity*, structural ambiguity et unreferenced ambiguity.

D'après nos expérimentations sur un corpus de descriptions de randonnées, entre 45 et 70% des toponymes (en fonction de la ressource utilisée : IGN, Geonames ou Openstreetmap) présents dans la ressource ont plus d'un référent. Nous avons proposé d'utiliser une méthode de clustering par densité spatiale (DBSCAN) afin d'identifier le plus grand groupement de points distincts (Moncla et al., 2014). Dans un premier temps, pour permettre de lever les ambiguïtés de type « referent », la définition des clusters fournit une méthode de désambiguïsation basée sur une analyse spatiale pour identifier le cluster qui regroupe le plus grand nombre de toponymes distincts en terme de distance géographique. Puis dans un second temps, afin de répondre partiellement au problème posé par l'incomplétude des bases de données géographiques nous proposons d'utiliser la boîte englobante de ce cluster pour estimer la localisation des noms de lieux non référencés. L'approximation de la localisation des lieux non répertoriés peut être affinée grâce à l'interprétation des informations extraites du texte et notamment des relations spatiales.

Nous avons également proposé une méthode de désambiguïsation utilisant les informations annotées grâce au concept d'entité nommée étendue et permettant le sous-typage des toponymes (Nguyen et al., 2013). La combinaison de ces méthodes de désambiguïsation permet d'obtenir de bons résultats et est particulièrement adaptée à des corpus de descriptions de déplacements.

Reconstruction automatique d'un itinéraire

La dernière étape de notre traitement consiste à utiliser les informations extraites du texte afin de reconstruire automatiquement l'itinéraire. Nous proposons un modèle de graphe générique pour la reconstruction automatique d'itinéraires (Moncla et al., 2015), où chaque nœud représente un lieu et chaque segment représente un chemin reliant deux lieux. L'originalité de notre modèle est qu'en plus de tenir compte des éléments habituels (chemins et points de passage), il permet de représenter les autres éléments impliqués dans la description d'un itinéraire, comme par exemple les points de repères visuels. Un calcul d'arbre de recouvrement minimal à partir d'un graphe pondéré est utilisé pour obtenir automatiquement un itinéraire sous la forme d'un graphe (Moncla et al., 2014 ; Moncla et al. 2015). Chaque segment du graphe initial est pondéré en utilisant une méthode d'analyse multicritère combinant des critères qualitatifs et des critères quantitatifs. La valeur des critères est déterminée à partir d'informations extraites du texte et d'informations provenant de ressources géographiques externes (ordre des toponymes dans le texte, distance géographique, expression de perception associée au toponyme, etc.). Par exemple, nous combinons également les informations issues du traitement automatique de la langue comme les relations spatiales décrivant une orientation (ex : *se diriger vers le sud*) avec les coordonnées géographiques des lieux trouvés dans les ressources pour déterminer la valeur du critère « relation spatiale ». La figure 2 montre le résultat de la reconstruction d'itinéraires sur trois exemples de randonnées. Les lignes rouges représentent les traces GPS réelles des itinéraires et permettent d'évaluer le résultat de la reconstruction (visible en bleu).

Implémentation et évaluation

Nous avons implémenté notre chaîne de traitement dans une architecture Web modulaire. Chaque étape du traitement est implémentée au sein de modules indépendants (annotation morpho-syntaxique, extraction d'information et reconnaissance d'entités nommées, résolution des toponymes, désambiguïsation des toponymes et reconstruction de l'itinéraire). Nous avons également développé un démonstrateur en ligne et des Web services (http://erig.univ-pau.fr/PERDIDO/) permettant d'utiliser les différents modules de notre chaîne de traitement.

La figure 3 montre un aperçu du résultat d'annotation et de reconstruction d'itinéraires produit par notre chaîne de traitement pour une description de randonnée.

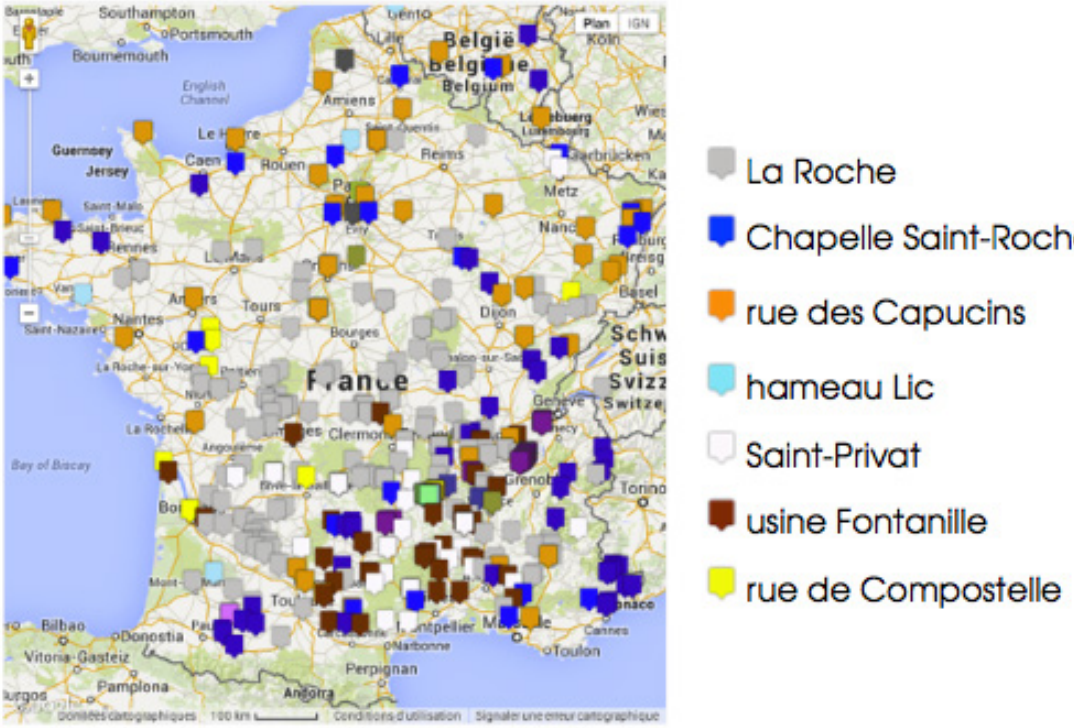

Figure 1 : Illustration du « referent ambiguity »

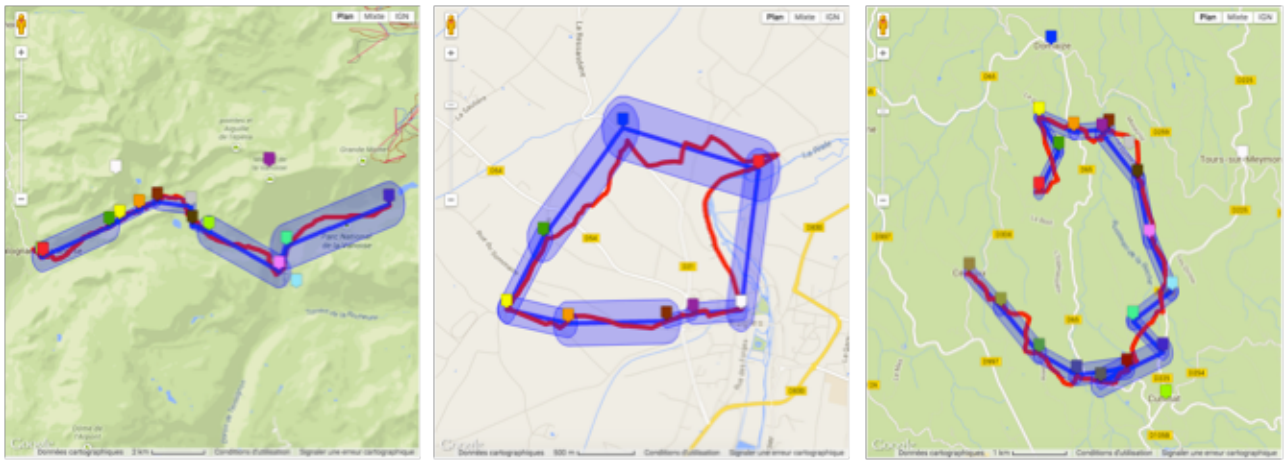

Figure 2 : Exemples de résultats de reconstructions d'itinéraires

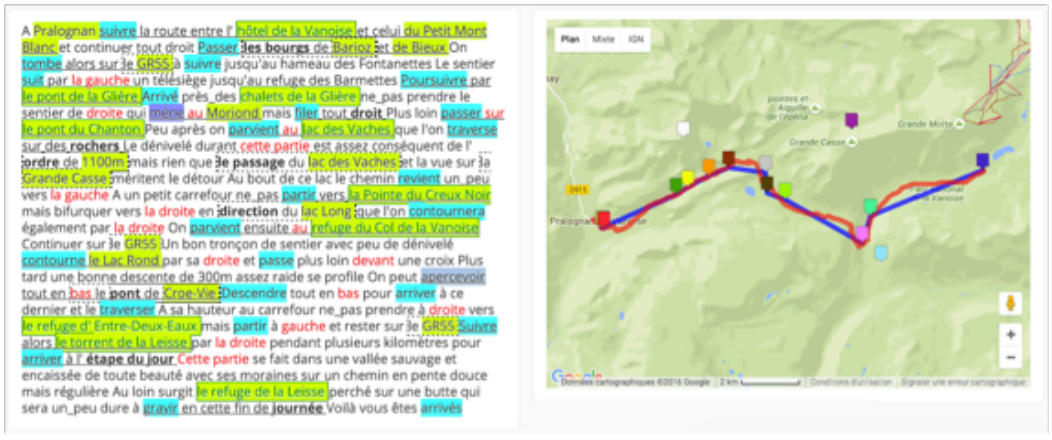

Figure 3 : Exemple graphique de résultat du processus d'annotation et de reconstruction de l'itinéraire

Nous avons construit un corpus d'évaluation composé de descriptions de randonnées pour les trois langues. Ces documents ont été automatiquement collectés sur des sites internet de partage de randonnées et sont associés à leurs traces GPS. Ces traces GPS permettent l'évaluation automatique des résultats de la reconstruction automatique produite par notre système. Nous avons également annoté manuellement ce corpus afin d'évaluer en détail chaque étape du traitement. Pour l'étape de reconnaissance des entités nommées étendues, nous obtenons 82% d'entités bien reconnues (en tenant compte de différents cas d'erreurs : insertion, suppression, substitution, etc.) contre 45% avec le détecteur d'entités nommées CasEN (Maurel et al., 2011). Concernant l'étape de reconstruction de l'itinéraire, nous obtenons un score de 72% de bonne reconstruction en comparant les résultats avec les traces GPS et un score de 96% en comparant le résultat automatique avec l'ordre des points de passage construit manuellement. Notre méthode permet donc une bonne interprétation des informations présentes dans le texte et propose une première approximation de l'itinéraire décrit. Des améliorations sont envisagées comme la prise en compte du relief, de la géométrie des objets géographiques et du réseau routier en milieu urbain afin de proposer une approximation plus fine de l'itinéraire.

Bibliographie

Leidner, J. L. (2007) *Toponym Resolution in Text : Annotation, Evaluation and Applications of Spatial Grounding of Place Names.*

Smith, D. A. and Mann, G. S. (2003) *Dootstrapping toponym classifiers.* In Proceedings of the HLT-NAACl workshop on Analysis of geographic references. Stroudsburg, USA.

Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol-Taravella, I., and Nouvel, D. (2011). *Cascades de transducteurs autour de la reconnaissance des entités nommées.* TAL, 52(1):69–96.

Moncla, L. and Gaio, M. (2015) *A Multi-Layer Markup Language for Geospatial Semantic Annotations.* Proceedings of the 9th Workshop on Geographic Information Retrieval (GIR'15).

Moncla, L., Gaio, M., Nogueras-Iso, J., and Mustiere, S. (2015) *Reconstruction of itineraries from annotated text with an informed spanning tree algorithm.* International Journal of Geographical Information Science (IJGIS), 22 pages.

Moncla, L., Renteria-Agualimpia, W., Nogueras-Iso, J., and Gaio, M. (2014) *Geocoding for texts with fine-grain toponyms : an experiment on a geoparsed hiking descriptions corpus.* Proceedings of the 22nd ACM SIGSPATIAL Conference, pp. 183-192 Dallas, Texas, USA.

Moncla, L., Gaio, M., and Mustiere, S. (2014) *Automatic itinerary reconstruction from texts.* Proceedings of the 8th GIScience Conference. pp. 253-267, Vienna, Austria.

Nguyen, V. T., Gaio, M., and Moncla, L. (2013) *Topographic subtyping of place named entities : a linguistic approach.* Proceedings of the 15th AGILE Conference, 5 pages, Leuven, Belgium.