REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'enseignement supérieur et de la recherche scientifique
Université Mentouri de Constantine
Faculté des sciences de l'ingénieur
Département d'Informatique

Mémoire en vue de l'obtention du diplôme Magistère en informatique

Numéro d'ordre : 356/mag/2009
Numéro de série : 013/inf/2009

Data mining spatial :

La Désambiguïsation des Toponymes

Présenté par : Imene Bensalem

Encadrée par : Dr. Mohamed Kireddine Kholladi

Le jury est composé de :

Président: Dr. Alloua Chaoui

Rapporteur: Dr. Mohammed Kireddine Kholladi

Examinateur : Dr. Saidouni Djamel Eddine

Examinateur : Dr. Salim Chikhi

Soutenu le 29 octobre 2009

Résumé

L'espace géographique est une dimension omniprésente. La façon de se référer à un lieu dans cet espace, peut être formelle basée sur les coordonnées spatiales, ou informelle, que nous employons dans la langue naturelle en utilisant les toponymes (les noms des lieux). La présentation formelle est la base de tous les traitements spatiaux que peut effectuer la machine. Ces traitements ne sont pas possible en utilisant les toponymes. Les informations géographiques sont parmi les informations qui peuvent être extraites du texte en utilisant les techniques du traitement automatique des langues naturelles, mais malheureusement, elles ne peuvent être exploitées que si les lieux géographiques sont représentés d'une manière formelle, ce qui n'est pas souvent le cas dans les documents textuels. La conversion entre la représentation formelle et la représentation informelles des lieux géographiques est donc une nécessité pour pouvoir bénéficier des informations géographiques extraites du texte. La désambiguïsation de toponyme associe aux occurrences de toponymes dans le texte leurs représentations formelles. Cette tâche est problématique à cause de l'ambiguïté des toponymes. En effet un toponyme peut être le nom de plusieurs lieux dans le monde. La désambiguïsation des toponymes est une tâche primordiale dans une multitude d'application entre autre le data mining spatial. Ce mémoire traite le problème de la désambiguïsation de toponymes en présentant une nouvelle heuristique qui utilise une source d'évidence qui n'a pas encore été exploité dans les méthodes le l'état de l'art.

Mots clé : désambiguïsation de toponymes, informations géographiques, relations arborescentes, data mining spatial.

مُلخص

انفضاء انغشاف ً بْتعذي راخف ًك ميكا ٌ. ونى الإشارة رانى يكا قُ ً ْز اانفضاء ني ا شيك م سٌا ضَتاس ورخذاو الإحنتك ٍإخان كا َح، أوتشك م غثٍ سٌا ضَ ألا و هوأس ًاء ان قٍَع انغشاف ٍح، و ْزانشك مٍْ الأكثش ذذا ال فً راه غحان طٍ عٍح. انشك مانش ٌطض ًٍْأساسك مان ًعانداخ ونلس ىتاخ فلضطنعٍحان ًٌ ٍك آنٍ حانقٍ اوتٍ ًا. تذٍ أ ٌ ْزان ًهنٍح اخن ٍسدي ك ُتخاس ورخذاوأس ًاء الأيك ٍ.ذس ًر فٍ ُ ٍاخيعلندح راه غاخ انطٍ عٍتخاس ورث اج ي ع ؤٍاخ غشاف ٍح ي ٍان ُص ٍص، وك ٍ للأسف لا ٌك ٍ يعلندح ْزان ًعٍيات و لا الإفادج يٍْ ال إراكا ًَدان قٍَع انغشاف ٍحيوث حشك م سٌا ضً، و ْزا شك م ىَاس الاسوخذاوف ًان ُص ٍص كٍ ًَايكٍ تح ت راه غحان طٍ عٍح. إ ٌانرس ٌم ت ٍانشك مانش ٌا ضً وغير انس ٌا ضًن ٍرث ٍمان اقٍَع انغشاف ٍح تاخ يأشا ضروس ٍيانسفدجي ٍان ًع ؤٍاخ انغشاف ٍحانس ًَرخشخ ي ٍ انُضوص. ٍضٍ رأس ًاء الأيك ٍٍي عٍَ ٍحست ط الأسد ًاءانغشاف ٍحان ًٍخدج ًانُص ٍصتوتًٍ ًه ٌ انشٌا ضً .رنغش ْزان ًًٍح إشكلن ٍحتانُس ٍحآل ٍنة و رنكتسثة غ ٍَ ضأس ًاء الأيك ٍ. فً اراق عق ذ ٌك إلسى انإزذ أ ٌ ٌث ًمعذج أيك فٍ ًانعاى . ٍضٍ رأس ًاء الأيك ٍٍْإخشاء ي ًَتانُس ٍتح هعذ ٌذ ي ٍانطٍ قٍاخي ٍت ٍُ ًَانرق ٍة فً انثٍا ًَاخان كا ًَح. ونٍ ناول ْز الأطروح حشكً ٍضٍ رأس ًاء الأيك ٍن و رنك تاقوشاذ اسخصري ٍح خذٌ ذجذس ورخذو يصلدسا لإزانح انغ ًَضٍ ى رًٌ ى اس ورخذي ًاي قٍث م فً الأسلن ٍٍةان ًٍخدج ز انٍا.

انكه ًاخانش ئٍس ٍح : ٍضٍ رأس ًاء الأيك ٍ، انًٍ ع ؤٍاخ انغشاف ٍح،ان الق اخ ثنس دش ٌح،انرق ٍة فً انثٍا ًَاخان كا ًَح.

Abstract

The geographical space is an ubiquitous dimension. Referring to locations in this space can be formal, based on the spatial coordinates, or informal, that we use in natural language using toponyms (place names). The formal presentation is the basis of all special processing that can make the machine. These processing are not possible using toponyms. Geographic information can be extracted from the text using natural languages processing techniques, but unfortunately it cannot be exploited unless the geographical locations are represented in a formal way, which is often not the case in textual documents. The conversion between the formal and the informal representations of geographical locations is a necessity to benefit from geographic informations extracted from the text. Toponym Disambiguation associates occurrences of place names in the text with their formal representations. This task is problematic because of the ambiguity of place names. In fact a toponym may be the name of several places in the world. Toponym Disambiguation is an essential task in a variety of applications among other spatial data mining. This thesis addresses the problem of toponym disambiguation by presenting a new algorithm that uses a source of evidence that has not yet been exploited in the state of the art methods.

Keywords: Toponym Disambiguation, geographic information, arborescent relationships, Spatial Data Mining.

Remerciement

Tout d'abord, Louange et Remerciement éternel et immense à Allah, Seigneur de l'univers, pour sa charité et sa générosité infinie envers moi.

Je tiens à remercier infiniment mes parents pour leurs encouragements et leur soutien aux moments de joie et de détresse. Mes remerciements particuliers à ma mère qui n'a jamais cessé de prier pour moi, et je ne suis arrivée ici qu'avec la « baraka » de sa prière.

Je remercie mon encadreur Dr. Mohammed Khireddine Kholladi d'avoir accepté l'encadrement de ce travail, et je tiens à remercier les membres du jury Dr. Alloua Chaoui, Dr. Saidouni Djamel Eddine, et Dr. Salim Chikhi d'avoir pris la peine de l'évaluer.

Je suis énormément reconnaissante à Abdelhamid Baha (BAAZ entreprise), Khawla Chaib (ingénieur en informatique) pour les discussions précieuses à propos des données géographiques qui ont permis d'enrichir mes connaissances pour rédiger ce mémoire.

De même que je suis reconnaissante à Saloua Chettibi (magister en informatique) et les ingénieurs (par ordre alphabétique) Amina Moualkia, Hanène Zitouni, Meriem Kemmouch de m'avoir annoté le corpus CSTR que j'allais utiliser avant de m'opter pour le sujet de désambiguïsation des toponymes.

Je tiens à remercier les chercheurs Simon Overell (Imperial collage London), Davide Busaldi (Universidad Politécnica de Valencia, Espagne) et Nicola Stokes (University College Dublin, Ireland), d'avoir répondu à mes questions et fournir plus d'explications sur leurs travaux. Des remerciements particuliers à Simon Overell qui m'a proposé d'évaluer ma méthode en utilisant le corpus GeoSemCor, et à Davide Buscaldi de m'avoir envoyé une version originale de son article (Buscaldi et Rosso 2008) et aussi d'avoir partagé le corpus GeoSemCor gratuitement sur le Web.

Je remercie énormément le chercheur Andras Csomai (Google) pour ses conseils précieux et pour sa suggestion de programmer avec le langage Perl qui m'a fait gagner beaucoup de temps.

Je suis très reconnaissante à mes collègues et amies (par ordre alphabétique) Khouloud Meskaldji et Sara Boutamina de m'avoir corrigé la langue de l'article (Bensalem et Kholladi 2009c).

Je remercie mes amies (par ordre alphabétique) Hanène Zitouni, Khouloud Meskaldji, Naouel Ouafek pour leur soutien moral.

Un remerciement particulier à mon oncle Khalil qui m'a soutenu matériellement lors des journées scientifiques en informatique à Oran ; afin de présenter mon article. Ainsi qu'à ma sœur Abir et mon frère Walid qui m'ont cédé constamment leurs tours à utiliser le PC.

Mes remerciements aussi aux Dr. Mourad Bouznada et Dr. Allaoua Chaoui de m'avoir aidé à surmonter certaines contraintes administratives.

Et finalement, je remercie tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

Imene

Liste des figures

Liste des tableaux

Table des matières

Introduction générale

> *Presque tout ce qui se passe, se passe quelque part*
> (Longley, et al. 2005)

Contexte de la recherche

L'espace géographique est une dimension omniprésente. Chacun d'entre nous connaît au moins son lieu de naissance, lieu de résidence, lieu de travail, lieux où habitent les parents et les amis, les lieux qu'il a visité et d'autres dont il a entendu parler,…, etc. En plus, généralement, il n'est pas suffisant pour nous de connaître les lieux mais nous voulons toujours plus d'information sur ces lieux. Quotidiennement on se renseigne sur la météo de notre ville ; si on veut voyager on s'informe plus sur la destination ; nous lisons les journaux pour s'informer sur les évènements d'actualité de certains lieux,…, etc. Brièvement, nous vivons sur la surface de Terre, il est donc naturel que l'ensemble de nos activités, nos expériences, nos connaissances et, nos décisions soient liées à des lieux sur l'espace géographique.

La façon de se référer à un lieu, peut être *formelle,* basée sur les *coordonnées spatiales* comme la longitude et la latitude ou d'autres systèmes de *géo-référencement*, ou *informelle*, employée dans la *langue naturelle* en utilisant les *toponymes* (les noms des lieux) et les adresses postales (Hill 2006).

La représentation formelle est comprise par la machine car est elle précise et peut subir des calcules mathématiques. Le fait de connaître les coordonnées spatiales permet à un *système d'informations géographiques* de calculer les distances, les surfaces, et les directions, et d'effectuer des analyses comme la détection des relations spatiales (ex. le chevauchement et l'inclusion), ce qui n'est pas possible en utilisant les toponymes (Hill 2006).

Cependant, l'Homme préfère et comprend la représentation informelle. Les toponymes sont des mots fréquemment employés dans la *langue naturelle* orale ou

écrite. Chacun de nous connaît son adresse postale, et peut identifier les lieux des évènements par les toponymes, mais peu sont en mesure de préciser les coordonnées spatiales des endroits qu'ils connaissent (Longley, et al. 2005).

Le traitement automatique de la langue naturelle (TALN) est devenu un besoin indispensable pour bénéficier des grandes quantités de données textuelles stockées dans les pages web, les bibliothèques numériques, les rapports officielles, etc. Les informations géographiques sont parmi les informations qui peuvent être extraites du texte, mais malheureusement, elles ne peuvent être exploitées efficacement par la machine sauf si les lieux géographiques sont représentées d'une manière formelle, ce qui n'est pas souvent le cas dans les documents textuels. En fait, il a été estimé qu'au moins 70% des documents textuels contiennent des références aux lieux géographiques sous forme de toponymes (MetaCarta, Inc.).

La conversion entre la représentation formelle et la représentation informelle des lieux géographiques est donc une nécessité pour pouvoir bénéficier des informations extraites d'un texte où la mention des lieux géographiques est considérée importante, comme dans les textes d'actualité, de l'histoire, les biographies, et les rapports de voyage, etc.

À l'instar de plusieurs mots de la langue naturelle, les toponymes sont des mots ambigus, c.à.d. un seul toponyme peut être le nom de plusieurs lieux dans le monde (plusieurs *référents*). Si l'Homme ne pense même pas cette ambiguïté, celle-ci est considérée une problématique pour la machine.

La *Désambiguïsation des Toponyme* (DT) —aussi appelée la *Résolution des Toponymes*— est la tâche d'attribuer un emplacement géographique unique à un nom de lieu ambigu qui apparaît dans un contexte donné. Une fois un toponyme est désambiguïsé il sera possible de le présenter d'une manière formelle, par exemple, par la latitude et la longitude.

Motivation

Les méthodes de la désambiguïsation des toponymes utilisent le contexte comme source d'évidence principale. Les éléments du contexte les plus exploités pour résoudre un toponyme sont les toponymes qui apparaissent avec lui dans le même texte.

Une analyse de l'état de l'art de la DT nous a permis de remarquer que beaucoup de méthodes supposent une certaine proximité géographique entre les référents des toponymes du même contexte, et les résolvent ainsi sur cette base. Certaine méthodes comme (Leidner, Sinclair et Webber 2003) et (Smith et Crane 2001) supposent une *proximité spatiale* entre les référents des toponymes, donc elles résolvent les toponymes par les référents les plus proches entre eux en terme de distance géométrique. D'autres méthodes comme (Buscaldi et Rosso 2008) supposent une proximité dans l'arbre hiérarchique des lieux du monde que nous appelons une *proximité arborescente*. Ces méthodes résolvent les toponymes par les référents les plus proches entre eux dans l'arbre hiérarchique des lieux.

La relation arborescente la plus exploitée dans les méthodes de DT de la littérature est la *méronymie* (c.-à-d. la relation *est-partie-de*). En fait, La quasi-totalité des méthodes basées sur la proximité arborescente sont basées sur la découverte de ce type de relations entre les référents des toponymes du même contexte. Par exemple, si les toponymes du contexte sont {Constantine, Algérie} les méthodes basées sur la méronymie résolvent ces toponymes ambigus respectivement en {Constantine>Algérie, Algérie>Afrique} au lieu par exemple de {Constantine>Michigan>USA, Algérie>Massachusetts>USA}[1], car il y a une relation de méronymie entre les référents du premier ensemble (Constantine est méronyme de Algérie [2]), et ce n'est pas le cas dans le deuxième ensemble.

[1] Ces référents sont obtenus du glossaire géographique Getty disponible en ligne dans l'adresse : http://www.getty.edu/research/conducting_research/vocabularies/tgn.
[2] C'est-à-dire Constantine est partie de l'Algérie.

Cependant, d'autres relations arborescentes comme l'*holonymie* (l'inverse de la *méronymie*) et les *relations non hiérarchique* n'ont pas été –jusqu'à présent– bien exploitées comme sources d'évidence. Par ailleurs, le rôle de la détection des relations arborescentes dans la désambiguïsation des toponymes n'a pas été encore étudié.

Contribution

Notre contribution se résume en 4 points :

1. Classifier les relations géographiques qui peuvent contribuer à la désambiguïsation des toponymes en *relations arborescentes* et *relations spatiales,* et proposer une nouvelle vue du problème de la désambiguïsation des toponymes en considérant les relations arborescentes (avec touts leurs types) comme sources d'évidence.

2. Introduire la métrique de la *Densité Géographique* qui quantifie le degré des relations arborescentes entre les référents des toponymes.

3. Proposer une heuristique capable de résoudre les toponymes ambigus dans un texte en se basant sur la découverte de toutes les relations arborescentes qui existent éventuellement entre eux.

4. Étudier l'effet de la découverte des relations arborescentes dans la désambiguïsation des toponymes en comparant les performances de notre méthode à celles de quelques autres méthodes, entre autre une méthode basée sur la découverte des relations spatiales.

En plus, nous avons réalisé un état de l'art des méthodes de désambiguïsation des toponymes selon notre propre point de vue.

Position de la désambiguïsation des toponymes par rapport à d'autres domaines

Une fois les toponymes qui existent dans un texte sont identifiés puis désambiguïsés, ils peuvent être utiles dans une multitude d'applications. Par exemple, dans un moteur de recherche, les réponses aux requêtes contenant des toponymes deviennent grâce à la DT plus précises, car le système de *recherche d'information* devient capable de distinguer les lieux qui portent le même nom, et donc définir la pertinence d'un document pour une requête sur cette base. La *visualisation des collections de documents* sur des cartes géographiques pour des fins d'*analyse* ou de *navigation* devient aussi possible grâce à l'étiquetage de chaque document par les coordonnées géographiques des toponymes qu'il renferme. La DT permet aussi d'intégrer sans ambiguïté des données géographiques en provenance du texte dans des bases de données géographiques. Ces dernières peuvent en suite subir une analyse en utilisant entre autre le *data mining spatial.*

Le *data mining spatial* (DMS) est une analyse approfondie qui sert à découvrir des relations et des modèles implicites dans les grandes quantités de données spatiales. L'intégration de données de plusieurs sources est une opération de préparation des données pratiquement présente dans tout projet du data mining (spatial ou autre). La désambiguïsation des toponymes se situe donc dans la phase de prétraitement des données dans le processus du data mininig spatial, notamment, si les sources des données à intégrer sont des documents textuels.

La désambiguïsation des toponymes se situe dans l'intersection de deux disciplines qui sont le traitement automatique de la langue naturelle (TALN) et les systèmes d'information géographique (SIG). Chacune de ses deux disciplines lui offrent un éventail de techniques.

Plan du mémoire

Notre mémoire s'articule de la manière suivante :

Notre départ dans le monde de la recherche pour réaliser ce mémoire était dans le domaine du data mining spatial, mais nous somme arrivée à une contribution dans la désambiguïsation des toponymes. Le chapitre 1 explique en détail la position de la DT par rapport au DMS en passant par la définition des données géographiques qui sont le point central qui relie les deux domaines.

Le 2ème chapitre présente les différents types de l'ambiguïté des toponymes et positionne notre recherche par rapport à ces types. En outre, en raison de la nature multidisciplinaire de ce mémoire, nous présentons dans ce même chapitre des notions de base dans tous les domaines qui possède une relation avec la désambiguïsation des toponymes comme le traitement automatique de la langue naturelle, les systèmes d'informations géographique et la recherche d'information afin de préparer le terrain pour la suite du mémoire.

Dans le chapitre 3 nous discutons l'état de l'art des différents travaux sur la désambiguïsation des toponymes en distinguons 4 composants intrinsèques pratiquement à toute méthode qui sont : le contexte, les heuristiques, les connaissances, et les ressources. En plus, nous proposons des critères de classification des heuristiques et des connaissances, et nous pensons que cette classification pourrait réduire la grande diversité entre les méthodes de sorte qu'elle les organise dans des catégories génériques, et par conséquent elle aide à leur comparaison et assimilation.

Dans le chapitre 4 nous discutons certaines lacunes dans les heuristiques de la DT, notamment la non exploitation des différentes relations possibles entre les toponymes du même contexte, et nous proposons une heuristique qui remédie à cette lacune. Les performances de notre méthode sont ensuite comparées à celles

d’autres méthodes est des conclusions sont tirées en analysant les résultats de comparaison.

Finalement, nous terminons par une conclusion générale qui présente un résumé de notre recherche et un ensemble de perspectives.

Chapitre 1

Du Data Mining Spatial à la Désambiguïsation des Toponymes

1.1 Introduction

Notre recherche –afin de réaliser ce mémoire– a commencé par l'exploration d'un large domaine qui est le data mining spatial (DMS), or elle a abouti à une contribution dans un domaine spécifique qui est la désambiguïsation des toponymes.

En effet, la désambiguïsation des toponymes est une tâche indépendante en elle-même mais elle peut être considérée comme une étape d'une importance primordiale dans plusieurs domaines. Le fait que le DMS fût le domaine de notre départ, il aurait constitué une forte raison de consacrer ce chapitre à la démonstration de sa relation avec la contribution principale du présent mémoire.

Ce chapitre s'articule comme suit : les sections 1.2 jusqu'à 1.5 présentent un aperçu sur le data mining spatial. Les données géographiques –qui sont le point commun entre le DMS et la DT– sont en suite l'objet de la section 1.6. La section 1.7 explique la relation entre le DMS et la DT et on termine par une conclusion qui récapitule brièvement les principaux points discutés.

1.2 Définition et objectifs du data mining spatial

En raison de la grande quantité (habituellement, téraoctets) de données spatiales, il est coûteux et souvent irréalistes pour les utilisateurs de les examiner en détail. Le *data mining spatial* (fouille de données spatiales en français) vise à automatiser un tel processus de découverte de connaissances (Ng et Han 1994).

Le data mining spatial implique l'application d'outils informatiques pour révéler des patterns intéressants dans des objets et des événements répartis dans l'espace géographique et dans le temps (Miller et Han 2001). Il est défini aussi comme l'extraction de connaissances, de relations spatiales, ou d'autres patterns intéressants qui ne sont pas explicitement stockées dans les bases de données spatiales (Han et Kamber 2006). Son objectif est d'automatiser le processus de compréhension des données spatiales par des représentations concises qui font

apparaitre la sémantique des données. Ces représentations sont appelées : connaissances, et elles sont sous forme de relations spatiales, ou relations entre les données spatiales et non spatiales.

Une fois les connaissances sont découvertes par le DMS, elles peuvent être utilisées pour la construction des bases de connaissances spatiales, la réorganisation des bases de données spatiales, et l'optimisation de requêtes spatiales (Han et Kamber 2006).

1.3 Exemples historiques fameux de l'exploration des données spatiales

Shekhar & Chawla(2003) ont cité quelques exemples bien connus qu'ils ont eu lieu avant l'invention de l'ordinateur, mais ils bien illustrent le type de connaissances découvertes par le data mining spatial :

1. En 1855, lorsque le choléra asiatique soufflait à Londres, un épidémiologiste a marqué tous les lieux où la maladie a frappé (ce sont les données spatiales) sur une carte, et a découvert que les lieux forment un cluster (cela est la connaissance découverte) dont le centre s'est avéré être une pompe à eau. Lorsque les autorités gouvernementales éteignaient la pompe à eau, le choléra a commencé à s'estomper. Plus tard, les scientifiques ont confirmé la nature des eaux d'origine de la maladie.

2. En 1909, un groupe de dentistes ont découvert que les habitants de Colorado Springs ont exceptionnellement des dents saines, et ils ont attribué ça au niveau élevé de fluor naturel dans l'eau potable locale. Les chercheurs ont ensuite confirmé le rôle positif du fluor dans la lutte contre la carie dentaire. Maintenant, toutes les municipalités dans les États-Unis assurent que l'eau potable est fortifié avec du fluorure.

Dans ces exemples les scientifiques ont découvert des corrélations entre les données : le choléra et l'eau d'une pompe, et le fluor et la santé des dents. Sans

inspection minutieuse et approfondie d'un grand nombre de données, il est impossible de découvrir ces connaissances. Le rôle du data mining spatiales est d'automatiser la découverte de telles corrélations (Shekhar and Chawla 2003).

1.4 Les tâches du data mining spatial

Les tâches du DMS sont généralement une extension des tâches du Data mining (DM) en intégrant les données et les relations spatiaux. On trouve ainsi les règles associatives spatiales, le clustring spatial, la classification spatiale, l'analyse des tendances spatiales, et l'analyse des cas singuliers. Une brève description de ces tâches est présentée ci-dessous.

1.4.1 Les règles associatives spatiales

Une règle associative est une implication de la forme « si A alors B » (Gardarin 1999) ou plus formellement notée : A⇒B, [s%; c%] où A et B sont des ensembles de prédicats spatiaux et non spatiaux, s% est le support de la règle, et c% est sa confiance (Han et Kamber 2006). Les règles associatives servent à trouver des associations entre des propriétés des objets et celles de leur voisinage (Aufaure, Yeh et Zeitouni 2000).

Exemple

La règle suivante est une règle associative spatiale :

Est-un(X, "école") ∧ proche-de(X, "station de bus") ⇒ proche-de (X, "marché") [20%; 80%].

Cette règle stipule que 80% des écoles qui sont proches des stations de bus sont également à proximité des marchés, et que 20% des données appartenant à un tel cas.

1.4.2 Les collocations spatiales

Ce sont un type spécifique des règles d'association. Elles représentent des sous-ensembles d'objets géographiques qui apparaissent fréquemment proches les uns des autres dans une carte géo-spatiale (Han et Kamber 2006, Miller 2007). Ces objets géographiques sont représentés par des attributs booléens qui indiquent leur présence ou leur absence dans un endroit dans la surface de la Terre. Des exemples des objets géographiques booléens incluent les espèces végétales, les espèces animales, les types de routes, les cancers, la criminalité, et les types d'activités économiques (Shekhar, Zhang, et al. 2004).

La Figure 1-1 (voir Page 13) montre un ensemble de données qui consistent à des instances des objets spatiaux booléens, chacun d'eux est représenté par une forme distincte. Un examen attentif révèle le pattern de collocation {■, ♦}.

Exemple

Un exemple en écologie : la tâche des collocations spatiale peut découvrir que le crocodile du Nil et le pluvier égyptiens vivent dans les mêmes endroits géographiques.

1.4.3 Le clustering spatial

Le clustering est une méthode de classification automatique non supervisée qui regroupe des objets dans des classes. Son but est de maximiser la similarité intra-classes et de minimiser la similarité interclasses.

La transposition au domaine spatial des méthodes de clustering s'appuie sur une mesure de similarité d'objets localisés suivant leur distance métrique. Néanmoins, la finalité du clustering en spatial n'est pas tant de former des classes que de détecter des concentrations anormales (par exemple, détecter un point chaud dans l'étude de criminalité, ou des zones à risque en accidentologie) (Aufaure, Yeh et Zeitouni 2000). Voir (Ng et Han 1994) pour plus de détails sur le clustering spatial.

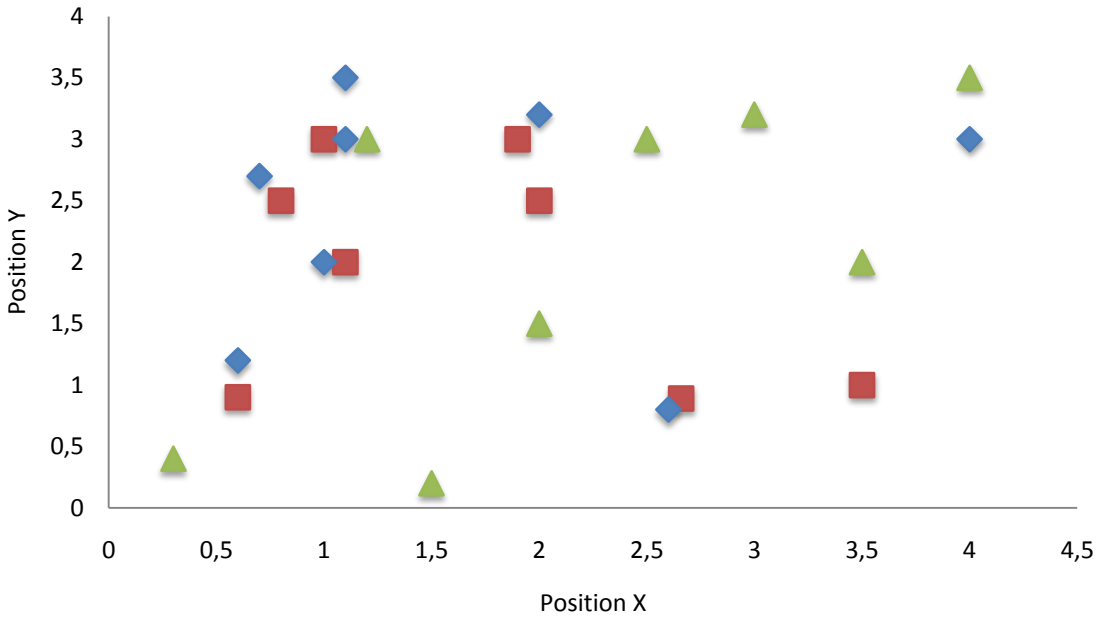

Figure 1-1. Exemple de collocations spatiales. Le pattern {■, ♦} est une collocation spatiale

Exemple

Le clustering est utilisé pour déterminer les "points chauds" dans l'analyse de criminalité et le suivi de maladies. L'analyse des points chauds "Hot spot analysis" est le processus de chercher des clusters d'évènements denses et inhabituels à travers le temps et l'espace. De nombreux organismes de justice pénale dans le monde profitent des avantages fournis par les technologies informatiques pour identifier les points chauds de la criminalité afin de prendre des stratégies préventives, comme le déploiement de patrouilles dans les zones de points chauds (Shekhar, Zhang, et al. 2004).

1.4.4 La classification spatiale

La tâche de classification consiste à attribuer un objet à une classe parmi un ensemble donné de classes. Cette attribution est faite sur la base des valeurs d'attribut de cet objet. Dans la classification spatiale les valeurs des attributs des

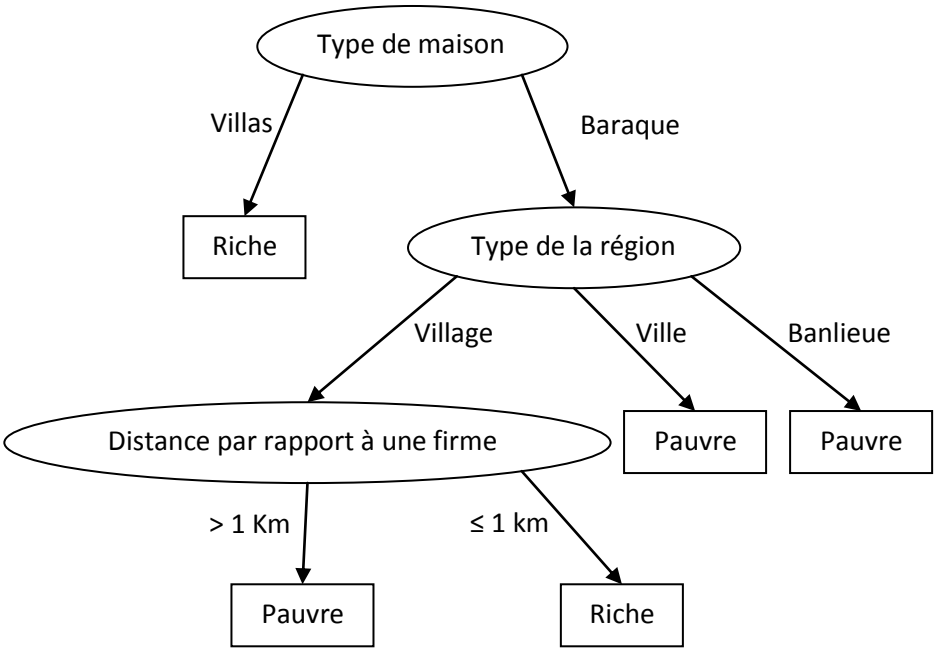

Figure 1-2. Arbre de décision pour la classification des régions en riches vs pauvres

objets voisins d'un objet peuvent également être pertinents pour sa classification, donc elles doivent être prises en considération (Azimi and Delavar 2007).

Cette tâche est réalisée par l'apprentissage supervisé qui, à partir de classes fournies partiellement en extension (un échantillon de la base de données), induit une description en intention (un modèle générique qui relie les attributs) permettant de classer les prochaines données (Aufaure, Yeh et Zeitouni 2000).

Exemple

Supposons que nous souhaitons classifier les régions d'une wilaya en riches versus pauvres. Pour ce faire, il faut identifier les facteurs importants liés à l'espace qui détermine la classification d'une région. Beaucoup d'attributs peuvent révéler intéressants pour cette classification, comme, le type de la région (village, banlieue, ville), type de maison qu'elles contiennent (villas, Baraques), et être à proximité d'une firme. Un modèle de classification est représenté sous forme d'un arbre de

classification (voir Figure 1-2)[1] ou d'un ensemble de règles, appelées aussi arbre de décision et règles de décision respectivement.

1.4.5 L'analyse des tendances spatiales

La tendance spatiale est un changement régulier d'une ou de plusieurs attributs non-spatiales lors du déplacement en dehors d'un objet donné (Azimi and Delavar 2007).

Les techniques souvent utilisées pour l'analyse de tendances spatiales sont la régression et l'analyse de corrélations.

Exemple

Analyser la tendance du taux de chaumage selon la distance par rapport à une métropole ou une capitale, ou la tendance du changement du climat ou de la végétation selon la distance par rapport à la côte.

1.4.6 L'analyse des cas singuliers

Les cas singuliers ou encore appelés valeurs aberrantes et extrêmes (*outliers* en anglais) sont des objets qui ne respectent pas le comportement général ou le modèle de données (Han et Kamber 2006).

Shekhar et al (2004) définissent un cas singulier spatial comme un objet spatialement référencé dont les valeurs des attributs non-spatiaux sont inconsistants avec celles des autres objets à l'intérieur d'un certains voisinage spatial.

Exemple

Un taudis (gourbi) dans un cartier de villas est un objet spatial aberrant en se basant sur l'attribut non spatial « type de maison ».

[1] Cette exemple est imaginaire, c.à.d. il ne représente pas une vraie étude sur des données réelles.

Nous avons présenté dans cette section des méthodes d'extraction de pattern. Cependant la validité des ces patterns n'est pas un but facile à atteindre. L'application triviale des tâches du data mining peut conduire à de faux résultats. En effet, les tâches du data mining ne sont pas « stand-alone » mais elles doivent s'exécuter au sein d'un processus bien déterminé, ce qui est l'objet de la section suivante. Une des étapes de ce processus est ensuite l'entrée vers le domaine de la désambiguïsation des toponymes.

1.5 Le processus de découverte de connaissance[2]

Nous présentons dans cette section un nouveau concept qui est la *découverte de connaissance dans les bases de données* en montrant ses étapes et sa relation avec le data mining. Ce que nous intéresse -bien sur- dans ce mémoire est le data mining spatial et la découverte de connaissances spatiales. Cependant les points discutés dans cette section ne se limitent pas aux données spatiales, mais concernent plutôt le data mining et la découverte de connaissances dans leurs sens génériques indépendamment des type de données sur lesquelles ils s'appliquent (relationnelles, spatiales, textuelles, multimédia…). C'est pour cette raison que nous avons choisi dans cette section d'utiliser les termes *data mining* et *découverte de connaissance* sans la spécification « spatial ».

1.5.1 Définition et étapes

La *découverte de connaissances dans les bases de données*, plus connu avec son acronyme anglais *KDD* (*Knwoledge discovery in databases*) est le processus non trivial d'identification de modèles valides, nouveaux, potentiellement utiles, et compréhensibles dans les données[3] (Fayyad, Piatetsky-Shapiro and Smyth 1996).

Le terme processus signifie que le KDD se compose de plusieurs étapes. Ces étapes peuvent être résumées en trois phases globales, à savoir : la préparation des

[2] Des parties de cette section ont été publiées dans (Bensalem et Kholladi 2008)

données, le data mining, et l'évaluation des modèles. Ces phases sont définies brièvement ci-dessous. Toutefois, les détails ne sont pas l'objet de ce mémoire. Voir (Han et Kamber 2006) pour une ample explication.

La préparation des données : elle comprend la collecte, l'intégration, la transformation, le nettoyage, la réduction, et la description des données.

Le data mining : consiste à appliquer des méthodes issues de la statistique, et de l'apprentissage automatique pour découvrir des modèles importants et utiles sur les données. Parmi les méthodes du DM, la classification, clustering, les règles associatives, etc. (voir la section 1.4).

L'évaluation des modèles : consiste à estimer l'erreur et la précision sur les modèles extraits, et mesurer leur utilité, leur originalité et leur intelligibilité. Un modèle est considéré comme une connaissance s'il est utile, inconnu auparavant, et dépasse un certain pourcentage de précision.

1.5.2 Le sens large et le sens étroit du data mining(Bensalem et Kholladi 2008)

La préparation des données, et aussi l'évaluation des modèles (les phases respectivement avant et après l'application des tâches du DM) sont des phases d'une importance primordiales. La phase de préparation de données seule contribue de 75 à 90% à la réussite du projet de fouille (Pyle 2003). C'est pourquoi il n'est pas question de négliger ces étapes dans la réalité. Ignorer les phases de préparation des données ou d'évaluation des modèles rendrait inutile le DM et nous met en danger d'obtenir des modèles étrangers à la réalité.

Ce lien étroit entre le data mining et les phases antérieures et postérieures est la raison derrière l'émergence d'un autre point de vue sur sa définition. Certains chercheurs comme (Han et Kamber 2006) définissent le data mining comme

[3] Par analogie au KDD, la Découverte de Connaissance Géographique DCG (en anglais Geographique knowledge discovery (GKD)) est le processus d'extraction d'informations et de connaissances à partir

l'ensemble des phases de découverte de connaissances et non pas seulement la phase d'extraction de patterns[4]. Par conséquent, il existe deux sens du terme *data mining* (voir Figure 1-3), dont l'un est un sens large : tout le processus de découverte de connaissances, tandis que l'autre est un sens étroit : l'étape d'extraction de patterns dans le processus de découverte de connaissances (Bensalem et Kholladi 2008).

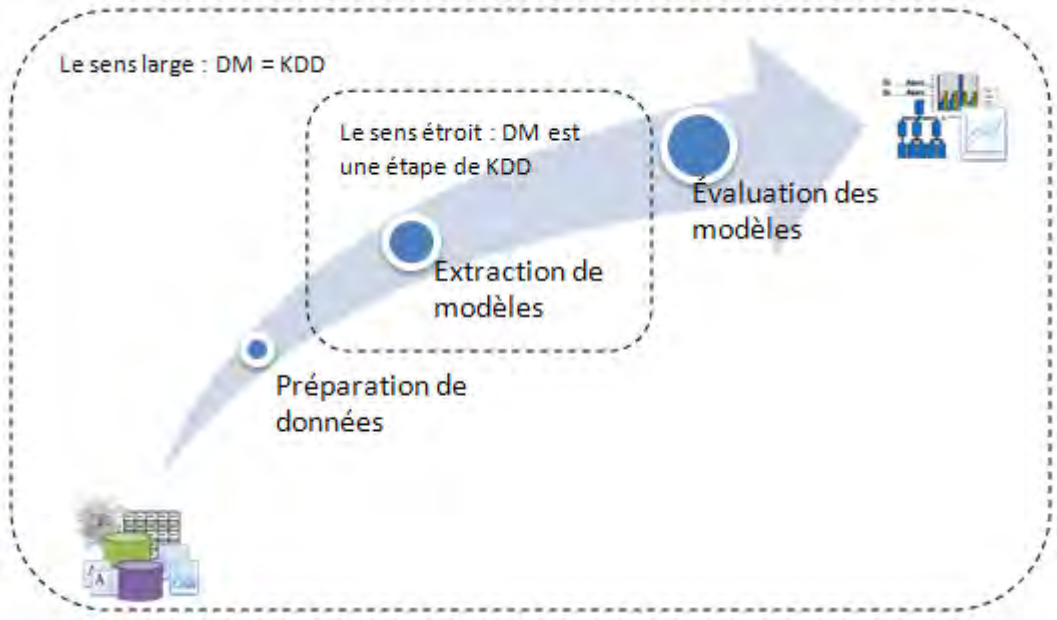

Figure 1-3. La relation entre le data mining et le KDD

Dans le reste de ce chapitre nous utilisons le terme data mining dans son sens large car c'est l'étape de collecte de données[5] qui vas nous permettre de montrer la relation du DMS avec la désambiguïsation des toponymes.

Comme nous avons déjà mentionné, le data mining spatial est une extension du data mining classique (c.-à-d. celui appliqué sur les données alphanumériques, relationnelles ou transactionnelles) avec une adaptation aux données spatiales. Les

des grandes bases de données géo-référencées (Miller, 2007).

[4] Par analogie, Shekhar & Chawla (2003) voient que le data mining spatial est un processus qui contient toutes les phases de découverte de connaissances géographiques.

[5] La collecte de données est une étape dans la phase du prétraitement de données.

données spatiales sont donc un concept clé pour le data mining spatial, et elles le sont également pour la désambiguïsation des toponymes.

La section suivante donne une vue globale sur ce type de données et plus particulièrement sur les données géographiques. Toutefois, on se limite aux aspects que nous considérons indispensables pour la compréhension du reste du mémoire. Quelques aspects des donnée géographiques ont été complètements omis, comme, la représentation raster et vectorielle, et les relations topologiques; les autres sont présentées d'une façon plus ou moins détaillées. Pour des détails plus amples sur les informations géographiques et des domaines en relation voir (Longley, et al. 2005).

1.6 Les données géographiques

Les données géographiques et spatiales sont de plus en plus nombreuses. Avec l'avènement du Web, la manipulation des données géographiques spécifiquement n'est plus exclusive aux communautés scientifiques et professionnelles mais elle est devenue une tâche presque quotidienne ou probablement indispensable dans la vie de l'Homme moderne.

1.6.1 Spatiale ou géographique : quelle est la différence ?

Les données spatiales concernent tous les phénomènes où les entités pouvant être intégrés à l'intérieur de certain espace formel qui génère des relations implicites entre elles. Cet espace peut être non géographique comme les surfaces des autres planètes et l'espace de l'univers. Une image médicales est un exemple de donnés spatiales ou l'espace de référence est le corps humain.

Les données géographiques concernent un cas particulier où les entités sont géo-référencées c.-à-d. elles se réfèrent à la surface de la Terre ou à ces proximités (Longley, et al. 2005, Miller 2007).

Les informations de l'environnent collectées par les capteurs numériques comme la température et la pression sont un exemple typique des informations géographiques. Les images satellitaires de la terre comme celles du fameux Google Hearth sont aussi un exemple bien connu des informations géographiques manipulées dans le Web. Les évènements d'actualités sont aussi des informations géographiques car ils se produisent dans des lieux déterminés dans la Terre.

Brièvement, toute information qui peut être liée à un endroit est une information spatiale. Si cet endroit est un emplacement dans la Terre, on parle alors d'une information géographique. Une information géographique est donc une information spatiale, mais le contraire n'est pas toujours vrai.

L'adjectif spatial (idem pour géographique) est ajouté à toute opération ou objet qui manipule les données spatiales, comme, requête spatiale, analyse spatiale, data mining spatial, base de données spatiale,…, etc. Par exemple, la requête « Quels sont les noms des librairies de Constantine? » est une requête géographique car elle contient une donnée géographique qui est « Constantine ».

Nous nous intéressant dans ce mémoire aux données géographiques spécifiquement. Néanmoins, nous utilisons l'adjectif « spatial » soit comme synonyme de « géographique » ou si le dit contexte n'est pas exclusif aux données géographiques.

1.6.2 Caractéristiques des données géographiques

Les données géographiques ont plusieurs caractéristiques qui les différencient des données alphanumériques simples. Le texte suivant mentionne certaines de ces caractéristiques. Il convient de noter que cette liste de caractéristiques n'est pas exhaustive.

- Les données géographiques sont **multidimensionnelles**, car deux coordonnées doivent être spécifiées pour définir un emplacement, par exemple la latitude et la longitude (Longley, et al. 2005).

- Les objets géographiques peuvent avoir de **multiples représentations** géométriques ; une rue par exemples peut être représentée par une surface ou une ligne selon les besoins.
- Les informations géographiques sont **complexes**. Elles sont composées d'une donnée spatiale, éventuellement des données temporelles, et un ensemble d'attributs (données attributaires). La section suivante fournit plus de détails sur ce point.
- **L'importance de la notion de précision** liée notamment aux procédures de collecte et de saisie de données(Laurini 1996). En effet, la qualité des résultats de l'analyse et des requêtes spatiales est liée à la précision des données.
- Les informations géographiques se manipulent par un outil logiciel appelé un **système d'informations géographiques** (SIG).

1.6.2.1 Les composants d'une information géographique

Comme nous avons déjà mentionné, une information géographique comprend trois composants principaux : une donnée spatiale, une donnée temporelle, et des données attributaires.

1.6.2.1.1 Les données spatiales

Une donnée spatiale renvoie à *l'emplacement géographique* d'une entité ainsi que sa *forme géométrique*. D'un point de vue SGBD, c'est une donnée liée à un système de coordonnées spatiales et dont son type est l'un des types géométrique fournis par le SGBD ou définis par le système d'information géographique (SIG).

L'emplacement d'un objet est représenté par un *localisant* (Laurini 1996) qui est une information permettant de localiser un objet dans l'espace. Le localisant joue le rôle d'un identifiant de l'objet géographique, et il est spécifié par rapport à l'un des systèmes de géo-référencement comme les adresses postales et les coordonnées géographiques. Voir le chapitre suivant pour plus d'informations à propos du géo-référencement.

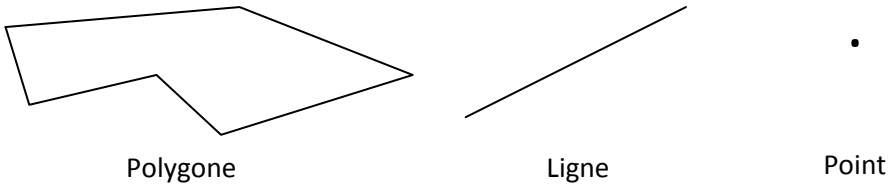

Figure 1-4. Les types géométriques élémentaires d'une donnée spatiale

Les formes géométriques élémentaires associées aux objets géographiques sont : le point, la ligne et, la surface (voir Figure 1-4). Ces formes sont des types abstraits de données géométriques qui peuvent être manipulées par des extensions de SQL.

- Le **point** est un type de base composé de deux ou trois coordonnées (X, Z) ou (X, Y, Z) selon la dimension utilisée 2D ou 3D respectivement. Un objet de type point représente par exemple le barycentre d'une ville dans une grande échelle.

- Une **ligne** est composée d'une liste de points. Elle peut représenter par exemple une route ou une rivière.

- Une **surface** est composée d'une chaîne fermée de lignes connectées, ayant un intérieur et un extérieur. Elle peut représenter par exemple un pays, un village, …etc. Une surface fait généralement référence à un polygone.

Il convient de noter que le localisant est la plus importante donnée d'une information géographique car il constitue la base de plusieurs avantages comme : La capacité de visualiser les objets sur la carte, de lier différents types d'information au fait qu'elles se réfèrent à au même lie, et de mesurer les distances et les superficies. Sans localisant, les données sont censées être non-spatiales et n'auraient aucune valeur au sein d'un système d'information géographique (Longley, et al. 2005).

```
Create table wilaya_algeria
(
Nom_wilaya   VARCHAR(20),
Population   INTEGER,
Position     POLYGON,
)
```

Nom_wilaya	Population	Position
Oran	897 700	POLYGON(4.97 44.42, 4.96 44.41, 4.95 44.417,…,4.90 44.37)

Figure 1-5. Exemple d'une table d'informations géographiques

1.6.2.1.2 Les données temporelles

Les informations géographiques ne contiennent pas obligatoirement des données temporelles. Toutefois le temps est une données importante dans certains domaines comme la géophysique et la météorologie (Longley, et al. 2005).

1.6.2.1.3 Les attributs

Sont des données alphanumériques classiques décrivant les caractéristiques quantitatives ou qualitatives de l'entité géo-référencée.

Certains attributs sont physiques ou environnementaux, comme la température d'un lieu, tandis que d'autres sont sociaux ou économiques, comme la population d'un pays. D'autres attributs représente une mesure de quelque chose dans un endroit et éventuellement dans le temps, par exemple, la température atmosphérique, tandis que d'autres représentent un classement en catégories, par exemple, les catégories d'utilisation de terrains, qui distinguent entre les terrains d'agriculture, d'industrie, ou résidentiels (Longley, et al. 2005).

Exemple

La Figure 1-5 (voir page 23) représente une table de données géographiques. La willaya d'Oran (le nom est une donnée attributaire) a une population de 897700 (données attributaire), et elle est représentée sur la carte par un polygone dont chaque sommet est représenté par des coordonnées spatiales (l'attribut position est une données spatial de type polygone)[6].

1.6.2.2 Sources de données géographiques

Les données géographiques peuvent être collectées de plusieurs sources, ou bien achetées auprès d'un fournisseur privé ou public.

Les sources connues des données géographiques sont les cartes, les sondages, les SIG, les images satellitaires, etc. Avec l'avènement du Web et des bibliothèques numériques, et le développement des techniques du traitement automatique des langues naturelles (TALN), une nouvelle source a commencé à prouver son utilité ; cette source est les documents textuels.

Nous proposons de classifier les sources de données géographiques selon le « type de données ». Ce critère de classification les divise en 2 catégories principales: sources fournissant des données structurées, et sources fournissant données non structurées. Le Tableau 1-1 illustre cette classification.

Ce qui nous intéresse dans ce mémoire est l'obtention des données géographiques à partir du texte. Dans la section suivante nous présentons des brèves descriptions de quelques travaux dans ce sujet.

[6] La valeur de la donnée spatiale n'est qu'un exemple et ne représente pas la position spatiale réelle d'Oran.

Tableau 1-1. Classification des ressources d'informations géographiques selon le type de données

			Exemples
Type de données que fournis la source	Données structurées		Bases de données spatiales Glossaires géographiques Fichiers plats Tableau de données
	Données non structurées	Images	Images satellitaires Photos aériennes Images obtenues à partir caméras vidéo au sol Cartes géographiques scannées
		Texte	Pages Web Collection de document : rapport professionnel, article de presse…

1.6.3 Des exemples de travaux sur l'utilisation du texte comme une source de données géographiques

Dans cette section nous présentons quelques travaux dans la littérature dont les données géographiques sont extraites du texte en langue naturelle puis utilisées dans des applications différentes.

1.6.3.1 Extraction des descriptions des villes pour la mise à jour d'un SIG urbain

Borges, Laender, Medeiros, Silva, et Davis (2003) ont utilisé le Web comme une source importante d'informations géographiques urbaines. Ils ont proposé un environnement qui permet d'extraire des données géographiques à partir des pages Web (comme les noms des villes, des rues, des boulevards, et autres), les convertir au format XML, puis les utiliser pour mettre à jour une base de données géographique d'un SIG urbain.

1.6.3.2 Data mining spatial sur des données géographiques extraites des pages web

Dans (Morimoto, et al. 2003) les auteurs ont présenté un système d'extraction de connaissances spatiales à partir des collections de pages web contenant des informations géographiques comme les adresses et les codes postaux. Pour chaque information géographique, ils ont appliqué des techniques du *géocodage* (voir le chapitre suivant pour plus d'informations sur le géocodage) pour calculer ses coordonnées géographiques. Ensuite, ils ont extrait les concepts-clés des pages web, puis formé une table d'associations géographiques dont chaque tuple contient les concepts-clé d'une page web et les coordonnées géographiques des lieux qu'elle renferme. Finalement des techniques du data mining spatial sont appliquées pour trouver des patterns spatiaux par exemples les collocations spatiales.

1.6.3.3 L'extraction et la visualisation des événements

Li, Srihari, Niu, et Li (2003) ont construit un entrepôt dynamique de connaissances à partir des documents textuelles (articles d'actualités et guide de touristes). Le but de la construction de cet entrepôt est de supporter plusieurs applications comme le data mining, et la visualisation et l'analyse des évènements. Parmi les informations contenues dans cet entrepôt des profiles des personnes et des descriptions des évènements. Ces derniers sont des informations géographiques du fait qu'elles sont composées d'une donnée spatiale qui est le lieu de naissance dans les profiles de personnes et le lieu d'occurrence dans les évènements.

1.6.3.4 Base de données géographique pour la conscience de la situation

L'extraction des évènements[7] à partir des documents textuelles à été utilisé aussi pour crées une base de données géographiques pour la conscience de situation[8] (Kalashnikov, Ma, et al. 2006, Kalashnikov, Ma, et al. 2006). La base de donnée est

[7] Les évènements sont des informations géographiques.

construite pour être analysée probablement par le data mining, ou tout simplement pour l'interrogation[9].

Le Tableau 1-2 résume les travaux présentés ci-dessus.

Tableau 1-2. Quelques travaux qui utilisent les documents textuels comme une source d'informations géographiques

	Sources textuelles	**Buts d'extraction des informations géographiques**
(Borges, et al. 2003)	Page web	Mise à jour une base de données géographique d'un SIG urbain
(Morimoto, et al. 2003)	Page web	Data mining spatial
(Li, et al. 2003)	Articles d'actualités et guide de touristes	Génération des profiles de personnes Visualisation et analyse des évènements Text mining
(Kalashnikov, Ma, et al. 2006)	Les registres de communications transcrites et les rapports déposés par les premiers intervenants après la catastrophe du 9/11. Articles de journaux et rapports de blog portant sur le tsunami de l'Asie.	Construire une BD des évènements pour la conscience de situation

1.6.3.5 Discussion

Après avoir examiné un ensemble de travaux sur l'utilisation du texte comme une source d'informations géographiques, nous avons pu tirer les remarques suivantes :

[8] La conscience de situation (situational awareness (SA)) est la perception des éléments de l'environnement dans un volume de temps et d'espace, la compréhension de leur signification, et la projection de leur état dans le futur proche.

- Les informations géographiques souvent extraites du texte sont : les évènements, les adresses et les codes postaux, les noms des lieux, les noms des routes, les numéros de téléphone,…etc.
- Les informations extraites soit elles sont utilisées pour construire une base de données comme le cas de l'extraction des évènements et la génération des profiles de personnes, soit pour enrichir une base de données géographiques déjà existante.
- Les bases de données géographiques construites à partir des documents textuels avaient des utilisations variées dans la littérature entre autre l'analyse et la visualisation des évènements et le data mining.
- L'extraction des entités géographiques à partir des documents textuels utilisent des techniques pour identifier les informations géographiques dans le texte et d'autres pour relier ces informations à une position unique sur la Terre.

1.7 La relation entre le data mining spatiales et la désambiguïsation des toponymes

Nous avons montré dans la section précédente que les documents textuels peuvent servir comme une source de données géographiques. En plus, dans certains travaux comme (Morimoto, et al. 2003) le data mining spatial a été utilisé pour tirer des connaissances à partir des informations géographiques provenant du texte.

La question qui se pose maintenant est : quel est la relation de tout ça avec la désambiguïsation des toponymes qui est le sujet principal de ce mémoire ?

[9] D'après une communication personnelle avec Dmitri V. Kalashnikov, le premier auteur des deux articles cités ci-dessus.

Tableau 1-3. Comparaison entre les toponymes et les coordonnées géographiques

Toponymes	Cordonnées géographiques
Données attributaires	Données spatiales
Non formels (nominales)	Formelles
Ne peuvent pas subir les calculs géométriques	Permettent les calculs géométriques
Manipulés beaucoup plus par l'Homme dans le texte et la parole	Manipulées beaucoup plus par la machine, notamment par les SIG

En effet, l'utilisation du texte comme source de données (géographiques et non géographiques) pâti d'un grand problème qui est l'ambiguïté des sens des noms propres. Généralement, cette ambiguïté consiste à l'utilisation d'un seul nom pour représenter des entités différentes.

Les toponymes c.-à-d. les noms des lieux sont parmi les noms propres qui peuvent être extraits du texte, notamment pour construire une base de données géographiques. À l'instar des autres types de noms propres, les toponymes sont des noms très ambigus (voir le chapitre suivant). Constantine, par exemples, est le nom de 17 lieux dans le monde[10].

L'ambiguïté des toponymes est un problème pour le data mining spatial pour deux raisons, d'un coté, elle **réduit la qualité de données**, qui est un facteur important pour la réussite du data mining[11], et d'un autre côté c'est **un obstacle à l'intégration de données** de plusieurs sources, qui est une étape importante pour la préparation des données du DMS.

En outre, les toponymes sont des données attributaire **non formelles**. Il est donc nécessaire de les convertir en données formelles comme la latitude et la longitude dans le but d'obtenir une base de données géographiques au sens du mot c.-à-d.

[10] D'après Getty Thesaurus of Geographic names online http://www.getty.edu/research/conducting_research/vocabularies/tgn (consulté le 6 mai 2009)

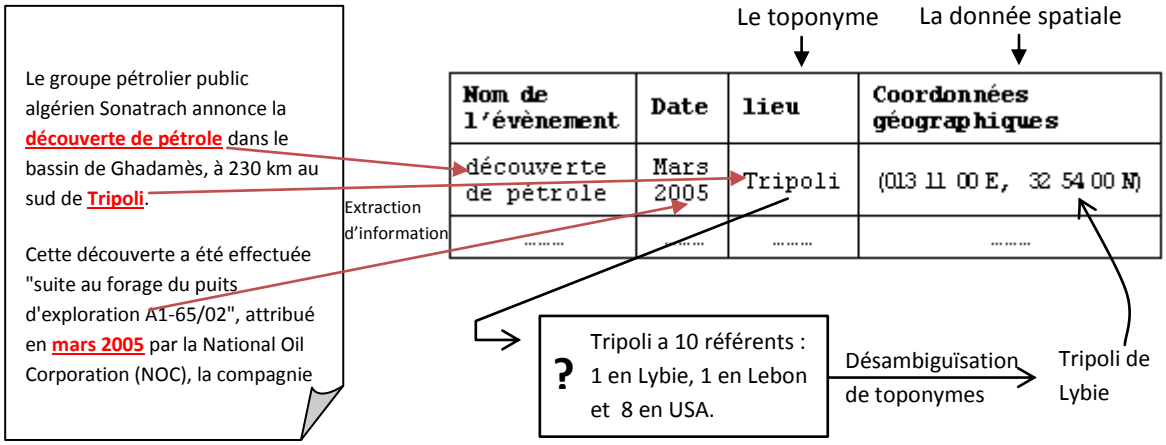

Figure 1-6. Le rôle de la désambiguïsation des toponymes dans la construction d'une base de données géographiques à partir du texte

une base de données qui contient des données spatiales (voir Section 1.6.2.1). Contrairement aux toponymes, Ces dernières, peuvent d'un côté, subir des calculs géométriques, qui sont les opérations de base des tâches du DMS et d'un autre coté, elles sont précises, ce qui est une caractéristique centrale pour la réussite du data mining spatial. Le Tableau 1-3 (voir Page 29) résume les différences qui existent entre les coordonnées géographiques qui sont des données spatiales et les toponymes qui sont donnée attributaire.

La désambigüisation des toponymes peut être considérée comme une étape de prétraitement de données dans le processus du DMS permettant de déterminer le lieu à lequel il se réfère chaque toponyme ambigu extrait de la source textuelle. Autrement dit, la désambiguïsation des toponymes permet d'attribuer à un toponyme, qui est une donnée ambiguë non formelle, une position unique dans la Terre, qui est une donnée précise. Cette dernière peut être convertie en une représentation formelle (spatiale) qui est indispensable pour les traitements spatiaux notamment le data mining spatial. La Figure 1-6 est une illustration de ce point.

[11] L'application du data mining (spatial ou autre) sur des données ambiguës va sûrement engendrer des résultats erronés.

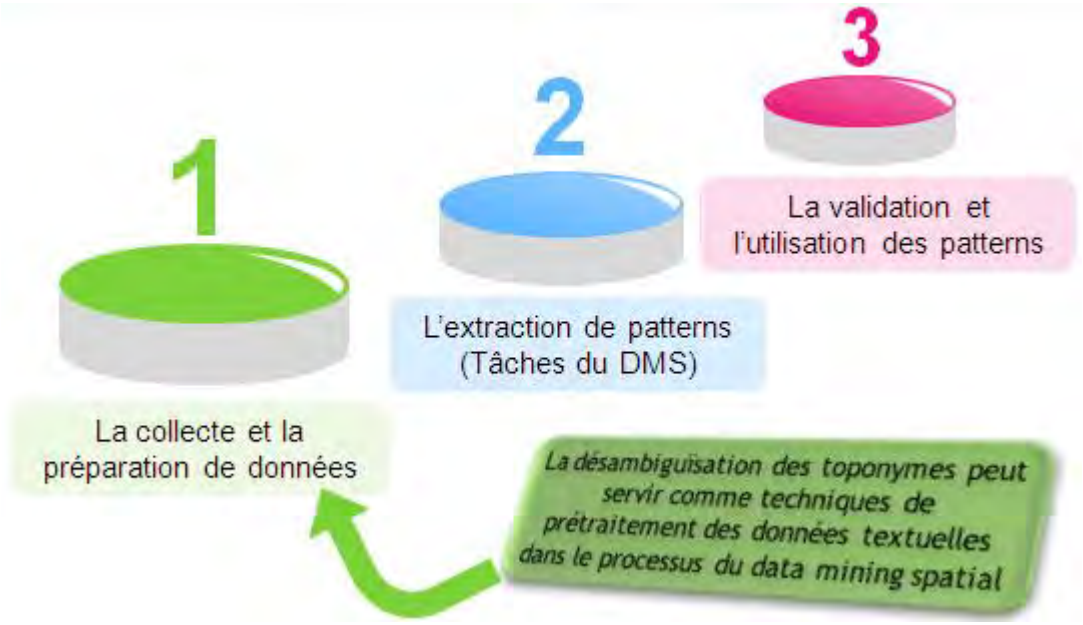

Figure 1-7. La position de la désambiguisation des toponymes dans le processus du data mining spatial

Brièvement, la relation du data mining spatial et la désambiguïsation des toponymes se résume dans la phrase suivante : la désambiguïsation des toponymes est une technique indispensable dans la phase de préparation des données du data mining spatial dans le cas où le texte en langue naturelle est utilisé comme une source des données géographiques. La Figure 1-7 montre la position de la DT dans le processus du DMS.

1.8 Conclusion

La préparation de données en vue de construire une base de données géographiques est une phase d'une importance primordiale, en particulier si cette base va subir une analyse en utilisant par exemple le data mining spatial, car la qualité de l'analyse dépond de la qualité des données.

Les données géographiques peuvent être collectées depuis plusieurs sources. Les données extraites du texte en langue naturelle y compris les toponymes sont généralement ambigües et non formelles, ce qui est un problème pour les traitements spatiaux comme le data mining spatial.

La désambiguïsation des toponymes peut servir comme technique de préparation de données géographique afin que ces dernières soient exploitables par les tâches du data mining spatial. Cette technique devient utile au sein du processus du DMS dans le cas où le texte en langue naturelle est la source des données géographiques à lesquelles le DMS sera appliqué.

Nous avons présenté dans ce chapitre un aperçu sur le data mining spatial et nous avons expliqué le rôle de la désambigüisation des toponymes dans son processus.

Il convient de noter que la désambiguïsation des toponymes est une tâche indépendante en elle-même, et le data mining spatial n'est pas le seul champ de son application. Le chapitre suivant fournit plus de détails sur cette tâche et discute sa relation avec d'autres domaines.

Chapitre 2

La Désambigüisation des Toponymes : Notions de Base

2.1 Introduction

Dans ce chapitre nous définissons davantage la tâche de désambiguïsation des toponymes, nous présentons ses différentes applications et nous précisons sa position par rapport à d'autres domaines.

2.2 Les toponymes

Nous avons mentionné précédemment que le terme *toponyme* désigne un nom de lieu. Dans cette section nous parlons d'avantage sur les toponymes, leurs types, et une de leurs caractéristiques qui est l'ambiguïté.

2.2.1 Définition

Un toponyme peut être défini comme un nom de lieu ou un nom géographique, ou encore un nom propre d'une localité ou d'une région ou d'une certaine autre partie de la surface de la Terre ou de ses objets naturels ou artificiels (Hill 2006). Brièvement, un toponyme est un nom propre qui sert à désigner un emplacement géographique.

Le sens du terme *toponyme* est vaste. Le Tableau 2-1 montre les différents types de toponymes.

Tableau 2-1. Les types de toponymes

Type de toponyme	Détails
Noms des lieux habités	Noms des villes, villages…
Noms des divisions géographiques, politiques et administratives	Noms des continents, pays, wilayas, daïra, états (comme en USA), capitaux,…
Noms des reliefs (oronyme)	Noms des montagnes, plaines, plateaux…
Noms des voix de communications (odonyme)	Noms des routes, rues…
Noms des étendu d'eau (hydronyme)	Noms des océans, mers, rivières…
Fabrication humaine (artéfact)	Noms des tours, villas, centre, université…

2.2.2 L'ambiguïté des toponymes

L'ambiguïté est inhérente aux langues naturelles. Les toponymes –autant que termes de la langue– sont très ambigus. En fait, l'ambiguïté des toponymes a 2 types : l'ambiguïté géo/géo, l'ambiguïté géo/non-géo.

L'**ambiguïté géo/géo** se pose lorsqu'un toponyme représente plusieurs lieux (Amitay, et al. 2004), par exemple, selon les gazetteer[1] Getty[2] et Geonames[3] Constantine est le nom de 5 lieux habités dans le monde (voir Figure 2-1).

Figure 2-1. Les référents de Constantine dans le monde

L'**ambiguïté géo/non-géo** apparaît lorsqu'un toponyme se réfère à d'autres types d'entités (ex. Arafat est le nom d'un lieu à coté de La Mecque et aussi le nom de l'ex-président de Palestine) ou possède d'autres sens (ex. java un langage de programmation et Java une ile indonésienne).

[1] Un gazetteer est un terme anglais qui désigne traditionnellement un dictionnaire de toponymes qui organise des informations sur les lieux géographiques. Nous avons choisi dans ce mémoire d'utiliser cette appellation anglaise car il n'y a pas une traduction unique et précise en français. Voir le chapitre 3 pour plus d'informations sur les gazetteer.
[2] http://www.getty.edu/research/conducting_research/vocabularies/tgn
[3] http://www.geonames.org

2.3 La désambiguïsation des toponymes

2.3.1 Définition

La *Désambiguïsation des Toponymes* (DT) a plusieurs appellations dans la littérature : *Résolution des Toponymes* (Leidner 2007), *Normalisation des Locations* (Li, et al. 2003), *Grounding* ou *Localisation* (Amitay, et al. 2004). La DT est une tâche qui adresse l'ambiguïté des toponymes de type géo/géo et elle est définie dans la littérature avec plusieurs points de vue. Nous avons choisi de présenter les trois définitions ci-dessous.

La désambiguïsation des toponymes est :

« *La tâche de déterminer quelle place l'on entend par une occurrence d'un nom de lieu* » (Amitay, et al. 2004).

« *La tâche d'attribuer un emplacement à un nom de lieu ambigu* » (Li, et al. 2006).

« *Un cas particulier de la désambiguïsation des sens des mots (DSM)*[4]*, qui est une tâche du traitement automatique des langues naturelles, elle s'agit de déterminer le sens d'un mot ambigu dans un contexte donné* » (Stokes, et al. 2008).

2.3.2 Étapes

La plupart des méthodes de DT comprennent 2 phases principales : (1) l'extraction des référents candidats et (2) le choix du référent correct (voir Figure 2-2).

[4] Voir section 2.4.3 pour plus d'informations sur la DSM

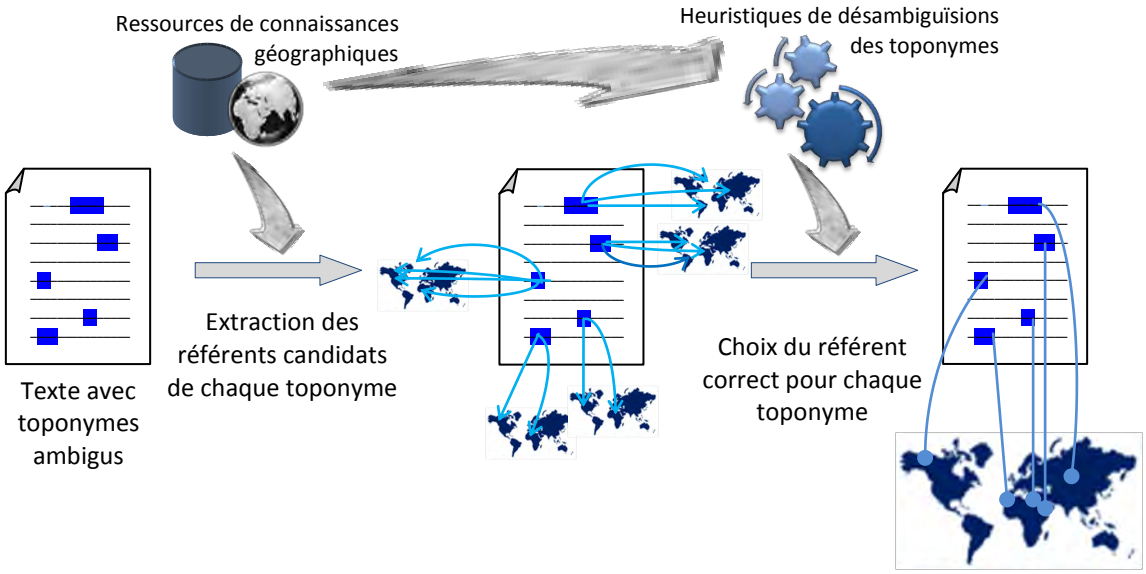

Figure 2-2. Les étapes de la désambiguïsation des toponymes

La **première phase** consiste à déterminer les référents possibles de chaque toponyme dans le texte à main. L'obtention de ces différents référents (ce qu'on appelle les référents candidats) s'appuie habituellement sur des *ressources structurées* qui contiennent des listes prédéfinies de sens pour chaque toponyme. Les gazetteers et les ontologies sont des exemples de ces ressources.

La **deuxième phase** consiste à l'application d'un ensemble d'heuristiques en vue de déterminer parmi l'ensemble des candidats le référent le plus susceptible d'être le sens voulu par le toponyme ambigu. Ces heuristiques utilisent principalement les *connaissances* fournies par le *contexte* et des *ressources externes* comme sources d'évidence.

2.3.3 Terminologie

Dans cette section nous définition des termes intrinsèques à la désambiguïsation des toponymes qui sont : contexte, connaissances, ressources. Plus de détails sur ces éléments se trouvent au chapitre suivant.

2.3.3.1 Le contexte

Le contexte est le texte en langue naturelle où le toponyme à résoudre apparaît. Le contexte est la source d'évidence principale et intuitive dans les méthodes de DT. Les toponymes du contexte sont des informations souvent utilisées pour résoudre un toponyme ambigu du même contexte (voir Section 3.3 pour des informations plus amples).

2.3.3.2 Connaissances

Une *connaissance* –dans le contexte de la DT– est toute information qui peut aider à l'association des toponymes avec leurs référents correctes. Les connaissances peuvent être internes c.-à-d. en provenance du contexte, ou externe en provenance de sources hormis le contexte (voir Section 3.5).

2.3.3.3 Ressources

Toute source de connaissance hormis le contexte est appelée *ressource*.

Le Tableau 2-2 donne quelques exemples de ressources et les connaissances qu'ils fournissent (voir Section 3.6).

Tableau 2-2. Exemples des ressources utilisées dans les méthodes de DT et les connaissances qu'ils fournissent

Ressources	Connaissances
Gazetteers, dictionnaires, ontologies	Relations coordonnées spatiales définitions
corpus	Cooccurrences fréquences d'usage

2.3.4 Applications

Nous avons discuté dans le chapitre précédant l'utilité de la désambiguïsation des toponymes dans le domaine du data mining spatial. Cependant, le DMS n'est pas le

seul champ d'application de la DT, cette dernière est une technique utile dans plusieurs applications dans multiples domaines.

Dans cette section nous présentons quelques applications de la désambiguïsation des toponymes.

2.3.4.1 Indexation géo-spatiale des documents textuels

En se basant sur l'indexation et la recherche par mots clés seulement, la requête spatiale « chercher des articles à propos de Constantine » va récupérer tous les documents qui contiennent des occurrences du mot « Constantine » quelque soit la localisation géographique de Constantine (Constantine de l'Algérie, Constantine des États-Unis…).

Cependant, l'indexation spatiale des documents –dont la désambiguïsation des toponymes se trouve parmi ses techniques principales– permet le regroupement ou le raffinement des résultats de la requête préalablement mentionnée selon la localisation géographique de Constantine.

En outre, l'indexation spatiale permet aussi de récupérer des documents qui ne mentionnent pas explicitement Constantine mais plutôt ils contiennent des toponymes qui représentent ses communes comme par exemple Zighoud Youcef, Al-Khroub…etc. Un tel résultat de recherche est impossible à obtenir par l'indexation classique basée sur les mots clés. Voir Section 2.4.1 pour plus d'informations sur ce sujet.

Exemple réel

La société MetaCarta fourni des services d'indexation spatiale des pages web (MetaCarta, Inc 2008) ; et dernièrement son site web a mis au point le service GeoSearch News[5] qui est un service de recherche dans les informations de l'actualité en combinant les mots clés et les noms des lieux (Voir Figure 2-3).

[5] http://geosearch.metacarta.com

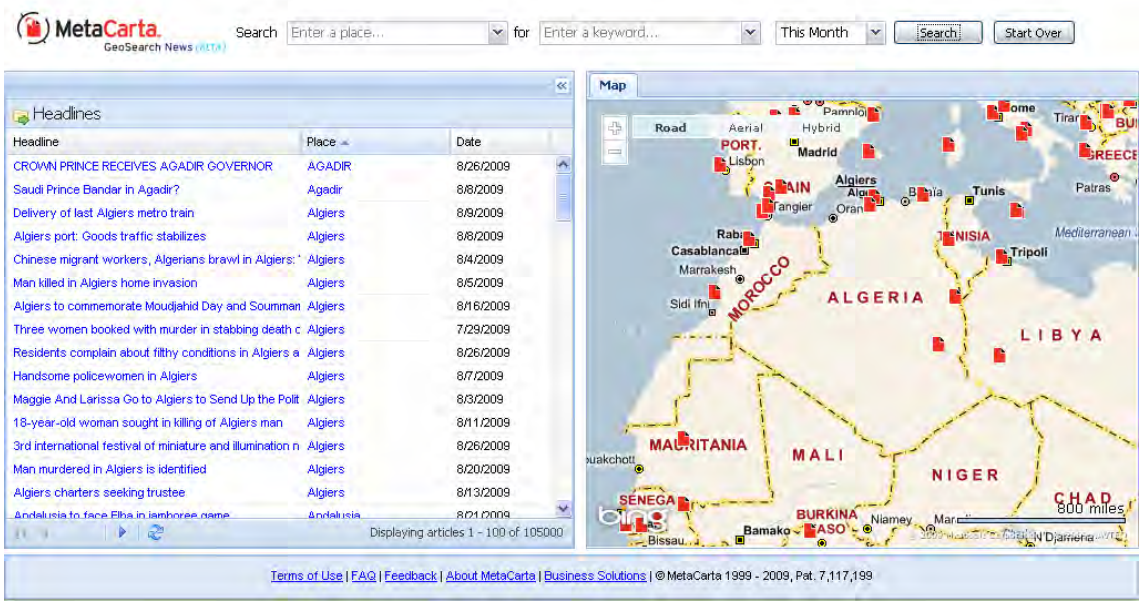

Figure 2-3. La page web GeoSearch News de MetaCarta: Recherche géo-spatiale dans l'actualité du monde

2.3.4.2 Navigation géo-spatiale

Il s'agit d'étiqueter des collections de documents textuels qui se trouvent soit dans le web ou dans les bibliothèques numériques avec les toponymes qu'ils renferment, puis, les afficher sur une carte géographique pour permettre une navigation avec une dimension géo-spatiale. Cela facilite le parcours des documents qui mentionnent le même emplacement géographique.

Par exemples, si un article de presse contient le toponyme Mila. Il sera estampé dans la carte dans Mila, mais cela après la désambiguïsation des toponymes qui décide s'il s'agit de Mila>Algérie ou Mila>Northumberland>Virginie>États-Unis.

Exemples réels

1. La Figure 2-5, (voir Page 34) montre le site Google Maps[6] qui fourni une navigation géo-spatiale dans les articles de Wikipedia[7].
2. Le site AuthorMapper[8] permet une navigation géo-spatiale dans la bibliothèque numérique de Springer[9] selon les lieux des universités des auteurs (voir Figure 2-4).

[6] http://maps.google.com
[7] http://www.wikipedia.org
[8] http://www.authormapper.com
[9] http://www.springer.com

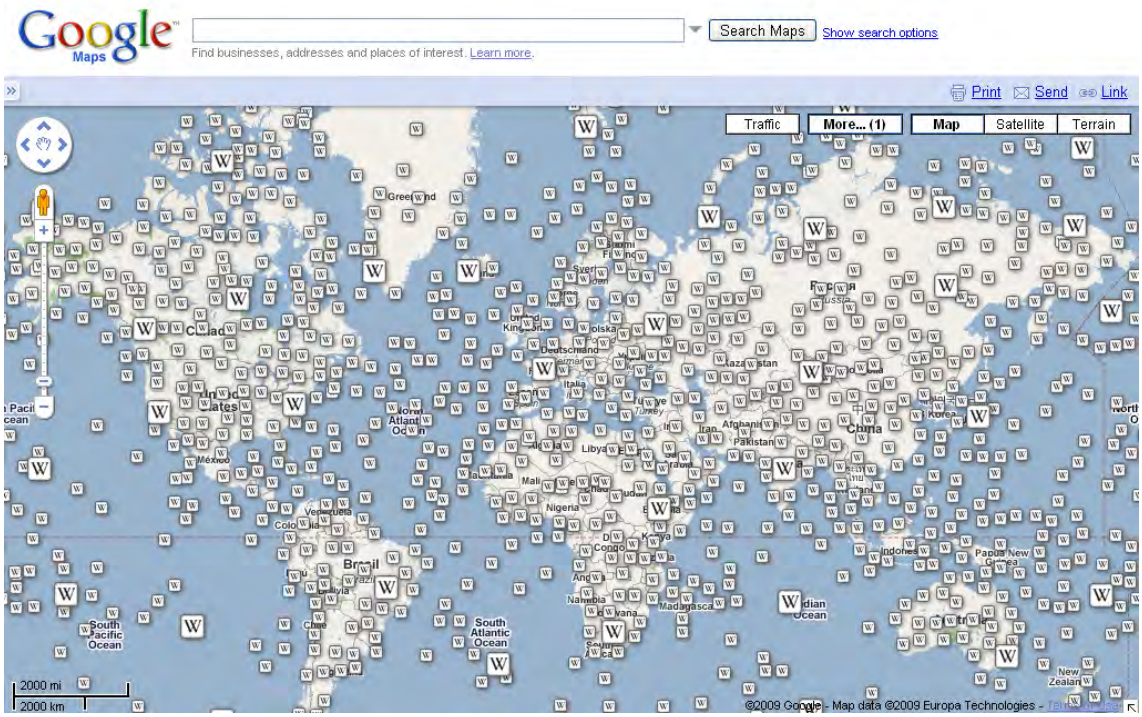

Figure 2-5. Naviguer dans les articles de Wikipedia à travers Google Maps

Figure 2-4. AuthorMapper: navigation géo-spatiale dans la bibliothèque Springer

2.3.4.3 Analyse visuelle des évènements

Figure 2-6. Biocaster: suivie des éclosions des maladies dans le monde

Il s'agit de projeter les évènements extraits du texte dans une carte selon l'endroit où se sont passés. Cela permet une analyse rapide des évènements rapportés dans un grand ensemble de documents textuels. Ça aide par exemple à détecter les évènements identiques, les suivre (c'est-à-dire ce qui s'est passé ensuite dans le même endroit), et les regrouper.

Exemples réels

1. Biocaster[10] (voir Figure 2-6) est un système de surveillance mondiale de la santé qui sert à détecter et à suivre les éclosions de maladies infectieuses à partir d'une analyse continue des documents signalés dans plus de 1700 flux RSS. Le système visualise les évènements de maladies dans Google Maps après leur géo-localisation (Collier, et al. 2008).

[10] http://biocaster.nii.ac.jp

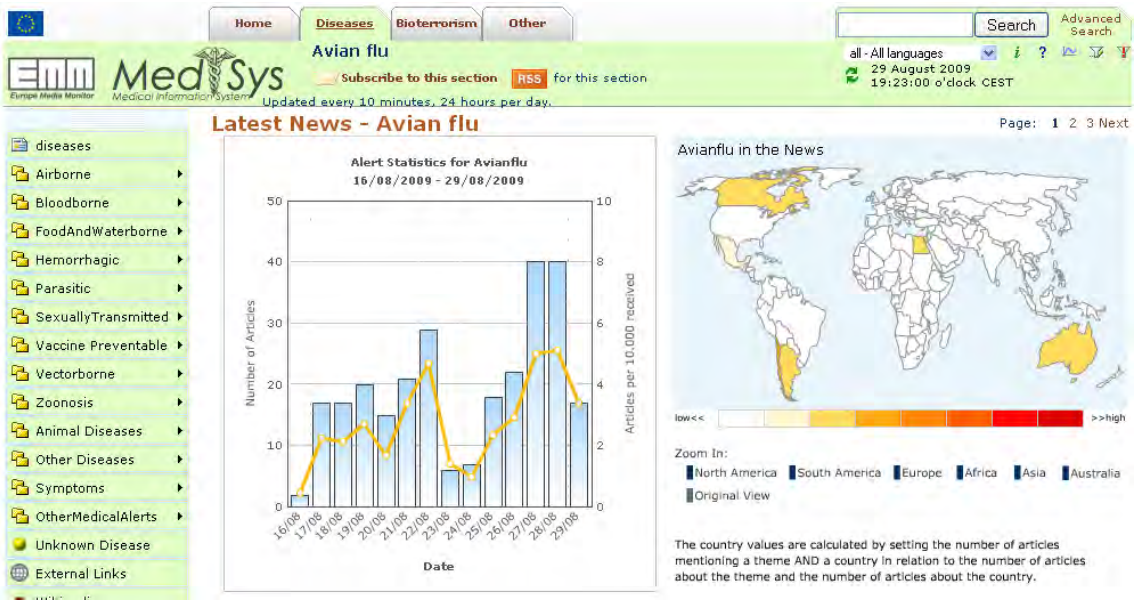

Figure 2-7. La page du service MedISys : Système d'analyse des informations médicales

2. Le site de EMM[11] (Europ Media Monitor) fournit un ensemble de services d'analyse des évènements écrits en plusieurs langues par leur visualisation sur la carte ou en utilisant des graphes de statistique. La Figure 2-7 représente la page MedISys : le service d'analyse des informations médicales.

2.4 Domaines en relation avec la désambiguïsation des toponymes

La désambiguïsation des toponymes est un domaine qui relie l'espace et le texte (Leidner 2007). Conséquemment, ses techniques sont issues principalement de deux disciplines qui sont le traitement automatiques des langues naturelles (TALN) qui s'occupe du traitement des données textuelles et les systèmes d'informations géographiques (SIG) qui s'occupent du traitement des données spatiales (voir Figure 2-8). Par ailleurs, la désambiguïsation des toponymes une tâche importante dans plusieurs domaines à savoir la recherche d'information géographique et l'extraction d'information.

[11] http://emm-labs.jrc.it

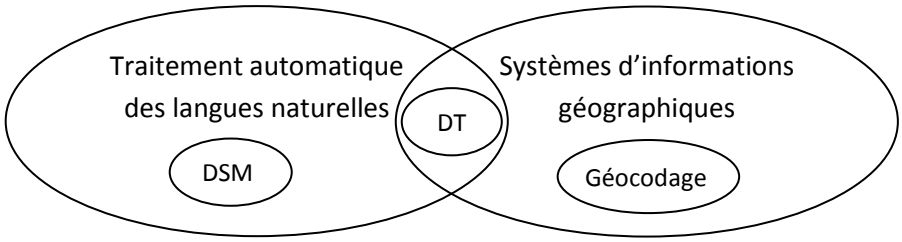

Figure 2-8. Position de la DT par rapport à d'autres domaines

Dans cette section, nous faisons un tour d'horizon sur les domaines en relation avec la désambiguïsation de toponymes.

2.4.1 Recherche d'information géographique

2.4.1.1 La Recherche d'information

Une définition classique de la recherche d'information (RI) (Rijsberg 1979) est : la discipline qui fournit des techniques d'indexation de texte et des mécanismes de recherche.

Un problème typique de la recherche d'information est de sélectionner les documents pertinents parmi une collection de documents en fonction de la requête de l'utilisateur. Cette requête est souvent sous forme de quelques mots-clés décrivant l'information voulue(Han et Kamber 2006).

Contrairement aux systèmes de gestion de bases de données (SGBD), qui mettent l'accent sur la recherche et le traitement des données structurées comme les bases de données relationnelles, la recherche d'information concentre sur la recherche et l'organisation d'informations non structurées, particulièrement les documents textuels(Han et Kamber 2006).

La recherche d'information a deux procédures principales : l'*indexation* et la *recherche*. Au temps de l'*indexation*, une collection de documents est traitée document par document et les *termes clés* de chaque document sont extraits puis stockés dans un *index*. Au temps de la *recherche*, un utilisateur encode un besoin

d'information dans une *requête*, qui est analysée par le *système de recherche*. Ce dernier sélectionne les documents dont leurs termes clés correspondent aux termes clés de la requête, et une fonction de classement classe les documents en ordre décroissant de pertinence à l'égard de la requête (Leidner 2007).

2.4.1.2 La recherche d'information avec une dimension géographique

L'espace est une dimension très intuitive pour la recherche d'information, une étude faite sur le moteur de recherche Excite[12] a montré que 18.6% des requêtes sont liées à la géographie, et 79.5% des requêtes géographiques contiennent des toponymes(Sanderson et Kohler 2004). Le problème ici est que les systèmes de RI classiques traitent les termes géographiques, entre autre les toponymes, comme tous les autres termes.

La *recherche d'information géographique* (RIG) est un nouveau domaine, d'abord décrit et baptisé par Ray Larson(1996)(Hill 2006). La RIG diffère de la RI par la reconnaissance et la modélisation explicite de l'espace dans le cadre des procédures d'indexation et de recherche d'information (Leidner 2007). Dans un système de RIG, non seulement les termes clés qui sont indexés mais aussi les termes géographiques avec leurs positions unique dans la Terre appelées *empreintes spatiales* (spatial footprint). La recherche dans ce cas, est basée sur la comparaison de l'empreinte spatiale d'une requête avec les empreintes spatiales des documents. Généralement, la comparaison n'est pas exacte, mais elle est basée plutôt sur un certain degré de chevauchement.

La Figure 2-9 montre le chevauchement de l'empreinte spatiale d'une requête géographique et les empreintes spatiales de quatre documents. Les documents A, B, C illustrés dans cette figure sont pertinents pour la requête, tandis que D ne l'est pas.

[12] http://www.excite.com

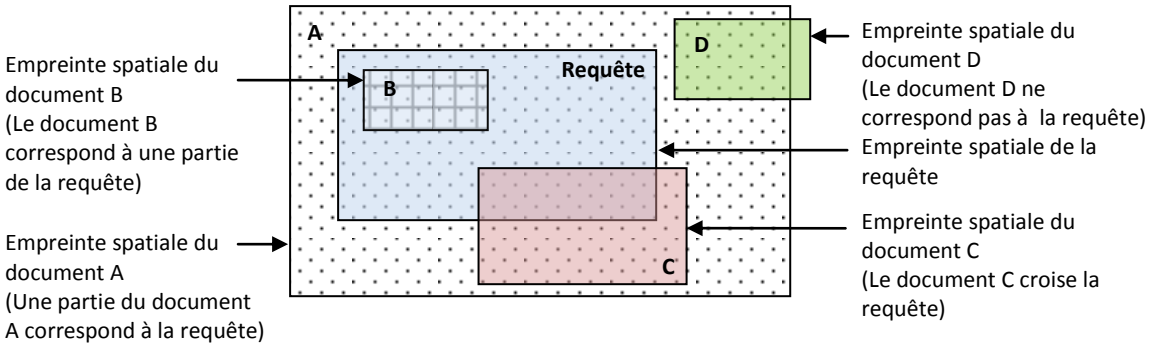

Figure 2-9. Les différents types de chevauchements entre l'empreinte spatiale d'une requête et les empreintes spatiales des documents

La création des empreintes spatiales des documents passe essentiellement par deux étapes qui sont : l'identification des toponymes dans le texte puis la désambiguïsation des toponymes.

La désambiguïsation des toponymes est donc une tâche d'une importance primordiale dans le processus de la recherche d'information géographique. Elle est appliquée au niveau de la *recherche* pour désambiguïser les toponymes de la requête, et au niveau de l'*indexation* pour désambiguïser les toponymes des documents textuels (voir Figure 2-10).

2.4.2 Extraction d'information

L'*Extraction d'Information* (EI) est le nom donné à tout processus qui sert à identifier et à classifier –à partir d'un ensemble de classes prédéfinies– les instances des noms et des relations qui se trouvent dans des documents textuels (Cowie and Lehnert 1996). Elle peut être définie aussi comme la transformation des textes en langage naturel (comme les articles de presse, les brevets, les pages web, etc.) en des représentations structurées prédéfinies. Une fois extraites, les informations peuvent ensuite être stockées dans des bases de données pour être interrogées, analysées, fouillées, etc.(Gaizauskas, et al. 1997).

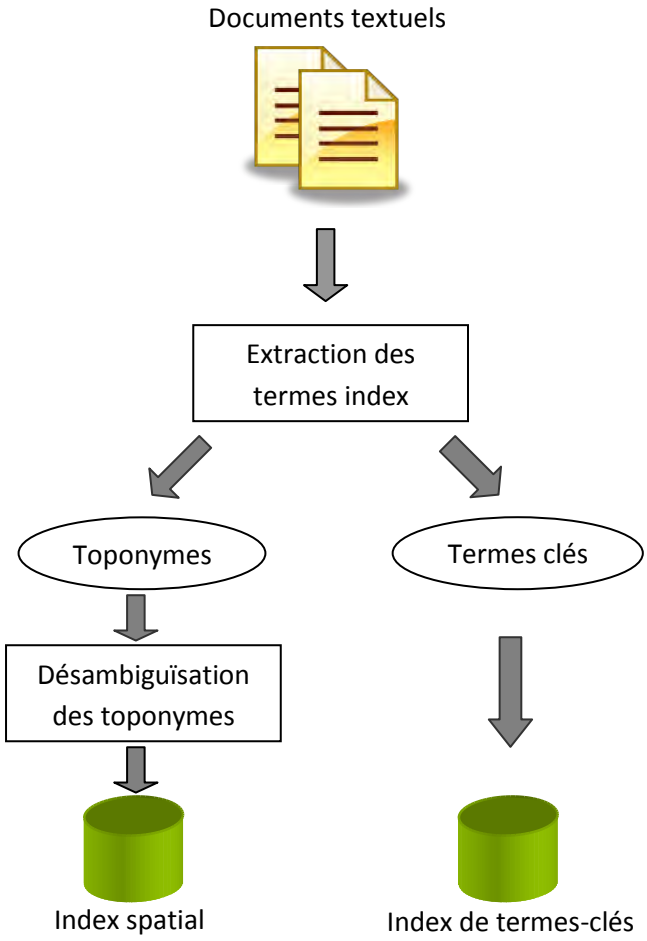

Figure 2-10. Pipeline spatial dans la procédure d'indexation dans un système de recherche d'information géographique

La figure 1-6 (voir Chapitre 1, page 30) illustre la création d'une base de données géographiques des évènements à partir du texte, a travers l'extraction d'information.

Conceptuellement, l'EI englobe trois sous-tâches: la *reconnaissance des entités nommées*, la *désambiguïsation des entités nommées*, l'*extraction de relations* (Bunescu 2007). Ces opérations sont décrites brièvement dans les sous-sections suivantes.

2.4.2.1 Reconnaissance des entités nommées

La reconnaissance des entités nommées (REN) (Chinchor 1998) consiste à identifier dans le texte les mentions des noms propres, des expressions de temps, et des expressions numériques, comme le montre le Tableau 2-3.

Tableau 2-3. Catégories des entités nommées selon (Chinchor 1998)

Catégories des entités nommées	Sous catégorie
Noms des entités (Noms propres)	Personne Organisation Toponyme
Expressions temporelles	Date temps
Expression numériques	Expression monétaires Pourcentage

Exemple

Dans la phrase suivante: « Le prophète Mohamed est né le 12 Rabi`a al Awal à La Mecque », le système de reconnaissance des entités nommées doit identifier 3 entités nommées : « Mohamed » autant qu'un nom de personne, « 12 Rabi`a al Awal » comme une date, et « La Mecque » comme un nom de lieu (un toponyme).

2.4.2.2 Désambiguïsation des entités nommées

L'identification des entités nommées, et en particulier celles de la première catégorie (c.à.d. les noms propres associés aux entités) n'est pas généralement suffisante pour obtenir des informations consolidables à partir du texte. Cela est dû à l'ambigüité qui est un caractère inhérent aux noms dans la langue naturelle. Un type de cette ambiguïté consiste à associer un nom à plusieurs entités. Par exemple, dans les phrases ci-dessous « Al Akkad » se réfère à deux personnes différentes, ce qui provoque une ambiguïté dans les informations extraites.

Al Akkad est le réalisateur des films « Le message » et le « Le lion du désert ».

Al Akkad est l'auteur du livre « génie de Mohamed ».

La *désambiguïsation des entités nommées* (Bunescu 2007) est la tâche qui permet l'identification de l'entité qui corresponde à une occurrence d'un nom dans un document textuel, cette tâche est un cas spécifique de la désambiguïsation des sens des mots (Section 2.4.3). Par exemple, en appliquant la désambiguïsation des entités nommées sur le nom « Al Akkad » dans les deux phrases ci-dessus ; Al Akkad de la première phrase est associé à l'entité : Moustafa Al Akkad, par contre celui de la deuxième phrase est associé à l'entité : Mahmoud Al Akkad.

La désambiguïsation des entités nommées est une sous tâche importante dans l'extraction d'information, en particulier, lorsque les informations extraites d'un certain document doivent être intégrées avec des informations sur la même entité en provenance d'autres documents ou de sources externes.

2.4.2.3 Extraction de relations

Une fois les entités nommées ont été correctement identifiées puis désambiguïsées, une étape supplémentaire dans l'EI est de trouver des relations prédéfinies entre ces entités. Par exemple dans la phrase « Al-Khawarizmi est un mathématicien originaire de Khiva, né vers 783 », un système conçu pour extraire les relations entre les personne et les lieux doit identifier la relation *né-à* qui relie le nom de personne Al-Khawarizmi et le toponyme Khiva. C'est le résultat de cette étape qui permet de construire des bases de données qui contiennent une description pour chaque entité extraite.

Les lieux géographiques sont parmi les entités extraites. Et le fait de les relier avec d'autres informations permet de construire des bases de données géographiques comme il a été discuté dans la Section 1.6.3 du le chapitre précédant.

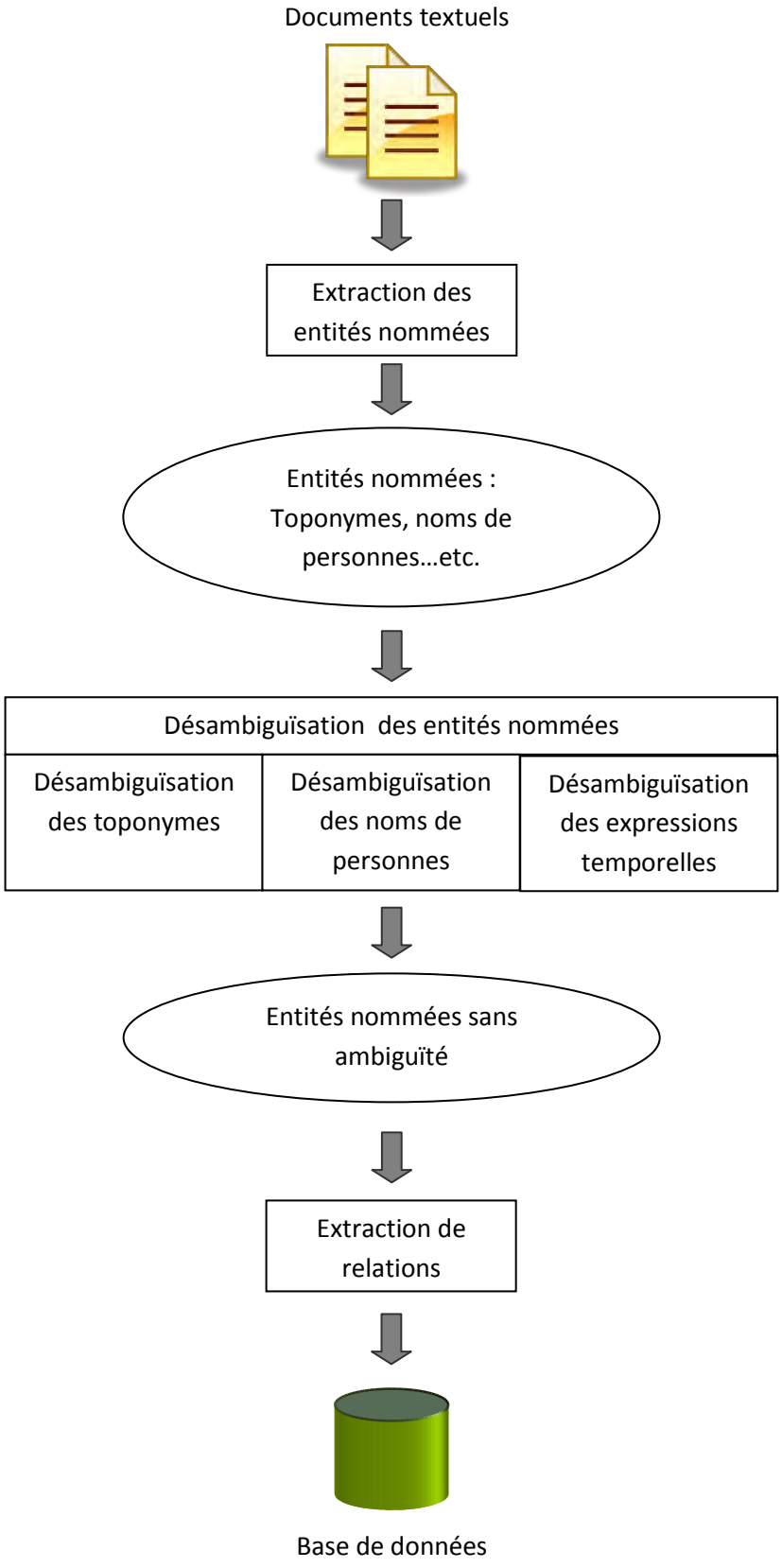

Figure 2-11. Le processus d'extraction d'information avec la tache de désambiguïsation des toponymes

2.4.2.4 Relation entre l'extraction d'information et la désambiguïsation des toponymes

La *désambiguïsation des toponymes* peut être considérée comme une spécification de la tâche de *désambiguïsation des entités nommées*. Par conséquent, c'est l'une des étapes importantes dans le processus d'extraction d'information cela est dans le cas où des toponymes ambigus sont parmi les informations extraites. La Figure 2-11 (voir Page 34) montre la position de la désambiguïsation des toponymes dans le processus de l'EI.

2.4.3 Désambiguïsation des sens des mots

2.4.3.1 Description du problème

La *désambiguïsation des sens de mots* (DSM)[13] est définie comme : la tâche de l'attribution automatique du sens le plus approprié à un mot polysémique[14] dans un contexte donné (Sinha et Mihalcea 2007).

Formellement, supposons que T est une portion de texte c.-à-d. une séquence de mots ($m_1, m_2 \ldots m_n$) ; et $Sens_D\ (m_i)$ est l'ensemble des sens $(s_1, s_2, \ldots, s_n)$ des mots m_i encodés dans un dictionnaire *D*. On peut décrire la DSM comme la tâche d'attribuer les sens s_i à l'ensemble ou certains des mots de *T*. Cela revient à identifier une fonction *F* qui associe les mots vers leurs sens. Tels que $F(i) \subseteq Sens_D\ (m_i)$, où $F(i)$ est un sous ensemble des sens du mot m_i qui sont appropriées dans le contexte *T*. La fonction *F* peut associer plus qu'un sens à chaque mot $m_i \in T$, mais en général, seulement le sens le plus approprié est sélectionné, c.-à-d. $|F(i)| = 1$ (Navigli 2009).

2.4.3.2 Relation de la DSM avec la désambiguïsation de toponymes

Les toponymes sont un type spécial de mots. Certain auteurs comme (Stokes, et al. 2008) considèrent la DT comme un cas spécifiques de la DSM où les mots à

[13] Traduction directe du terme anglais Word sens disambiguation (WSD). En effet, nous n'avons pas trouvé un terme conventionnel en français. Néanmoins, il existe des traductions variées comme : résolution de polysémie, désambiguïsation sémantique et désambiguïsation syntaxique.
[14] Un mot polysémique est un mot qui possède plusieurs sens.

désambiguïser sont les toponymes et leurs sens sont les lieux physiques que l'auteur du texte a fait entendu en les mentionnant.

Un autre point de vue dit que la DT est une étape au-delà de la DSM (Li, et al. 2003), car les méthodes de cette dernière ne peuvent résoudre que l'ambiguïté de type géo/non-géo, c.-à-d. elles peuvent déterminer si un nom est un toponyme ou non, mais elles ne sont pas en mesure de lui associer le lieu physique à lequel il se réfère. Et c'est ça le rôle de la DT.

La recherche d'information géographique et l'extraction d'information (discutés ci-dessus) sont des domaines qui utilisent la désambiguïsation des toponymes comme une tâche dans leurs systèmes. Par contre, la désambiguïsation des sens des mots est un domaine que la DT inspire beaucoup de techniques et de notions, telles que : les phases principales et les opérations de base (voir Section 2.3.2).

Sauf que, les méthodes de DSM emploient plus de ressources, et quand au contexte, il est représenté par la quasi-totalité des mots du texte et non pas par les toponymes. Le Tableau 2-4 résume les principales différences entre la DSM et la DT.

Tableau 2-4. Comparaison entre la Désambiguïsation des Sens des Mots et la Désambiguïsation des Toponymes

Désambiguïsation des Sens des Mots	Désambiguïsation des Toponymes
Trouver le sens voulu par la mention du mot dans un contexte donné	Trouver le référent voulu par la mention du toponyme dans un contexte donné
Concerne tous les types de mots : les noms, les verbes, les adjectif…	Concerne seulement les noms des lieux
Les ressources utilisées sont : les dictionnaires numériques, les thésaurus, les ontologies, les corpus	Les ressources utilisées sont : les gazetteers, les ontologies, les corpus, le Web
Le contexte est représenté par tous les mots	Le contexte est représenté par les toponymes

2.4.4 Géocodage

Le Géocodage est le nom communément donné au processus de conversion des adresses postales aux coordonnées de latitude et longitude, ou d'autres systèmes universels de coordonnées. Le Géocodage permet à n'importe quelle base de données contenant des adresses, de contribuer à un système d'information géographique (Leidner 2007).

Le terme géocodage dans ce sens est utilisé beaucoup plus par la communauté des SIG, mais pratiquement, il est utilisé aussi par d'autres communautés avec un sens plus large qui ne se limite pas aux adresses postales, à savoir, géocoder les montagnes, les rivières, les numéros de téléphones, les noms de domaines…etc.

On peut dire donc que la désambiguïsation des toponymes est un géocodage (avec son sens large) dont les données à géocoder sont des toponymes. Et d'un autre point de vue, la DT est une technique alternative au géocodage des adresses postales (le sens restreint), car les toponymes sont des données non structurées quand a les adresses postales sont des données structurées.

2.4.5 Géo-référencement

Le géo-référencement est le terme qui désigne l'opération de relier les informations aux emplacements géographiques, il s'agit d'établir une relation entre les informations (ex. documents, bases de données, cartes géographiques, images) et les emplacements géographiques à travers les toponymes ou les codes de lieux (ex. les codes postaux) ou le référencement géo-spatiale (ex. les coordonnées longitude et la latitude) (Hill 2006). Le Tableau 2-5 donne quelque exemple des systèmes de géo-référencement.

Tableau 2-5. Quelques systèmes de géo-référencement couramment utilisés

Système de géo-référencement	Type	Domaine de couverture sans ambiguïté	Exemple
Adresses Postales	Nominal	Globe	Université Mentouri, Route Ain El Bay, Constantine, Algérie
Code postal	Nominal	Pays	L'aéroport d'Alger : 16101
Latitude/longitude	Métrique	Globe	Tassili de Hoggar : 26° 19' 60" Nord 5° 0' 00" East
UTM	Métrique	Globe	Oran : x : 713981.9 y : 3952997.6 Zone : 30 Hémisphère du nord

On peut dire que la désambiguïsation des toponymes est un géo-référencement des documents textuels.

Le géocodage, le géo-référencement et la désambiguïsation des toponymes ont des sens similaires mais avec quelques différences. Le Tableau 2-6 présente une comparaison entre ces 3 tâches.

Tableau 2-6. Comparaison entre le géo-référencement, le géocodage et la désambiguïsation des toponymes

Tâche		Type d'information à relier à l'espace	Façon de représenter l'espace
Le géo-référencement		Tous les types d'informations	Tous les types de représentation de l'espace
Le géocodage	Sens restreint	Adresses postales	Coordonnées spatiales
	Sens large	Tous les types d'informations	Coordonnées spatiales
La désambiguïsation des toponymes		Toponymes apparaissant dans le texte	Représentation spatiale non ambigüe

2.5 Conclusion

Dans ce chapitre nous avons défini la tâche de désambiguïsation des toponymes, et nous avons montré ces différentes applications et sa position par rapport à d'autres domaines.

En conclusion, on dit que la tâche de désambiguïsation des toponymes est multidisciplinaire dans ses notions de base, ses techniques et aussi dans ses applications. Cette multidisciplinarité serait plus évidente en présentant les différents travaux dans ce domaine, ce qui est l'objet du chapitre suivant.

Chapitre 3

État de l'art

3.1 Introduction

Les chapitre précédent ont permis d'avoir une vue globale sur la désambiguïsation des toponymes, ils ont donc donné des réponses à deux questions principales dans la recherche qui sont le « quoi » et le « pourquoi » mais ils n'ont pas répondu à une troisième question de la même importance qui est le « comment ».

« Comment désambiguïser les toponymes ? » c'est donc l'objet de ce chapitre qui répond à cette question en présentant l'état de l'art des méthode de désambiguïsation.

Malgré le fait que les méthodes de désambigüisation des toponymes sont très différentes dans l'esprit (Leidner 2007)(dû à la nature multidisciplinaire), mais ils ont des éléments en commun que leur présence est incontournable et indispensable dans toute méthode de DT. Ce chapitre s'articule selon ces éléments qui sont le contexte, les heuristiques, les connaissances et les ressources.

Nous commençons d'abord dans la section suivante par une brève comparaison de notre point de vue avec celui de Leidner (2007) qui fût le premier à présenter un état de l'art élargie de la désambiguïsation des toponymes autant qu'une tâche indépendante de la DSM et de la REN. En suite nous présentons une synthèse des différents travaux dans la DT en faisant des comparaisons et des classifications des méthodes selon la présentation du contexte, les heuristiques, les connaissances et les ressources.

3.2 Les méthodes

Les méthodes de désambiguïsation des toponymes comprennent deux phases principales qui sont : l'obtention des référents candidats d'un toponyme, et le choix du référent correct (voir Section 2.32) mais elles se distinguent principalement par la deuxième phase.

Étant donné que la littérature de la désambiguïsation des toponymes est dispersées à travers plusieurs disciplines (RI, TALN (DSM, IE), SIG) Un simple examen des méthodes peut donner l'impression qu'elles sont complètement différentes les unes des autres, notamment, dans la deuxième phase. Cependant, un examen plus approfondi permettra d'en tirer des facteurs en commun.

Leidner (2007) –dans le cadre de sa thèse[1] – a analysé une dizaine de méthodes de l'état de l'art de DT[2] et il a remarqué que plusieurs *moyens d'évidence* et *sources de connaissances (ressources)* se reproduisent dans des méthodes différentes. En outre, il a résumé les *moyens d'évidence* en dix-sept heuristiques et connaissances de base qu'il a ensuite regroupé dans une taxonomie distinguant entre les connaissances (ou les heuristiques) linguistiques et les connaissances (ou les heuristiques) du monde.

Après avoir analysé les méthodes de DT présentées dans l'état de l'art élaboré par Leidner (2007) et aussi d'autres méthodes plus récentes, nous avons reformulé les moyens d'évidence de Leidner (2007) mais selon notre point de vue et avec notre propre classification. Contrairement à Leidner, nous distinguons entre les heuristiques et les connaissances[3], et sur la base de cette distinction nous considérons la majorité des moyens d'évidence que Leidner a tiré comme des heuristiques ; nous élaborons donc deux taxonomies différentes pour les heuristiques et les connaissances.

Ainsi, les méthodes de DT peuvent être vues comme des **heuristiques** (des algorithmes) qui servent à désambiguïser les toponymes ambigus apparaissant dans un certain **contexte textuel,** en manipulant des **connaissances** extraites de ce contexte et des **ressources** externes. La Figure 3-1 illustre le rôle de ces différents éléments.

[1] La thèse de Leidner (2007) est la première thèse qui a adressé la désambiguïsation des toponymes autant que tâche indépendante. Cette thèse est une référence de base dans ce domaine et elle est publiée aussi comme livre.

[2] Les travaux analysés dans (Leidner 2007) sont publiés entre 1999 et 2006.

[3] Cette distinction n'est pas parfois évidente car il existe des connaissances qui sont inhérentes à certaines heuristiques.

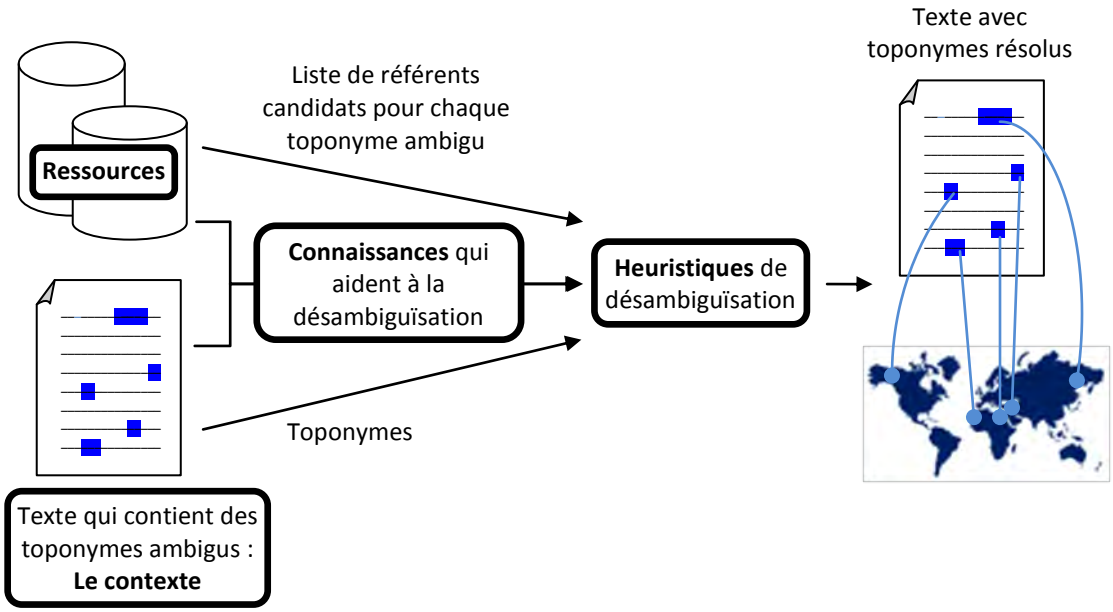

Figure 3-1.Les éléments principaux des méthodes de désambiguïsation des toponymes

Il est donc clair que les méthodes de désambiguïsation des toponymes sont toutes composées de quatre éléments principaux qui sont : le contexte, les heuristiques, les connaissances, et les ressources. Dans les sections suivantes nous présentons un état de l'art des méthodes de DT organisé selon ces quatre composants. En lisant les sections de ces éléments le lecteur (selon ces besoins) ne se trouverait pas contraint de suivre leur ordre de rédaction.

3.3 Le contexte

Le contexte est le texte en langue naturelle qui contient le(s) toponyme(s) à désambiguïser. Il est donc naturelle que l'opération de manipuler le contexte soit présente dans toute méthode de DT.

Deux types d'informations qui peuvent être tirées du contexte :

1. Les toponymes (ou d'autres mots pertinent) qu'il contient,
2. Des informations statistiques ou linguistiques sur le toponyme à résoudre tels que la position dans le texte, la fréquence d'occurrence, … etc. (voir Section 0 pour plus d'informations sur les connaissances)

L'utilisation du contexte pour associer les mots à leurs sens est une idée intuitive dont l'origine est dans le domaine de désambiguïsation des sens des mots. Cependant, le contexte dans les méthodes de désambiguïsation des toponymes est représenté généralement par les toponymes qu'il contient et non pas par tous les mots du texte.

La **taille du contexte** dans les méthodes de DT varie de quelques toponymes autour du toponyme ambigu jusqu'à tous les toponymes du texte d'un document.

Supposons qu'un document contient le texte ci-dessous[4] (Les toponymes sont soulignés).

> « La ville de La Mecque, se situe à l'ouest de l'Arabie saoudite, sur les pentes de la chaîne d'Al-Sarawat, entre les massifs du Hedjaz et de l'Asir, plus précisément dans la vallée de l'Oued Ibrahim au pied de collines de 60 m à plus de 500 m de hauteur. Le port de Djeddah n'est distant que de 65 kilomètres.
>
> La partie est de la ville se situe entre 194 et 310 m au-dessus du niveau de la mer. La partie ouest à 400 m, se caractérise par la présence de certains monts qui peuvent atteindre jusqu'à 900 m d'altitude comme le mont Jabal Tarki (qui est la plus haute montagne de La Mecque) et le Jabal Khandama qui culmine à 914 m. »

Le Tableau 3-1 illustre les différentes tailles du contexte, en supposons que le toponyme « Asir » (dans le texte ci-dessus) est le toponyme à désambiguïser.

Tableau 3-1. Les différentes tailles du contexte

Taille du contexte	Explication	Exemple
n-grams	une séquence de n toponymes, y compris le toponyme cible (le toponyme à désambiguïser)	Hedjaz, **Asir**, Oued Ibrahim (n=3)
Fenêtre (taille ±n)	Une fenêtre de taille ±n veut dire n toponyme à droite et n toponyme à gauche du mot cible.	Al-Sarawat, Hedjaz, **Asir**, Oued Ibrahim, Djeddah (n=2)

La suite du tableau est dans la page suivante

[4] Ce texte est un extrait de : La Mecque. (2009, août 25). *Wikipédia, l'encyclopédie libre*. Page consultée le 10:21, septembre 6, 2009 à partir de http://fr.wikipedia.org/w/index.php?title=La_Mecque&oldid=44178292.

Phrase	Tous les toponymes de la phrase qui contient le toponyme cible.	La Mecque, Arabie saoudite, Al-Sarawat, Hedjaz, **Asir**, Oued Ibrahim.
Paragraphe	Tous les toponymes du paragraphe qui contient le toponyme cible.	La Mecque, Arabie saoudite, Al-Sarawat, Hedjaz, **Asir**, Oued Ibrahim, Djeddah, Djeddah.
Discours	Tous les toponymes du texte qui contient le toponyme cible.	La Mecque, Arabie saoudite, Al-Sarawat, Hedjaz, **Asir**, Oued Ibrahim, Djeddah, Djeddah, Jabal Tarki, La Mecque, Jabal Khandama

Buscaldi et Rosso (2008c) ont comparé la précision et le recall[5] de deux heuristiques de DT en utilisant des tailles différentes de contexte. DC représente l'heuristique de la densité conceptuelle (Buscaldi et Rosso 2008a) (Voir H9 H6ci-dessous), et MAP représente l'heuristique de (Smith and Crane 2001) (Voir H6 ci-dessous).

Les graphes de la Figure 3-2 (Construit à partir des données fournies par (Buscaldi and Rosso 2008c)) montrent que le recall (le pourcentage des toponymes résous

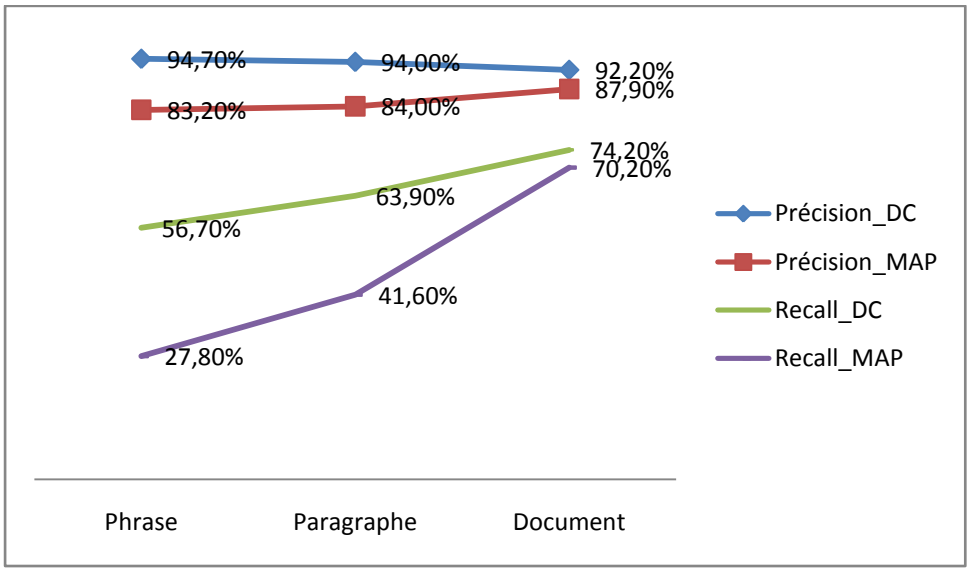

Figure 3-2. L'effet de la taille du contexte sur la performance de désambiguïsation des toponymes

[5] La précision et le recall sont des mesures de performance des méthodes de désambiguïsation des toponymes. Voir le chapitre suivant pour plus d'informations sur ces mesures.

correctement par rapport à tous les toponymes) augmente avec des valeurs significatives en augmentant la taille du contexte. Tandis que la précision diminue dans une méthode et augmente dans une autre mais avec des valeurs non importantes.

3.4 Les Heuristiques

3.4.1 Qu'est ce qu'une heuristiques de désambiguïsation de toponymes

Nous définissons les heuristiques de désambiguïsation des toponymes comme les algorithmes qui manipulent les connaissances disponibles dans le contexte du toponyme ambigu ou extraites des différentes ressources en vue d'émerger parmi un ensemble de candidats, le référent le plus susceptible d'être le sens voulu par le toponyme ambigu.

3.4.2 Classification des heuristiques de désambiguïsation de toponymes

Dans les sections suivantes nous présentons une classification des heuristiques de résolution de toponymes (Bensalem et Kholladi 2009a). Nous distinguons trois classes globales : *heuristique basées sur le contexte*, *heuristiques basées les règles de préférence*, et *heuristiques complémentaires*. Généralement, une méthode de désambiguïsation combine plusieurs heuristiques dans un ordre particulier, ou dans une procédure ou formule de calcule de poids, dans le but d'augmenter le nombre de toponymes résolus.

3.4.2.1 Désambiguïsation par le contexte

Les heuristiques que nous classifions sous cette catégorie cherchent des indices de désambiguïsation dans l'environnement textuel où le toponyme ambigu apparait.

Il existe deux approches générales de la désambiguïsation par le contexte, la première consiste à désambiguïser chaque toponyme séparément des autres. Il

s'agit d'extraire des mots (toponymes ou/et termes) particuliers du document qui contient le toponyme à résoudre, puis, choisir parmi les référents candidats le référent le plus relié à ces mots. Cette relation peut être spatiale (H1), linguistique (H2, H3) ou statistique (H4).

La deuxième approche consiste à résoudre plusieurs toponymes ambigus à la fois en effectuant des calculs géométriques sur les coordonnées spatiales des référents candidats des toponymes (H5, H6), ou en appliquant d'autres algorithmes comme le cas de des heuristiques H7 et H8.

Les heuristiques de cette classe utilisent des techniques issues des domaines suivants: le Traitement Automatique des Langues Naturelles (TALN) y compris la DSM (H1, H2, H3, H4, H7, H9), les Systèmes d'Informations Géographiques (SIG)(H1, H5, H6), et la théorie des graphes (H7H8).

H1 Distance aux voisins textuels non ambigus

La résolution d'un toponyme *t1* se fait par les étapes suivantes (Leidner 2007) :

1. Envisager un contexte de taille W toponymes non ambiguë de chaque côté du t1 dans le texte.
2. Attribuer comme interprétation de t1 le référent qui est géographiquement le plus proche de tous les toponymes de W.

Il convient de noter que cette heuristique n'est plus applicable dans le cas de l'absence des toponymes non ambigus dans le texte concerné.

Cette heuristique est basée sur les calculs géométriques dans la 2ème étape donc elle nécessite l'utilisation des coordonnées géographiques de référents de chaque toponymes. Et elle est utilisée par Smith & Crane (2001).

H2 Chevauchement entre les chemins hiérarchiques des référents et le texte

Il s'agit de calculer le chevauchement (les mots identiques) entre les noms des lieux qui composent le chemin hiérarchique de chaque référent du toponyme

ambigu et ceux qui se trouvent dans le contexte. À partir d'une étude empirique, Clough (2005) a choisi un contexte de chevauchement de 2 mots à gauche et 8 mots à droite du toponyme ambigu. Une partie de notre heuristique de désambiguïsation est aussi basée sur l'intersection du chemin hiérarchique avec le contexte (Bensalem et Kholladi 2009a). Voir le chapitre suivant pour plus de détails.

Une idée similaire est de **chercher la mention du toponyme supérieure**. Si *t1* est un toponyme à résoudre, et un deuxième toponyme *t2* apparaît d'ailleurs dans le même document, tel que, l'un des référents de *t1* est situé dans l'un des référents de *t2*. Alors attribuer à *t1* ce référent situé à *t2*. Cette heuristique est utilisée par (Hauptmann and Olligschlaeger 1999) et (Pouliquen, et al. 2004), et (Li, et al. 2006).

Exemple

Dans l'exemple[6] donné dans le Tableau 3-2, le référent choisi pour résoudre le toponyme « Constantine » est « Africa>Algeria» car « Algeria » (le toponyme supérieur de Constantine[7]) existe dans le contexte ce qui a permis de lui attribuer le plus grand score de chevauchement.

Tableau 3-2. Exemple sur l'application de l'heuristique H2

Les toponymes du contexte	Skikda, Algeria, Constantine	
Le toponyme à désambiguïser	Constantine	
Les chemins hiérarchiques des référents du Constantine avec leurs scores de chevauchement avec le contexte	North and Central America>United States> Kentucky> Breckinridge county	0
	North and Central America>United States> Michigan>Saint Joseph county	0
	Oceania>Australia> Queensland	0
	Africa>**Algeria**	1

[6] Les référents de l'exemple sont extraites du Gazetteer Getty.
[7] On dit aussi que Algeria est holonyme de Constantine.

H3 L'appariement des patterns

C'est une technique appliqué dans le TALN, elle est connue habituellement sous le nom anglais « patterns matching ». Dans le domaine de désambiguïsation des toponymes il s'agit de chercher dans le texte, des modèles prédéfinis – syntaxiques et/ou lexicales – sur les expressions qui contiennent le toponyme ambigu. Une fois le pattern est détecté dans le texte, les informations qu'il contient sont comparées aux référents candidats pour choisir le plus approprié parmi eux.

Nous distinguons deux types d'heuristiques de résolution de toponymes qui utilisent les techniques d'appariement de patterns : des heuristiques qui servent à extraire des *relations hiérarchiques* (H3.1), et des heuristiques qui servent à extraire le *type de toponyme* (H3.2). Chacune de ces heuristiques est expliquées ci-dessous. Il y a des patterns qui détectent à la fois les relations hiérarchiques et le type de toponyme, comme ceux utilisés dans (Li, Srihari, et al. 2003).

H3.1 Les patterns de relation hiérarchique

Ils capturent les toponymes contigus dans le texte. Ce type de patterns peut prendre l'un des formats suivants (Leidner 2007) :

t1, *t2*

t1/t2

t1 (t2)

Si exactement le cas où l'un des référents candidats *r1* du toponyme *t1* est situé dans *r2,* tel que *r2* est l'un des référents du toponyme t2. Alors attribuer r1 à t1.

Ce format est beaucoup utilisé dans les adresses, et il peut contenir deux toponymes ou plus.

Exemples

En détectant le pattern *t1(t2)* dans la phrase ci-dessous, le toponyme ambigu Tripoli sera résolu à Tripoli>Liban au lieu de Tripoli>Libye.

> À l'époque des Omeyyades, Tripoli (Libon) devint une importante base navale.

Cette heuristique est utilisée dans : (Hauptmann and Olligschlaeger 1999), (Smith and Crane 2001), (Li, Srihari, et al. 2003), (Rauch, Bukatin and Baker 2003), (Amitay, et al. 2004).

H3.2 Les patterns de type

Si un toponyme apparaît dans le texte à coté d'un nom qui indique son type (ex. ville, capital, pays, commune…) alors éliminer les référents candidats qui ne sont pas de ce type.

Exemple

Soit la phrase suivante qui contient le toponyme ambigu Washington :

> L'état de Washington est situé dans le nord-ouest des États-Unis.

Le pattern « état de *toponyme* » permet de résoudre le toponyme Washington à Washington l'état au lieu de Washington la capitale.

Cette heuristique est utilisée dans : (Li, Srihari, et al. 2003), (Rauch, Bukatin and Baker 2003) et (Schilder, Versley et Habel 2004), (Li, et al. 2006), (Stokes, et al. 2008).

H4 Modèle de cooccurrence

Une *cooccurrence* (aussi appelée *collocation*) fait référence à des mots souvent utilisés ensemble (Zheng, et al. 2007). L'idée de base derrière cette méthode est que la distribution d'un mot dans un contexte lexical (les mots et les expressions qu'il apparaît avec) est fortement révélatrice de sa signification (Pekar, Krkoska et Staab 2004).

Le modèle de cooccurrence renferme pour chaque sens d'un toponyme (c.-à-d. référent) les toponymes (ou les mots) fréquemment apparus avec lui. Un modèle de cooccurrence est construit à partir d'un corpus (Voir 3.6.2) puis appliqué sur le texte à main pour capturer les mots du contexte et ensuite, inférer à partir de ces mots le sens le plus approprié au toponyme ambigu.

Par exemple, on peut inférer à partir d'un corpus que le terme « palais d'Alhambra » est positivement corrélé avec le toponyme « Grenade » tel que le référent de ce dernier dans ce cas est Grenade>Espagne. Alors, en appliquant ce modèle de cooccurrence sur un nouveau texte, si « Alhambra » est mentionné à proximité du toponyme « Grenade », ce dernier sera attribué à la ville de Grenade dans l'Espagne au lieu de l'état de Grenade dans l'océan atlantique.

La désambiguïsation en utilisant les modèles de cooccurrences est inspirée des méthodes de désambiguïsation des sens des mots. Cette heuristique est implémentée en utilisant un éventail de technique à savoir, l'apprentissage automatique.

Cette heuristique a été utilisée par Overell et al. (2006a, 2006b, 2007) qui ont généré un modèle de cooccurrence à partir de l'encyclopédie libre Wikipedia. Elle a été utilisée aussi par (Smith et Mann 2003) qu'ils ont utilisé une méthode d'apprentissage supervisé pour construire un modèle de cooccurrence sous forme de classificateur.

H5 Espace géométrique (polygone / distance) minimaliste

Il s'agit d'attribuer à tous les toponymes qui émergent dans le même texte les référents qui diminuent le plus les distances spatiales bilatérales, et par conséquent, ils occupent ensemble espace géométrique le plus réduit. Cette heuristique prend en compte toutes les interprétations possibles pour chaque toponyme et fait des traitements d'optimisation à l'aide de la proximité géographique comme critère.

Cette heuristique est utilisée dans (Leidner, Sinclair et Webber 2003), (Rauch, Bukatin and Baker 2003), et (Amitay, et al. 2004).

H6 Contexte géographique unifié

Consiste à la sélection dynamique d'une zone géographique selon les toponymes contenus dans le texte, et ignorer les référents qui se trouvent en dehors cette

zone. Le contexte géographique est élaboré en calculant le centroïde (barycentre) géographique des référents des toponymes mentionnés dans le document, puis éliminer tous les référents candidats qui sont situés à plus d'une certaine distance loin du centre. Dans (Smith and Crane 2001) cette distance était définie 2 écarts-types. Cela peut être considéré comme une version dynamique de H16.

H7 Le chemin le plus court entre les référents

Une heuristique utilisée dans (Stokes, et al. 2008) consiste à désambiguïser les toponymes en cherchant le chemin le plus cours entre les référents candidats dans l'arbre hiérarchique de Getty : le Thésaurus des Noms Géographique[8]. Cette idée est déjà appliquée dans les méthodes de désambiguïsation des sens de mots mais en employant WordNet.

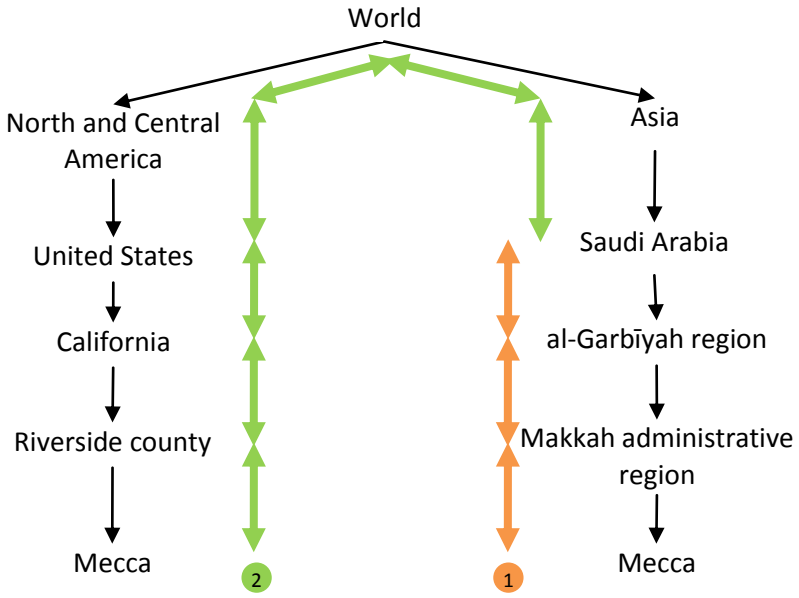

Figure 3-3. Chemins entre le toponyme ambigu Mecca et Saudi Arabia dans l'arbre hiérarchique du monde selon le gazetteer Getty : le chemin numéro 1 est le plus court car il contient 3 arc seulement.

Exemple

Soit « Mecca » le toponyme à résoudre sachant que le toponyme « Saudi Arabia » apparaît avec lui dans le même contexte. L'heuristique H7 résout « Mecca » à

[8] Getty Thesaurus of Geographic Names: http://www.getty.edu/research/conducting_research/vocabularies/tgn (dernière visite le 07/04/2009)

« Mecca>Saudi Arabia » comme c'est expliqué dans la Figure 3-3.

H8 Les nœuds de l'arbre couvrant maximum

Il s'agit de construire un graphe pondéré, où chaque nœud représente un sens d'un toponyme (un référent), et chaque arête représente une relation entre deux sens (voir Figure 3-4). Le poids représente la similarité entre chaque couple de référents. Le graphe est partiellement complet car il n'y a pas de liens entre les différents sens d'un toponyme. Les nœuds de l'arbre couvrant[9] de poids maximal (maximum weight spanning tree (MST))[10] sont les sens considérés les plus prometteurs pour les toponymes.

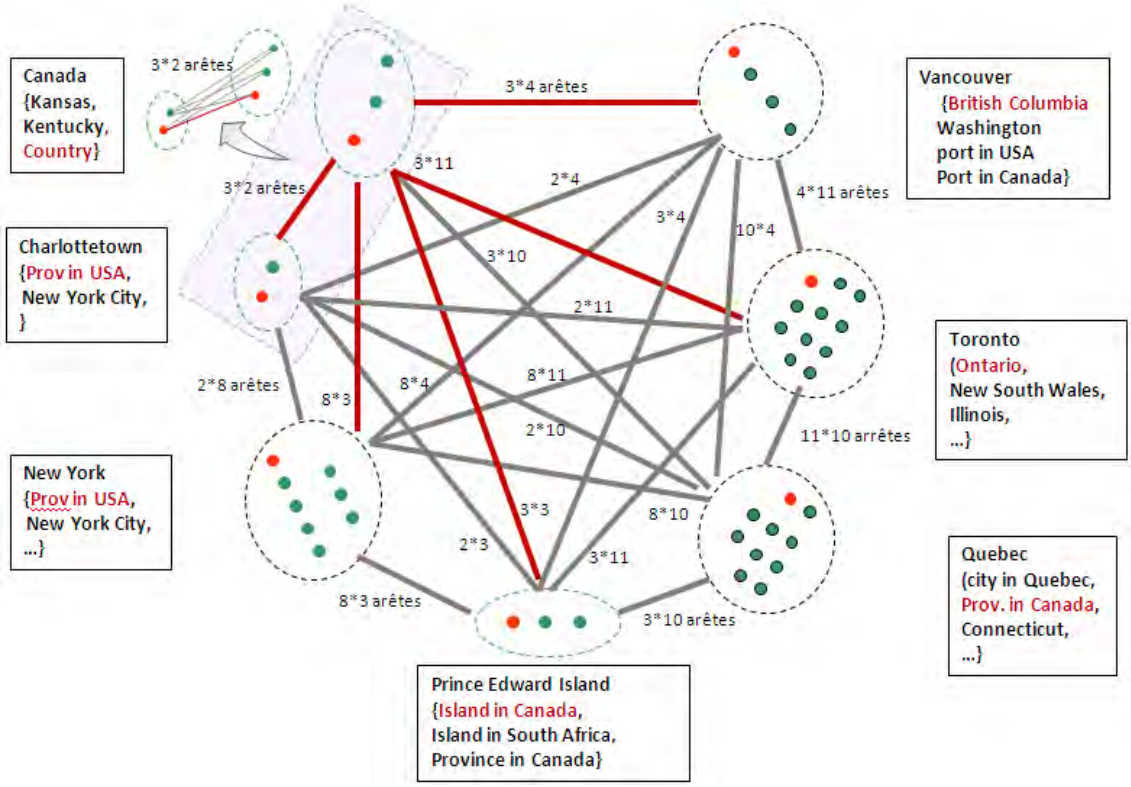

Figure 3-4. Le graphe des lieux et l'arbre couvrant maximum d'après (Li, Srihari, et al. 2003)

[9] Un arbre couvrant d'un graphe est un sous-graphe sans cycle qui connecte tous les sommets ensemble. Un graphe peut comporter plusieurs arbres couvrants différents.

[10] Un arbre couvrant de poids maximal est un arbre couvrant dont le poids est plus grand ou égal à celui de tous les autres arbres couvrants du graphe.

Cette heuristique est utilisée dans (Li, Srihari, et al. 2002) en appliquant l'algorithme *Kruskal* pour calculer l'arbre couvant maximum, puis optimisée par le même groupe d'auteur dans (Li, Srihari, et al. 2003) au moyen de l'algorithme *Prim* qui utilise un espace de recherche plus petit par rapport à *Kruskal.*

H9 La densité conceptuelle

La *Densité Conceptuelle* (DC) est une mesure de corrélation entre le sens d'un mot et son contexte. Elle a été présentée dans le domaine de DSM par Agirre et Rigau (1996) puis reformulée par Rosso et al. (2003). Cette dernière est ensuite adaptée à la désambiguïsation des toponymes par Buscaldi et Rosso (2008a).

La formule de calcul de DC est :

$$DC(m, f, n) = m^{\propto} \left(\frac{m}{n}\right)^{\log f} \quad (1)$$

Tel que m est le nombre de nœuds (synsets)[11] pertinentes dans la sous-hiérarchie composée des lieux du contexte, n est le nombre total de synsets dans la sous-hiérarchie, et f est le poids de la fréquence du sens (par exemple 1 pour le sens le plus fréquent, 2 pour le second, etc.).

Dans la méthode de Buscaldi et Rosso (2008a), la densité conceptuelle est calculée pour chaque référent candidat du toponyme ambigu. En suite, le référent qui maximise cette valeur (c.-à-d. la densité conceptuelle) est celui qui sera attribué au toponyme ambigu.

L'explication détaillée de cette heuristique est hors l'objet du présent chapitre, mais il est suffisant de dire que la densité conceptuelle est une quantification d'une certaine proximité entre les toponymes du contexte. C'est-à-dire que cette heuristique résout les toponymes ambigus par les référents les plus proches les uns aux autres. Cependant, la proximité quantifiée n'est pas spatiale comme le cas

[11] Les mots synonymes dans WordNet sont regroupés dans des nœuds appelée synset.

des heuristiques H1, H5, H6 mais elle est plutôt une proximité dans l'arbre hiérarchique des lieux du monde comme le cas de H2, H3.1., H7.

L'heuristique que nous proposons (Bensalem et Kholladi 2009a) se situe dans la même classe des heuristiques expliquée ci-dessus, c.-à-d. elle est basée sur le contexte, mais nous laissons son explication pour le chapitre suivant.

3.4.2.2 Désambiguïsation par les règles de préférences

Le choix d'un référent parmi les candidats dans cette classe d'heuristiques dépend principalement des préférences et des intuitions de l'Homme et il est complètement indépendant du contexte (le contraire des heuristiques de la première classe (Section 3.4.2.1)).

Chaque règle de préférence permet directement de choisir un référent parmi les candidats, ou d'affecter un poids à chacun d'eux, et celui qui a le plus grand score (la somme des poids attribués par plusieurs heuristiques) sera ensuite choisi comme le référent correct. Par exemple, les auteurs de (Li, et al. 2006), ont utilisé une approche qui attribue des scores de probabilité aux candidats en se basant sur plusieurs heuristiques comme H10 et H13.

Une règle de préférence peut être basée sur l'intuition humaine (H10, …, H14, H16) ou sur des statistiques effectué sur des corpus de référence (H15), ou sur des exigences de l'application (H17).

Certaines heuristiques de cette classe ne sont qu'une simplification du problème, c.-à-d. elles ne conduisent pas directement au référent voulu mais plutôt elles réduisent le nombre de référents candidats, c'est le cas de H16 et H17.

Nous expliquons dans ce qui suit les heuristiques de la catégorie règles de préférences.

H10 La plus grande population

Cette heuristique consiste à attribuer au toponyme ambigu le référent avec la plus grande population, en s'appuyant sur une source d'informations fiables.

Cette heuristique est utilisée dans (Rauch, Bukatin and Baker 2003), (Amitay, et al. 2004) et (Pouliquen, et al. 2004), (Li, et al. 2006).

H11 Le référent de niveau supérieur

Soit une taxonomie de toponymes dont la racine est le monde et les feuilles sont les villes[12].

Si un toponyme peut se référer à deux référents candidats, dont l'un est un pays, et l'autre est une ville, H11 choisit celui qui appartient à la classe la plus supérieure, dans ce cas c'est le pays qui sera choisi.

Cette heuristique est utilisée dans (Smith and Crane 2001), (Li, Srihari, et al. 2003), (Clough 2005) et (Stokes, et al. 2008).

H12 Le référent le plus connu

Le choix du référent correct est basé sur l'intuition humaine loin de toute connaissance fournie par les gazetteers ou d'autres ressources. Li, Srihari, et al. (2002, 2003) ont développé une procédure qui récupère le lieu le plus connu pour un toponyme ambigu en se basant sur les mécanismes de « ranking » des moteurs de recherche. Leur heuristique utilise le moteur de recherche *Yahoo!*[13].

Exemple

Si le toponyme « Cairo » est mentionné dans un texte, cette heuristique lui attribue le référèrent « Cairo>Egypte » au lieu de « Cairo>Alabama>USA » par exemple. Car les premiers résultats retournés par la requête « cairo » au moteur de recherche Yahoo! représentent le référent « Cairo>Egypte », comme c'est illustré dans la Figure 3-5.

[12] C'est ce que nous avons appelé l'arbre hiérarchique des lieux du monde.
[13] http://www.yahoo.com

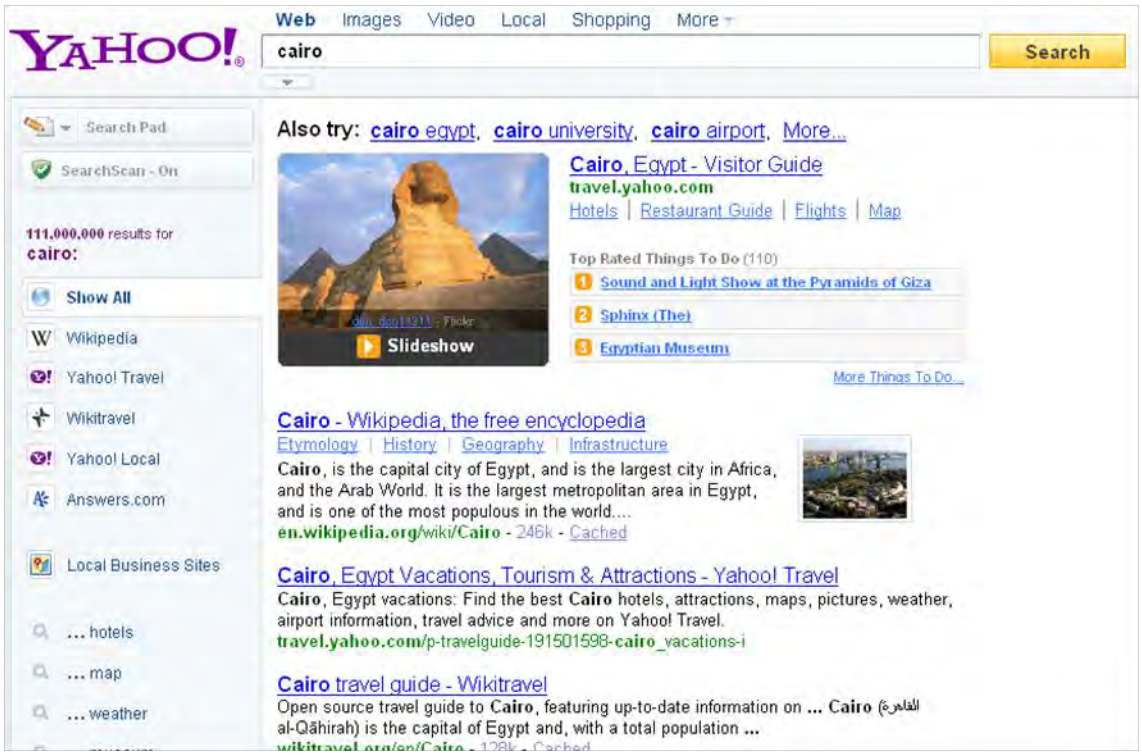

Figure 3-5. Les résultats de la requête "cairo" dans le moteur de recherche Yahoo!

H13 Préférer un type

Par exemple préférer les référents qui représentent des capitales, ou préférer les lieux habités que les divisions administratives…etc.

Exemple

Constantine peut indiquer la wilaya de Constantine ou la ville de Constantine, si le type préféré est « ville » alors c'est la ville de Constantine qui est choisie comme référent.

Cette heuristique est utilisée dans : (Li, Srihari, et al. 2003), (Li, et al. 2006).

H14 Ordre de préférence des ressources

Lors de l'utilisation parallèle de plusieurs gazetteers, il peut être utile de définir un ordre de priorité statique entre eux. Clough (2005) a prouvé l'efficacité de cette méthode en établissant un ordre de préférence entre 3 ressources géographiques selon leurs qualités.

H15 Le sens le plus fréquent dans un corpus

Il s'agit de choisir le référent qui est situé dans l'état ou le pays le plus fréquent. Ces fréquences d'occurrence sont calculées sur un corpus d'apprentissage.

Smith et Mann (2003) ont utilisé les résultats de cette heuristique comme référence pour mesurer les performances de leur méthode principale.

Stokes et al. (2008) ont supposé que l'emplacement le plus fréquent pour un toponyme est celui représenté par la page de Wikipedia[14] qui contient le plus grand nombre d'occurrences de ce toponyme. Le classement des pages de Wikipedia selon le nombre d'occurrence d'un toponyme est obtenu par le service web GeoNames[15]. L'intuition derrière cette heuristique[16] est que les contributeurs de Wikipédia ont tendance à écrire un article plus long (conséquemment avec plus de mentions du toponyme) pour l'emplacement le plus souvent associé à un toponyme ambigu.

Exemple

On ne s'attend pas d'avoir un long article sur Gaza située aux États-Unis que celui sur Gaza de Palestine, donc l'article de Gaza>États-Unis ne contient pas autant d'occurrence du terme Gaza, par conséquence il ne sera pas classé le premier dans les résultats de recherche fournies par GeoNames. Et donc c'est Gaza>Palestine qui sera attribué au toponyme Gaza.

H16 Supprimer les petites places

Il s'agit de réduire la taille de la ressource des lieux géographiques en fonction de la taille de la population. Cela diminue l'ambiguïté, mais bien évidemment c'est une simplification du problème plutôt que une véritable solution. Toutefois, Pouliquen et al. (2004) ont démontré que cette technique peut être utile dans certaines applications.

[14] http://www.wikipedia.org
[15] http://www.geonames.org
[16] Cette clarification avec l'exemple utilisé est obtenue par une communication personnelle avec Nicola Stokes (le premier auteur de l'article).

H17 Concentration sur une zone géographique

Cette heuristique consiste à ignorer les référents qui se trouvent en dehors d'un polygone ou d'une zone géographique (pays, région, continent…).

La zone géographique concernée est sélectionnée d'une manière statique, c. à. d. elle ne dépend pas formellement du texte mais plutôt, c'est une décision faite par l'utilisateur ou le concepteur du système de désambiguïsation. Cette heuristique peut être considérée comme la version statique de H6, et elle est utilisée dans (Pouliquen, et al. 2004).

3.4.2.3 Heuristiques complémentaires

Les heuristiques de cette classe ne conduisent pas toutes seules à la désambiguïsation des toponymes mais plutôt elles sont utilisées comme des procédures complémentaires dans les méthodes de désambiguïsation.

H18 Un référent par discours

Il s'agit de supposer que tous les toponymes identiques partagent le même référent, c'est-à-dire propager le sens des toponymes résolus à ceux qui ont la même forme dans le document. Cette heuristique a des origines dans le domaine de la désambiguïsation des sens des mots (Gale, Church et Yarowsky 1992).

Cette heuristique est utilisée par (Leidner, Sinclair et Webber 2003), (Li, Srihari, et al. 2003), (Amitay, et al. 2004), (Schilder, Versley et Habel 2004), (Pouliquen, et al. 2004), (Hauptmann and Olligschlaeger 1999).

H19 Attribuer les référents aux toponymes non ambigus

Un toponyme est dit non ambigu s'il a exactement un seul référent comme candidat. L'affectation des référents aux toponymes non ambigus est une étape triviale utilisée par tous les systèmes de désambiguïsation des toponymes, et habituellement invoquée en premier lieu avant le traitement des toponymes ambigus.

La Figure 3-6 résume les différentes heuristiques de l'état de l'art de désambiguïsation des toponymes. Et le Tableau 3-3 distribue ces heuristiques selon leurs références.

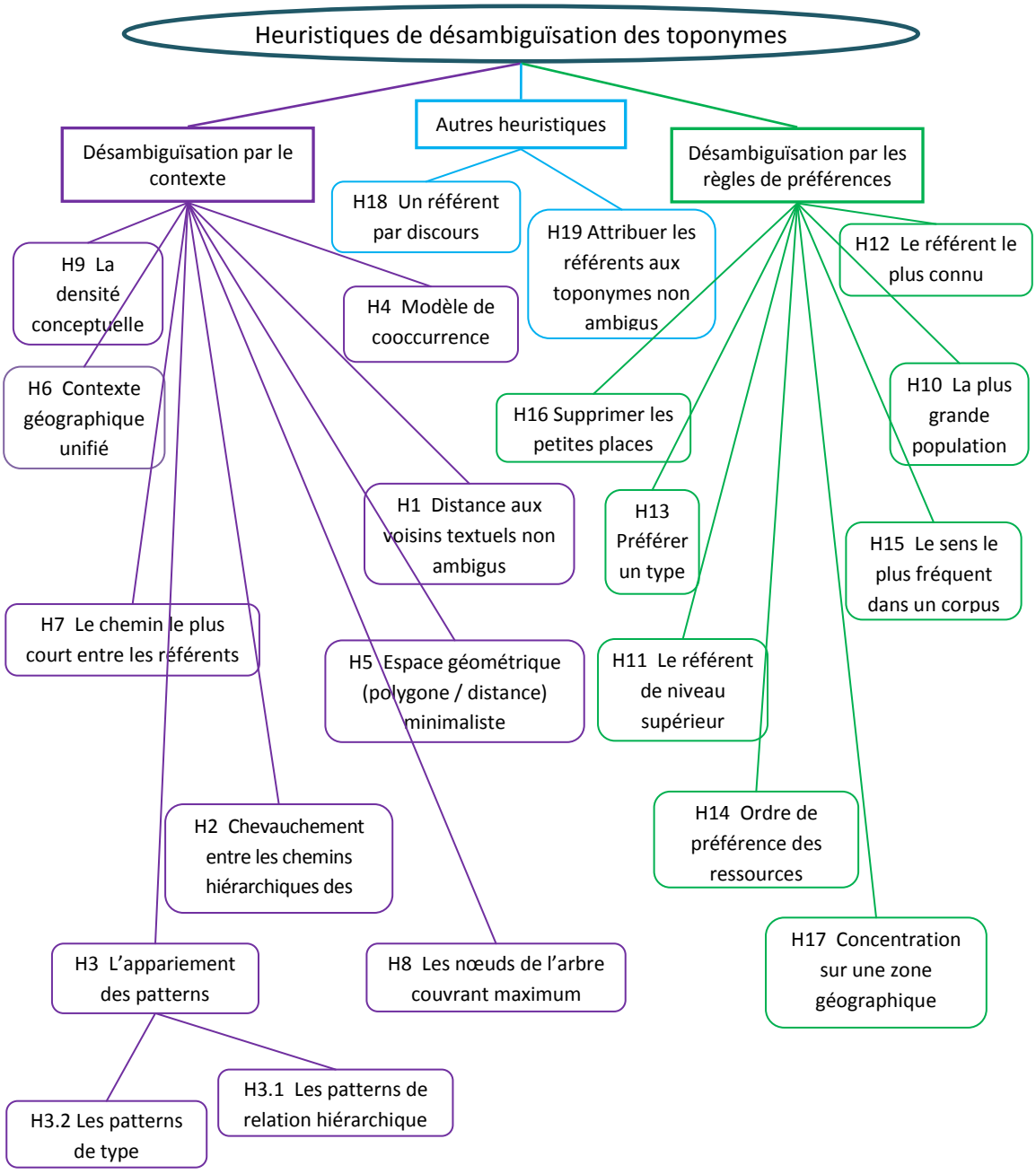

Figure 3-6. Classification des heuristiques de désambiguïsation des toponymes

Tableau 3-3. Distribution des heuristiques de désambiguïsation des toponymes utilisées dans la littérature[17]

	Désambiguïsation par le contexte										Désambiguïsation par les règles de préférences								
	H1	H2	H3.1	H3.2	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	H17	H18
(Hauptmann and Olligschlaeger 1999)		■	■																■
(Smith and Crane 2001)	■		■				■					■							
(Leidner, Sinclair et Webber 2003)						■													■
(Li, Srihari, et al. 2003)			■	■					■			■	■	■					
(Rauch, Bukatin and Baker 2003)			■	■		■					■								
(Smith et Mann 2003)					■											■			
(Amitay, et al. 2004)			■			■					■								■
(Pouliquen, et al. 2004)		■									■						■	■	■
(Schilder, Versley et Habel 2004)				■															■
(Clough 2005)		■		■								■			■				
(Li, et al. 2006)		■									■			■					
(Overell et Rüger 2007)					■														
(Buscaldi et Rosso 2008a)										■									
(Stokes, et al. 2008)				■				■			■	■				■			

3.5 Les connaissances

Les connaissances représentent l'ensemble des informations à propos du toponyme ambigu et ces référents candidats. Les connaissances sont les éléments de base derrière le choix du référent correcte. Sans connaissances il n'est pas possible ni pour l'homme ni pour la machine de déterminer le sens des mots ambigus (Navigli 2009) y compris les toponymes.

Nous présentons dans cette section une synthèse des connaissances manipulées dans l'état de l'art des méthodes de DT.

[17] La représentation de ce tableau est inspirée de (Leidner 2007, p.116), mais la signification des heuristiques est différente de celle de (Leidner 2007) comme j'ai déjà expliqué dans la section 3.2.

3.5.1 Classification des connaissances

D'après notre point de vue, les connaissances peuvent être classifiées selon 5 critères : la cible, la source, le domaine, la nature, et la méthode d'acquisition. Le Tableau 3-4 fourni une explication de ces critère avec les classes engendrées.

Tableau 3-4. Critères de classification des connaissances utilisées pour la désambiguïsation des toponymes

Critère	Classes	Explication
Cible	▪ Connaissances à propos du toponyme ▪ Connaissances à propos des référents	On veut dire par « cible » celui qui est concerné par les connaissances. Les noms des classes répondent à la question : cette connaissance est à propos de quoi ?
Sources	▪ Contexte ▪ Gazetteer (ou autre ressource qui joue le même rôle) ▪ Corpus	Classification selon la ressource à partir de laquelle une connaissance est obtenue. (Voir Section 3.6 pour plus de détails sur les sources de connaissances.)
Domaine	▪ Connaissances linguistiques ▪ Connaissances géographiques	Les connaissances linguistiques sont extraites du texte, tandis que les connaissances géographiques sont concrètes et concerne le monde réel.
Nature	▪ Valeur ▪ Relation	Exemple : pour un toponyme ambigu, le nombre de son apparition dans le texte (la fréquence d'occurrence) est une valeur. Cependant, les toponymes qui apparaissent avec lui dans le même contexte (les cooccurrences) sont des relations.
Méthode d'acquisition	▪ Direct ▪ Indirect	Il y a des connaissances obtenues directement depuis les ressources c.-à-d. elles sont brutes et d'autres sont calculées à partir des connaissances brutes (ex. la distance est calculée à partir des coordonnées spatiales), ou extraites du texte comme le type d'un toponyme (Voir H3.2 Les patterns de type)

Le critère de classification des connaissances le plus discriminant est la cible. Selon ce critère nous divisons les connaissances manipulées par les heuristiques de DT

en 2 classes : connaissances à propos du toponyme à résoudre et connaissances à propos des référents. Dans le reste de cette section nous présentons une vue d'ensemble sur ces deux classes de connaissances. Le schéma de la Figure 3-7 illustre cette description.

Dans cette sous-section le terme *toponyme* est utilisé pour désigner un mot qui représente un nom géographique mais qui n'a pas encore une représentation concrète dans le monde, tandis que le terme *référent* désigne un toponyme dont son sens est connu.

3.5.1.1 Connaissances à propos des toponymes

Les toponymes sans 'grounding' ne sont que des mots c.à.d. des unités lexicales, tant qu'ils sont ambigus, ils ne possèdent aucune relation avec le monde physique. Cela explique le fait que la quasi-totalité des connaissances pouvant être obtenues à propos des toponymes sont linguistiques. La seule connaissance géographique qui peut être obtenu à propos d'un toponyme est parfois le type de lieu à lequel il se réfère. En fait, le texte peut contenir une phrase qui indique que le toponyme mentionné est une ville ou une capitale…etc. Ce type de connaissance est obtenu par les patterns de type (Voir l'heuristique H3.2).

À partir du contexte on peut calculer des *valeurs* comme la fréquence d'occurrence d'un toponyme et la distance textuelle entre les toponymes comme dans la méthode de (Li, Srihari, et al. 2003), ou extraire les *cooccurrences*. Les cooccurrences dans le domaine de la DT sont les toponymes qui apparaissent avec le toponyme à résoudre dans le même contexte. La récupération des cooccurrences depuis le contexte est une procédure incontournable dans toutes les méthodes de DT basées sur le contexte (Section 3.4.2.1).

3.5.1.2 Connaissances à propos des référents

Un référent représente une entité physique dans le monde (un lieu), par conséquent beaucoup de connaissances géographiques peuvent lui être associées.

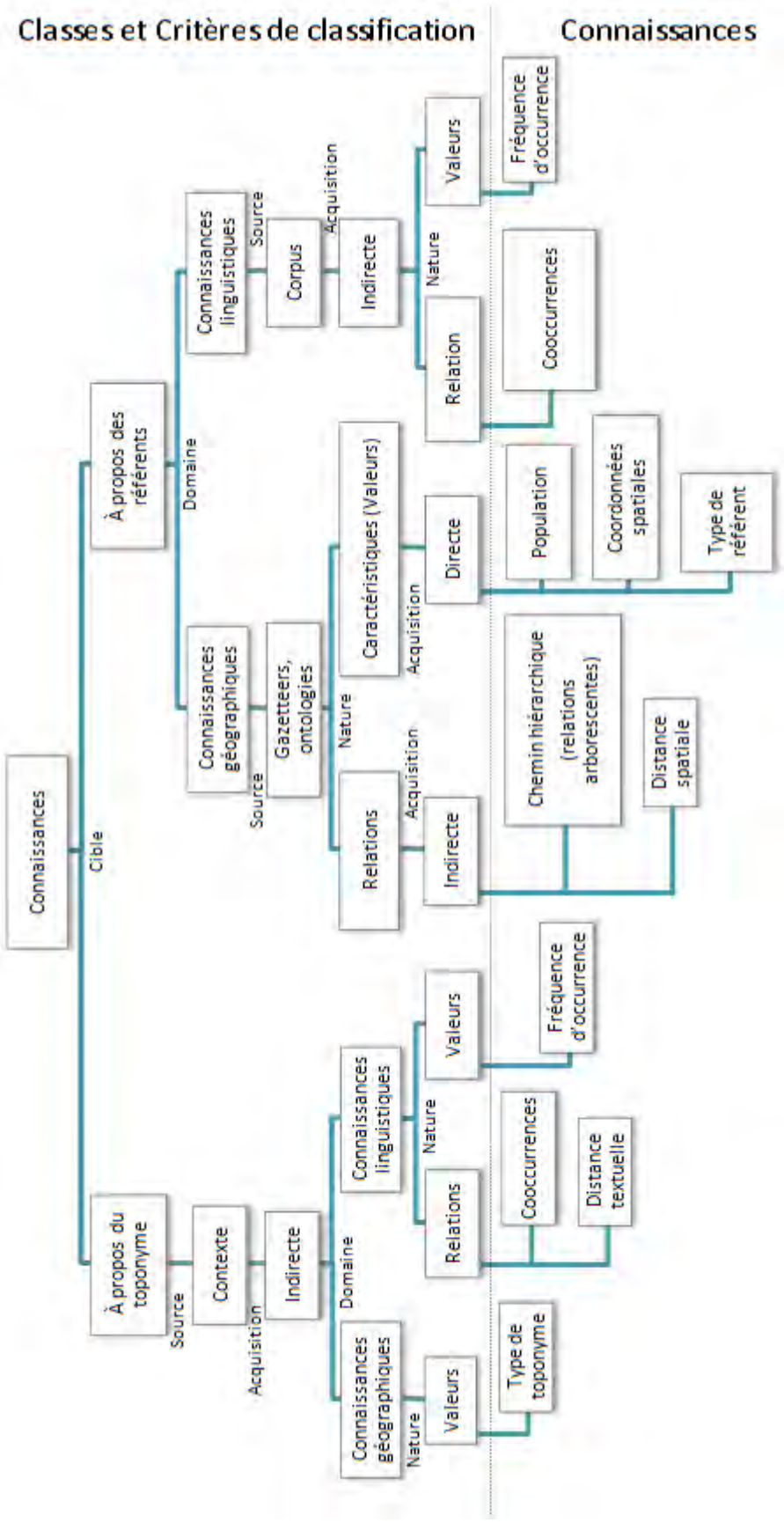

Figure 3-7. Taxonomie des connaissances utilisées pour la désambiguïsation des toponymes

Par exemple une fois le toponyme 'Constantine' est résous à 'Constantine> Algérie' on peut obtenir ses *caractéristiques* comme sa population, ses coordonnées spatiale,…, etc. et encore ses *relations* avec d'autres référents. Un exemple de relations est les distances spatiales entre les référents des toponymes du même contexte. Cette relation est exploitée dans (Smith and Crane 2001), (Leidner, Sinclair et Webber 2003), (Rauch, Bukatin and Baker 2003), et (Amitay, et al. 2004).

Un autre type de relations qui peut être exploité comme connaissance dans les méthodes de DT sont les relations arborescentes c.-à-d. les relations qui existent entre les référents dans l'arbre hiérarchique des lieux du monde (Bensalem et Kholladi 2009b). La relation arborescente la plus connue est '*est-partie-de*' (appelée aussi méronymie) à plusieurs niveaux, qui est généralement représentée sous forme d'un chemin hiérarchique. Par exemples le chemin hiérarchique : 'Jérusalem>Palestine>Asie' indique que Jérusalem *est-partie-de* Palestine, et Palestine *est-partie-de* l'Asie, et par conséquence Jérusalem *est-partie-de* l'Asie.

Contrairement aux relations basées sur la distance spatiale, les relations arborescentes ne sont pas exploitées explicitement dans les méthodes de DT (voir le chapitre 4 pour une ample discutions sur cette lacune).

Les connaissances linguistiques qui peuvent être obtenues à propos des référents sont généralement des valeurs statistiques, entre autre, la fréquence d'occurrence dans un corpus, et les cooccurrences c.à.d. les toponymes qui apparaissent fréquemment avec le référent cible (Voir l'heuristique H4 pour plus d'informations sur l'utilité des cooccurrences).

3.6 Les ressources

Toute source de connaissance hormis le contexte est appelée *ressource*. Les ressources ont deux rôles principaux dans la DT qui sont :

1. Fournir les différents référents d'un toponyme. Ce qui représente la première étape de la DT ;
2. Fournir des connaissances linguistiques et géographiques sur les référents (Voir aussi la taxonomie de connaissances dans la Figure 3-7).

Les ressources peuvent offrir des connaissances générales ou spécifiques à un domaine. Par exemple, WordNet[18] (voir Chapitre 4, p98) fourni des définitions et des relations pour plusieurs types de mots: les noms (y compris les toponymes), les verbes, les adjectifs, et les adverbes. Tandis que les gazetteer (Section 3.6.1) sont des ressources de connaissances sur les lieux géographiques seulement.

Les ressources utilisées dans l'état de l'art sont : les ontologies (Volz, Kleb et Mueller 2007), les corpus linguistiques (Smith et Mann 2003), les gazetteer.

3.6.1 Les gazetteers

Gazetteer est un terme anglais[19] qui représente traditionnellement un dictionnaire de toponymes. Maintenant, les gazetteers sont considérés comme un type de Systèmes d'Organisation des Connaissances (SOC), qui organisent des informations sur les lieux géographiques nommés (Hill 2006).

Une entrée dans un gazetteer contient au minimum 3 types d'informations (Leidner 2007) qui sont un toponyme avec son type et son empreinte spatiale:

Toponyme : nom d'un objet[20] géographique et éventuellement ses variantes historique ou vernaculaire (voir Section 2.2).

Type : c'est la catégorie de l'objet géographique à lequel se réfère le toponyme, par exemple : région administrative, pays, cité, montagne, pont, …, etc.

[18] http://wordnet.princeton.edu
[19] Nous avons choisi d'utilisé le terme gazetteer dans ce mémoire car il n'a pas une traduction unique et précise en français.
[20] On dit objet car le gazetteer peut contenir non seulement des noms de lieux comme les pays et les villes mais aussi des noms des montagnes, des rivières, des constructions …, etc.

Empreinte spatiale : représentation de la location référée par le toponyme dans un système de coordination par exemple la latitude et la longitude.

Les gazetteers diffèrent entre eux dans les types d'objets qu'ils renferment (ex. lieux habités, étendus d'eau, montagnes…) la couverture géographique (ex. le monde, un continent, un pays…), la granularité des lieux (ex. il peut contenir seulement les pays avec leurs villes comme il peut aller jusqu'aux villages, cartiers, rues..), et les détails de chaque entrée (population, longitude et latitude, code postale, superficie…) (Hill 2006) (Leidner 2007, Chapitre 4).

Les gazetteers sont utilisés dans les méthodes de DT pour 4 objectifs :

1. Identifier les toponymes dans le texte ;
2. Fournir la liste des référents candidats pour chaque toponyme ;
3. Fournir des connaissances géographiques à propos des référents ;
4. Annoter les corpus destinés à l'évaluation des méthodes de DT, ou ceux servant comme source de connaissance, notamment, dans les méthodes supervisées (comme (Smith et Mann 2003)). Voir Section 3.6.2 pour plus d'informations sur les corpus et leur annotation.

Le Tableau 3-5 montre les connaissances fournies par les gazetteers et les heuristiques qui les manipulent. Nous remarquons que les connaissances des gazetteers sont manipulées presque par tout les heuristique de désambiguïsation des toponymes.

Tableau 3-5. Les connaissances fournies par les gazetteers et les Heuristiques qui les manipulent

connaissances	Heuristiques qui les manipulent
Position géo-spatiale	H1 Distance aux voisins textuels non ambigusH1 H5 Espace géométrique (polygone / distance) minimaliste H6 Contexte géographique unifiéH6
Chemin hiérarchique	H2 Chevauchement entre les chemins hiérarchiques des référents et le **texte**

	H3.1 Les patterns de relation hiérarchique H7 Le chemin le plus court entre les référents H8 Les nœuds de l'arbre couvrant maximum
Population	H10 La plus grande population H16 Supprimer les petites places
Type de référent	H11 Le référent de niveau supérieur H3.2 Les patterns de type H13 Préférer un type

Le Tableau 3-6 fournit des informations sur quelque gazetteers utilisés dans la littérature de la DT.

Tableau 3-6. Exemple de gazetteers utilisés dans les méthodes de désambigüisation des toponymes

Nom	Nombre d'entrées	Site web	Utilisé par
The Getty Thesaurus of Geographic Names (TGN)	1.115.000	http://www.getty.edu/research/conducting_research/vocabularies/tgn	(Stokes, et al. 2008) (Li, et al. 2006) (Overell et Rüger 2007) (Clough 2005)
World gazetteer	inconnu	http://world-gazetteer.com	(Amitay, et al. 2004) (Stokes, et al. 2008) (Li, et al. 2006)
USGS Geographic Names Information System (GNIS)	1.836.264	http://geonames.usgs.gov	(Amitay, et al. 2004) (Volz, Kleb et Mueller 2007) (Garbin et Mani 2005)

3.6.2 Les corpus

Un corpus est une collection de textes utilisées pour apprendre des modèles de langue (Navigli 2009).

Les corpus dans le domaine de la DT sont des ressources textuelles[21] où tous les toponymes sont annotés avec des informations spatiales qui indiquent une position unique dans la Terre (Leidner 2007).

Les corpus sont utilisés dans les heuristiques de DT pour obtenir deux connaissances linguistiques: les collocations (c.-à-d. les cooccurrences fréquentes) (voir H4), et des statistiques linguistiques à propos de la distribution de l'occurrence des toponymes et leurs sens (ex. trouver le référent le plus fréquents pour un toponyme (voir H15)).

En plus de leur utilisation comme source de connaissances, les corpus sont utilisés aussi comme *terrain vérité* pour l'évaluation des méthodes de DT.

Dans un corpus de DT chaque toponyme doit être annoté par un label (tag) qui détermine le lieu à lequel il se réfère (ex. la latitude et la longitude). Les informations de l'annotation sont obtenues depuis les gazetteers. Conséquemment, l'utilisation d'un certain corpus pour l'évaluation impose l'utilisation du gazetteer avec lequel il est annoté.

Malheureusement, l'évaluation est encore problématique dans la communauté de recherche à cause du manque de corpus standards dédiés à la tâche de désambiguïsation des toponymes (Leidner 2007). Les méthodes de la littérature sont toutes évaluées sur des corpus différents.

3.6.3 Les ontologies

Volz, et al. (2007) ont présenté une approche de DT basée sur une ontologie et sa lexicalisation[22]. Dans leur approche, l'ontologie sert à identifier les toponymes dans le texte, à leur associer les référents possibles, et à fournir des connaissances pour la désambiguïsation.

[21] Un ensemble de documents qui contiennent du texte libre en langue naturelle.
[22] Création automatisée des listes qui comprennent tous les mots utilisés pour nommer respectivement les concepts, les relations, et les instances d'une ontologie.

L'ontologie dans l'approche de (Volz, Kleb et Mueller 2007) n'a joué pratiquement que le rôle d'un gazetteer mais seulement, elle a une structure différente où chaque type géographique est représenté par un concept (une classe), les référents sont les instances, et les toponymes sont le vocabulaire des instances. D'après notre point de vue, le vrai avantage des ontologies, est l'inférence des relations, mais malheureusement, cela n'a pas été exploité dans cette approche.

La méthode de Buscladi et Rosso (Buscaldi et Rosso 2008a) est basée sur l'ontologie WordNet qui a été utilisée pour fournir les différents sens d'un toponyme mais aussi pour calculer la densité conceptuelle (voir l'heuristique H9)

3.7 Conclusion

Nous avons articulé l'état de l'art sur quatre (4) axes, qui sont le contexte, les heuristiques, les connaissances et les ressources. Ces quatre composants sont les piliers de toute méthode de désambiguïsation des toponymes. Après cet état de l'art, nous avons remarqué que l'idée de désambiguïser les toponymes, par les référents les plus proches dans l'arbre hiérarchiques du monde, n'a pas été proposée auparavant. En effet, la seule relation arborescente entre les toponymes du même contexte qui a été exploitée explicitement pour la désambiguïsation est la méronymie (*est-partie-de*). Dans le chapitre suivant, nous proposons une nouvelle heuristique de DT qui désambigüise les toponymes ambigus du même contexte par les référents les plus proches les uns aux autres en termes de toutes les relations arborescentes qui peuvent exister entre eux.

Chapitre 4

Une nouvelle Heuristique de Désambiguïsation des Toponymes

Une partie de ce chapitre se trouve dans les articles (Bensalem et Kholladi 2009b)*et* (Bensalem et Kholladi 2009c)

4.1 Introduction

Nous présentons dans ce chapitre notre contribution principale dans ce mémoire qui est une nouvelle heuristique de désambiguïsation des toponymes basée sur le calcul de la plus forte relation arborescente entre les référents des toponymes du même contexte. Notre heuristique exploite la connaissance « relation arborescente » qui n'est pas exploitée d'une manière explicite dans les méthodes de l'état de l'art.

Nous commençons d'abord par présenter notre motivation, puis nous présentons notre méthode, en introduisant la mesure de la *Densité Géographique* que son calcul se base principalement sur les chemins hiérarchiques. Nous fournissons dans la section 04.3 les résultats d'évaluation de notre heuristique en la comparant avec une autre. Enfin, nous terminons par une conclusion qui résume les différents points discutés dans ce chapitre.

4.1.1 Aperçu sur les travaux antérieurs

Nous avons proposé dans le chapitre précédent (Section 3.4.2) une classification des heuristiques existantes de la désambiguïsation des toponymes. Cette classification a engendré deux catégories principales[1]: les *heuristiques de désambiguïsation par le contexte,* et *les heuristiques de désambiguïsation par les règles de préférence*.

Nous rappelons que les heuristiques de la première catégorie dépendent principalement des toponymes qui existent dans le même contexte dans lequel le toponyme à désambiguïser apparaît. Cela rend la tâche de désambiguïsation des toponymes similaire à la désambiguïsation des sens des mots (DSM) (Navigli 2009) qui est parmi les tâches connues du traitement automatique des langues naturelles (TALN). On veut dire par contexte (Section 3.3), le texte en langue naturelle qui contient le(s) toponyme(s) à désambiguïser. La taille de ce dernier

[1] Une troisième catégorie décrite dans le chapitre précédant contient des heuristiques complémentaires.

dans les méthodes de DT varie de quelques toponymes autour du toponyme ambigu jusqu'à tous les toponymes du texte du document.

Toutefois, les heuristiques de la deuxième catégorie désambiguïsent les toponymes en se basant sur des préférences et des intuitions de l'être humain. Par exemple, désambiguïser par les référents à plus grande population (Pouliquen, et al. 2004) (Amitay, et al. 2004) (Rauch, Bukatin and Baker 2003) ou par les référents les plus fréquents (Stokes, et al. 2008).

A titre d'exemple, si le toponyme à résoudre est 'Alexandrie', les deux heuristiques de la deuxième catégorie lui associent le référent 'Alexandrie>Égypte' au lieu de 'Alexandrie>Piémont>Italie' par exemple, car le premier lieu est le plus connu et le plus peuplé[2]. Tandis que le référent choisi par les heuristiques de la première catégorie peut être 'Alexandrie>Égypte' ou 'Alexandrie>Piémont>Italie' selon les toponymes qui apparaissent avec 'Alexandrie' dans le même contexte.

Dans le but de facilité la lecture de ce chapitre, le Tableau 4-1 (voir p. 90) rappelle la liste des heuristiques présentée dans le chapitre 3.

4.1.2 Les types de relations entre les toponymes du même contexte

En observant les heuristiques de la désambiguïsation des toponymes par le contexte, nous remarquons que derrière la plus part des heuristiques de cette classe, se cache une intuition qui consiste à supposer l'existence d'une certaine proximité géographique entre les référents des toponymes du même contexte.

H1, H5, H6 désambiguïsent les toponymes par les référents les plus proches en termes de distance, ce qui implique à faire des calculs géométriques en utilisant les coordonnées spatiales des référents. Cependant, les heuristiques H2, H9, H3.1, H7 désambiguïsent les toponymes par les référents les plus proches dans l'arbre

[2] Référents et statistiques de population selon *Word Gazetteer* : http://world-gazetteer.com (dernière consultation le 28 septembre 2009).

hiérarchique des lieux du monde. Dans ce cas, les référents doivent être représentés par leurs chemins hiérarchiques.

Tableau 4-1. Rappel des heuristiques de l'état de l'art de désambiguïsation des toponymes

Heuristiques de désambiguïsation par le contexte	**H1**	Distance aux voisins textuels non ambigus
	H2	Chevauchement entre les chemins hiérarchiques des référents et le texte
	H3	L'appariement des patterns
	H3.1	Les patterns de relation hiérarchique
	H3.2	Les patterns de type
	H4	Modèle de cooccurrence
	H5	Espace géométrique (polygone / distance) minimaliste
	H6	Contexte géographique unifié
	H7	Le chemin le plus court entre les référents
	H8	Les nœuds de l'arbre couvrant maximum
	H9	La densité conceptuelle
Heuristiques de désambiguïsation par les règles de préférence	**H10**	La plus grande population
	H11	Le référent de niveau supérieur
	H12	Le référent le plus connu
	H13	Préférer un type
	H14	Ordre de préférence des ressources
	H15	Le sens le plus fréquent dans un corpus
	H16	Supprimer les petites places
	H17	Concentration sur une zone géographique
Heuristiques complémentaires	**H18**	Un référent par discours
	H19	Attribuer les référents aux toponymes non ambigus

Nous appelons « relation spatiale » entre les référents toute relation géographique résultante de la proximité des distances, et « relation arborescente » toute relation résultante de proximités dans l'arbre hiérarchique des lieux du monde (Bensalem et Kholladi 2009b).

En outre, nous distinguons deux types de relations arborescentes: les *relations hiérarchiques*, et les *relations non hiérarchiques* (Bensalem et Kholladi 2009b).

Les relations hiérarchiques existent entre les lieux de la même branche dans l'arbre. Par exemple entre un pays et une de ses villes ; comme entre l'Algérie et Constantine dans la Figure 4-1.

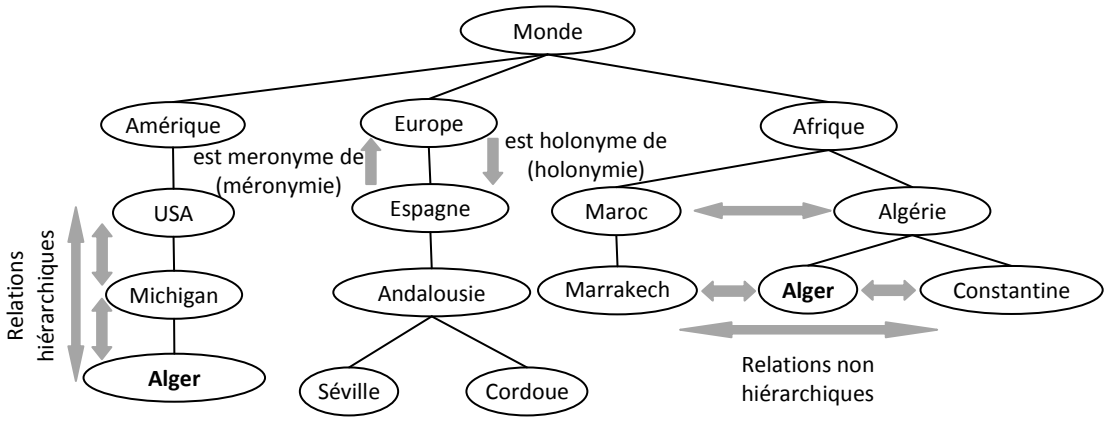

Figure 4-1. Une partie de l'arbre hiérarchique du monde (Alger est un toponyme ambigu)

Les relations non hiérarchiques sont celles qui existent entre les nœuds qui se trouvent dans des branches différentes mais qui ont une (ou plusieurs) racine commune. La racine commune peut être directe (ex. Andalousie par rapport à Séville et Cordoue) ou indirecte (ex. Afrique par rapport à Constantine et Marrakech).

Il existe deux sortes de relations hiérarchiques: la *méronymie* qui est la relation «*est-partie-de*» et *l'holonymie*[3] qui représente la relation « *contient-la-partie* ». Par exemple, nous disons que 'Algérie' est un *holonyme* de 'Constantine' et 'Constantine' est un *meronyme* de l'Algérie.

Un chemin hiérarchique d'un lieu est donc composé d'un ensemble de toponymes connectés les uns aux autres par des relations d'holonymie/méronymie. Par exemple 'Alger>Algerie>Afrique' et 'Alger>Michigan>USA>Amérique' sont des chemins hiérarchiques du toponyme ambigu 'Alger' (voir Figure 4-1).

La Figure 4-2 résume les différents types des relations géographiques qui peuvent exister entre les lieux du même contexte.

[3] L'holoymie et la méronymie sont des termes qui expriment des relations sémantiques et ils sont originaires de la discipline de la linguistique.

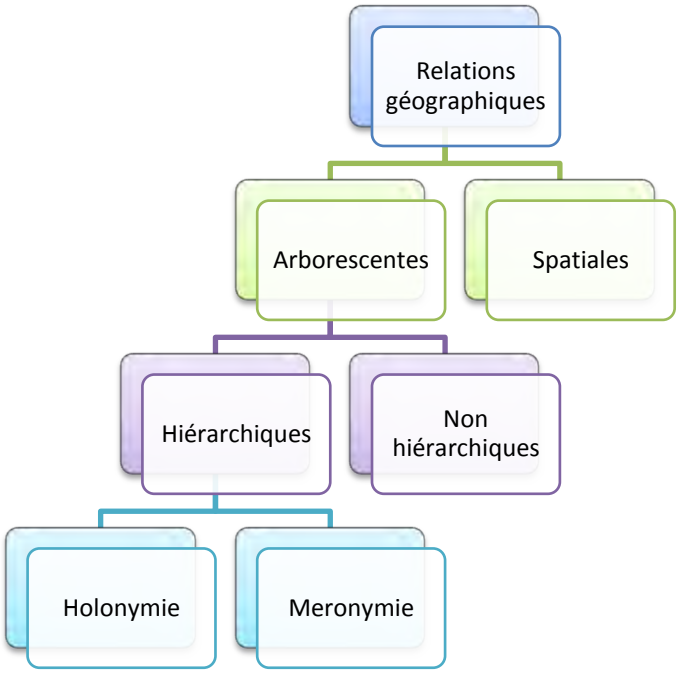

Figure 4-2. Les différents types de relations géographiques qui peuvent exister entre les lieux mentionnés dans le même contexte

4.1.3 Une nouvelle perspective au problème de la désambiguïsation des toponymes

Nous avons remarqué que la plupart des méthodes existantes fondées sur la proximité arborescente sont capables de résoudre les toponymes en recherchant ses holonymes dans le contexte c.-à-d. en cherchant ses relations de méronymie[4]. Clough (2005) quantifie l'existence des holonymes par le calcul du score du chevauchement (SC) entre le contexte et le chemin hiérarchique du référent (voir l'explication de H2 dans le chapitre 3). Le SC permet évidement d'identifier dans le contexte tous les holonymes d'un toponyme, qu'ils soient directes ou indirectes. L'heuristique de Pouliquen et al. (2004) cherche seulement la mention dans le contexte d'un holonyme du toponyme à résoudre sans le calcul du score du chevauchement.

[4] Étant donné que la relation de méronymie est la relation «est-partie-de », alors, chercher des relations de méronymie pour un toponyme consiste à trouver ses racine c.-à-d. ses holonymes.

Cependant, au meilleur de nos connaissances, les seules heuristiques qui essayent de chercher d'autres types de relations arborescentes entre les toponymes (c.-à-d. non seulement la relation de la méronymie) sont l'heuristique du plus court chemin (H7) de (Stokes, et al. 2008), et celle basée sur la densité conceptuelle (DC) de Buscaldi, et Rosso (2008a) (H9)[5].

Exemple

Si les toponymes : 'Alger', 'Constantine' sont mentionnés dans un texte, et Constantine est résous à 'Constantine >Algérie', il est possible en utilisant les heuristiques H7 et H9 de résoudre le toponyme ambigu 'Alger' par 'Alger>Algérie' car ce référent partage une racine commune avec 'Constantine' (donc une relation non hiérarchique avec 'Constantine') qui est l'Algérie (voir la Figure 4-1 p.91). Cependant, en utilisant les heuristiques basées sur le score du chevauchement avec le contexte on ne peut pas découvrir que le référent 'Alger>Algérie' est le plus relier à 'Constantine' par rapport aux autres référents de 'Alger'.

Stokes, et al. (2008) ont utilisé l'heuristique du plus court chemin (H7) comme une heuristique secondaire qui résout les toponymes ambigus par rapport aux ceux déjà résous par d'autres heuristiques. D'ailleurs, ils n'ont pas fourni des détails sur son principe.

L'heuristique de (Buscaldi et Rosso 2008a) quantifie des relations arborescentes entre les toponymes par le calcul de la densité conceptuelle, mais le principe de cette quantification n'est pas suffisament claire[6]. En outre la DC a été introduite pour la prmiere fois pour la désambiguisation des sens des mots (Agirre et Rigau 1996) (Rosso, et al. 2003) puis adaptée à la désambiguisation des toponymes (Buscaldi et Rosso 2008a), donc elle n'est pas conçue directement pour adresser le problème de DT.

[5] Sachant que cela reste notre point de vue sur leurs méthodes et ce n'est pas déclaré explicitement par les auteurs.

[6] Les auteurs n'ont pas déclaré que la DC est une mesure des relations arborescentes et n'ont pas expliqué sa formule dans ce sens.

En bref, contrairement aux relations spatiales, les relations arborescentes ne sont pas exploitées explicitement dans les méthodes de DT. En effet, à nos jours il n'y pas d'auteurs qui ont déclaré que leur méthode est basée sur la quantification des relations arborescentes de tous types entre les toponymes du même contexte. Ainsi, nous croyons que nous somme les premiers à voir le problème de désambiguïsation des toponymes dans cette perspective.

Nous proposons dans le reste de ce chapitre une nouvelle heuristique de désambiguïsation des toponymes basée sur le contexte. À la différence des autres heuristiques de cette catégorie, notre heuristique est conçue explicitement sur l'idée de chercher des relations arborescentes (hiérarchiques et non hiérarchiques) entre les toponymes du même contexte, et elle est basée sur une nouvelle mesure de corrélations arborescentes entre les toponymes que nous appelons la *Densité Géographique.*

4.2 Notre heuristique de désambiguïsation des toponymes

4.2.1 Notation

Tableau 4-2. Conventions de notation de l'heuristique de densité géographique

T l'ensemble des toponymes qui apparaissent dans un document D Chaque toponyme apparait une seule fois dans T. n est le nombre de toponymes.	$T = \{t_i \in D / i = 1...n\}$
G: un gazetteer. Chaque r_{id} est représenté par un ensemble de caractéristiques qui diffèrent selon le gazetteer utilisé. Dans cette heuristique nous avons besoin pour chaque lieu de : son identifiant, son nom et son chemin hiérarchique. On dit que le lieu r_{id} est un référent de t_i si t_i est le nom de r_{id}.	$G = \{ r_{id} / r_{id}$ est un lieu géographique dans la Terre $\}$

La suite du tableau est dans la page suivante

h_{id} est le chemin hiérarchique de r_{id} dans l'arbre d'hiérarchie de *G*. Chaque nœud de h_{id} est un référent $r_{id.k}$, tel que le premier nœud $r_{id.1}$ est l'extrême holonyme de r_{id} et le dernier nœud $r_{id.l}$ est r_{id}. tel que l est la longueur du chemin hiérarchique.	h_{id} =" r_{id1}>r_{id2}>…>r_{idl}"
Comp(h_{id}) sont les référents qui compose un chemin hiérarchique h_{id}.	*Comp*(h_{id}) ={ r_{idk}, $k=1..l$}
R_i : l'ensemble des référents du toponyme t_i.	R_i = { $r_{id} \in G$ / t_i est le nom de r_{id} }
H_i un ensemble composé des chemins hiérarchiques des référents de R_i	H_i = { h_{id} / $r_{id} \in R_i$ }
R est l'ensemble de tous les ensembles R_i, c.-à-d. l'ensemble des référents de tous les toponymes d'un document *D*.	*R* = { R_i , $i = 0..n$ }
H est l'ensemble de tous les ensembles H_i.	*H* = { H_i , $i = 0..n$ }
Comp(H_i): les composants de tous les $h_{id} \in H_i$ sans duplication des éléments	*Comp*(H_i)= $\cup$ *Comp*(h_{id})/ $h_{id} \in H_i$
Comp(H): l'ensemble des ensembles H_i.	*Comp(H)*= $\cup$ *Comp*(H_i)/ $H_i \in H$

4.2.2 Principe et méthode

Notre heuristique est basée sur l'hypothèse que les toponymes qui apparaissent ensembles dans le même document sont reliés géographiquement par des relations arborescentes qu'ils soient hiérarchiques ou non hiérarchiques.

L'heuristique proposée résout un toponyme par le référent qui est :

- Le plus relié géographiquement aux référents des autres toponymes, c.-à-d. celui qui possède relativement beaucoup de relations arborescentes avec les référents des autres toponymes (on peut dire que c'est une relation indirecte avec le contexte), et ;
- Le plus relié au contexte, c.-à-d. son chemin hiérarchique et le contexte contiennent relativement beaucoup de toponymes en commun (le même

principe de l'heuristique H2).

Ces deux caractéristiques sont quantifiées par le calcul de ce que nous appelons la *Densité Géographique* (Bensalem et Kholladi 2009a). Nous définissons donc la Densité Géographique (DG) comme une mesure de corrélation (directe ou indirecte) entre un référent d'un toponyme et le contexte de ce dernier.

La désambiguïsation des toponymes par le calcul de la densité géographique suit les étapes suivantes :

1. Extraire tous les toponymes du document *D* (taille du contexte = tous les toponymes du document).
2. Éliminer les duplications en appliquant l'hypothèse de « un sens par discours » (voir H18).
3. Déterminer la liste des référents candidats R_i pour chaque toponyme t_i. Chaque référent candidat r_{id} doit être représenté par son chemin hiérarchique h_{id}.
4. Calculer la densité géographique pour chaque référent candidat dans R_i de chaque toponyme t_i.
5. Attribuer à chaque toponyme t_i le référent r_{id} qui possède la plus grande densité géographique $DG(r_{id})$ parmi l'ensemble de ses référents candidats.

4.2.3 La densité géographique

Les connaissances principales sur lesquelles se base le calcul de la densité géographique sont les chemins hiérarchiques des référents candidats de tous les toponymes du contexte (c.-à-d. les éléments de l'ensemble *H*). Le chemin hiérarchique d'un référent est composé du référent lui-même, est ces holonymes[7] c.-à-d. sa racine directe, et ces racines indirectes.

La DG d'un référent r_{id} d'un toponyme ambigu t_i augmente lorsque :

[7] Dans l'intention de brièveté, désormais, le mot « holonnyme » seul suffira pour dire « holonymes directs et indirects » qui compose nt le chemin hiérarchique d'un toponyme.

(a) ce référent apparaît parmi les holonymes (les racines) des autres référents dans *R*–*Ri*, et /ou,

(b) ses holonymes sont parmi les référents candidats des autres toponymes (c.-à-d. dans *R*–*Ri*), et /ou,

(c) ses holonymes sont aussi des holonymes pour d'autres référents, et

(d) les toponymes qui composent son chemin hiérarchique existent partiellement ou totalement dans le contexte.

Les caractéristiques (a), (b) et (d) signifient la présence d'une relation hiérarchique entre le référent cible r_{id} et certains référents des autres toponymes, et (c) signifie la présence d'une relation non hiérarchique.

Les caractéristiques (a), (b) et (c) sont quantifiées par le calcul des fréquences du référent r_{id} et ses holonymes (c.-à-d. de $r_{id.1}$,..$r_{id.2}$, …$r_{id.l}$) dans les chemins hiérarchiques des référents de l'ensemble *R.* La fréquence d'un référent $r_{id.k}$ est la somme de ses poids dans chaque R_i (l'équation (2)).

Le poids P est une fonction booléenne qui indique l'existence ou l'absence d'un référent $r_{id.k}$ dans les chemins hiérarchiques d'un ensemble R_i (l'équation (3)). Par conséquent, La plus grande valeur que peut prendre une fréquence est égale à *n* : le nombre des ensembles R_i dans *R,* et ce qui représente aussi le nombre de toponymes dans le texte.

La caractéristique (d) est quantifiée par le calcul du score du chevauchement du chemin hiérarchique du référent r_{id} avec le contexte *D*, cela est représenté par la valeur SC(h_{id}, *D*).

La densité géographique *DG* (r_{id}, *R)* d'un référent candidat r_{id} est la somme de ces deux valeurs décrites ci-dessus (la fréquence des référents qui compose son chemin hiérarchique h_{id} et le score du chevauchement de ce dernier avec le contexte) (l'équation (1)).

$$\mathrm{DG}\,(r_{id}, R) = \sum_{k=1}^{l} (\text{Fréquence}\,(\,r_{id.k}\,, R)) \;+\; \mathrm{SC}(h_{id}\,, D) \tag{1}$$

$$\text{Fréquence}\,(\,r_{id.k}\,, R) = \sum_{i=1}^{n} \mathrm{P}(r_{id.k}, R_i) \tag{2}$$

$$\mathrm{P}(r_{id.k}, R_i) \;= \begin{cases} 0, \text{si le nombre de } r_{id.k} \text{ dans } \mathrm{Comp}(H_i) = 0 \\ 1, \text{si le nombre de } r_{id.k} \text{ dans } \mathrm{Comp}(H_i) \neq 0 \end{cases} \tag{3}$$

4.3 Évaluation

4.3.1 Description des ressources

L'évaluation des méthodes de la désambiguïsation des toponymes nécessite l'utilisation de deux ressources principales qui sont les corpus textuels et les inventaires de sens comme les gazetteers et les ontologies. L'évaluation est encore problématique dans ce domaine dû au manque de ressources standards qui permettent la comparaison entre les performances des différentes méthodes. Leidner (2004, 2006) a adressé ce problème mais malheureusement ses données ne sont pas disponible gratuitement[8].

Buscaldi et Rosso (Buscaldi et Rosso 2008a) ont évalué leur méthode basée sur la densité conceptuelle en utilisant l'ontologie WordNet comme un inventaire de sens, et le corpus GeoSemCor.

WordNet (Miller 1995) est une large base de données lexicale disponible aussi bien en anglais qu'en d'autres langues. Les mots dans WordNet sont relier les uns aux autres par une variété de relations sémantiques, parmi elles l'*holonymie* et sa relation inverse la *méronymie* qui sont les relations les plus significatives pour les toponymes.

[8] D'après une communication personnelle avec Jochen Leidner.

Les mots en WordNet sont groupés en 4 catégories : les noms, les verbes, les adjectifs et les adverbes. Les noms à leur tour sont classifiés en 26 catégories. Les toponymes se retrouvent parmi les noms de 2 classes: *Location* et *Object*. La classe *Location* contient des noms désignant une position spatiale, mais la classe, objet, contient des noms désignant des objets naturels.

Le corpus GeoSemCor –présenté pour la première fois dans (Buscaldi et Rosso 2008a)– est une version de SemCor (Miller, Leacock, et al. 1993) où chaque toponyme est annoté par son référent correct dans WordNet (voir Figure 4-3). Ce corpus est disponible gratuitement sur la page personnelle de Buscaldi[9]. Le Tableau 4-3 donne quelques informations à propos de GeoSemCor.

```
geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done
pos=NN lemma=atlanta wnsn=1 lexsn=1:15:00::>Atlanta</wf>
geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done
pos=NN lemma=georgia wnsn=1 lexsn=1:15:00::>Georgia</wf>
geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done
pos=NN lemma=atlanta wnsn=1 lexsn=1:15:00::>Atlanta</wf>
geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done
pos=NN lemma=georgia wnsn=1 lexsn=1:15:00::>Georgia</wf>
geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done
pos=NN lemma=atlanta wnsn=1 lexsn=1:15:00::>Atlanta</wf>
geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done
pos=NN lemma=atlanta wnsn=1 lexsn=1:15:00::>Atlanta</wf>
geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done
pos=NN lemma=atlanta wnsn=1 lexsn=1:15:00::>Atlanta</wf>
geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done
pos=NN lemma=georgia wnsn=1 lexsn=1:15:00::>Georgia</wf>
geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done
pos=NN lemma=savannah wnsn=1 lexsn=1:15:00::>Savannah</wf>
geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done
pos=NN lemma=texas wnsn=1 lexsn=1:15:00::>Texas</wf>
geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done
pos=NN lemma=georgia wnsn=1 lexsn=1:15:00::>Georgia</wf>
geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done
pos=NN lemma=georgia wnsn=1 lexsn=1:15:00::>Georgia</wf>
geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done
pos=NN lemma=georgia wnsn=1 lexsn=1:15:00::>Georgia</wf>
```

Figure 4-3. Les toponymes du fichier br-a01 du corpus GeoSemCor annotés avec leurs sens dans WordNet. La combinaison de `lemma` et `lexsn` permet de relier le toponyme avec son sens

[9] http://users.dsic.upv.es/grupos/nle/downloads.html

Tableau 4-3. Informations à propos le corpus GeoSemCor

Nombre total des toponymes	1210
Nombre des toponymes ambigus	498
Nombre de documents	123
Nombre moyen de toponymes par document	9,84
Nombre de toponymes sans duplications dans le même document	693
Nombre moyen de toponymes par document sans duplication	5,20
Nombre de toponymes dupliqué avec des référents différents dans le même document	13

Étant donné que WordNet n'est pas une source de connaissances purement géographiques, elle n'est pas aussi riche de toponymes et de référents pour chaque toponyme que les gazetteers. Le Tableau 4-4 fournit des toponymes pris du corpus GeoSemCor et des toponymes de quelques wilayas d'Algérie et compare leur nombre de référents récupérés du WordNet (version 2.1) et du Gazetteer Getty.

De son coté, GeoSemCor n'est pas compilé pour évaluer la tâche de DT, il est plutôt construit pour la tâche de désambiguïsation des sens des mots. Par conséquent, ces deux ressources ne sont pas vraiment adaptées à la tâche de désambiguïsation des toponymes.

Toutefois, Nous avons choisi d'évaluer notre heuristique en utilisant ces ressources. Cela est pour deux raisons. D'un coté, ce sont les seules ressources de DT gratuitement disponible[10], et de l'autre coté cela nous permet de comparer

[10] GeoSemCor est disponible dans l'adresse http://users.dsic.upv.es/grupos/nle/downloads.html et WordNet dans l'adresse http://wordnet.princeton.edu

notre méthode à celle de Buscaldi et Rosso (2008a) qui ressemble à la notre dans le fait qu'elle puisse détecter des relations non hiérarchiques entre les toponymes.

Tableau 4-4. Comparaison du nombre de référents pour certains toponymes dans WordNet et le Gazetteer Getty

Toponyme	Nombre de référents dans WordNet	Nombre de référents dans le gazetteer Getty
China	2	264
Georgia	3	74
New York	3	104
Paris	2	102
Palestine	2	44
Russia	4	14
Annaba	1	3
Constantine	1	17
Mila	0	4
Oran	1	14

4.3.2 Expérimentations

4.3.2.1 Objectifs et métriques d'évaluation

Nous avons implémenté notre méthode en utilisant le langage Perl et nous avons réalisé un ensemble d'expérimentations pour les buts suivants:

- Vérifier l'hypothèse de l'existence des relations arborescentes entre les toponymes du même contexte
- Étudier l'effet de la détection des relations de méronymie vs toutes les relations arborescentes sur les performances de la désambiguïsation des toponymes
- Comparer les performances de notre méthode avec d'autres.

L'estimation des performances des méthodes de désambiguïsation des toponymes se fait par les métriques utilisées dans les domaines de la recherche d'information

et le traitement automatique des langues naturelles. Ces métriques sont : la précision, le recall, la couverture, et F-mesure. Ils se calculent dans le domaine de la DT comme montré dans les équations (4), (5), (6) et (7) respectivement.

$$\text{Précision} = \frac{\text{nombre de toponymes résous correctement}}{\text{nombre de toponymes résous}} \quad (4)$$

$$\text{Recall} = \frac{\text{nombre de toponymes résous correctement}}{\text{nombre total de toponymes}} \quad (5)$$

$$\text{Couverture} = \frac{\text{nombre de toponymes résous}}{\text{nombre total de toponymes}} \quad (6)$$

$$\text{F} - \text{mesure} = \frac{2 * \text{Précision} * \text{Recall}}{\text{Précision} + \text{Recall}} \quad (7)$$

4.3.2.2 Résultats et analyse

Le Tableau 4-5 fourni les résultats d'expérimentations.

Tableau 4-5. Résultats d'évaluation en utilisant WordNet et GeoSemCor

	Précision	Recall	Couverture	F-mesure
DG (freq + SC)	88,2%	**87,4%**	**99,0%**	**0,878**
SC (H2)	**90,8%**	78,3%	86,3%	0,841
DC (H9)	89,9%	77,5%	86,2%	0,832
Map (H6)	87,9%	70,2%	79,9%	0,781

La ligne DG représente les résultats de notre méthode basée sur la densité géographique. Cette dernière –comme c'est expliqué précédemment– est la somme de la fréquence du référent et le score du chevauchement de son chemin hiérarchique avec le contexte. La ligne SC représente les résultats d'expérimentations avec le score du chevauchement seulement (voir l'heuristique H2). La ligne nommée DC représente les résultats de la méthode de Buscaldi et Rosso (2008a) qui est basée sur la densité conceptuelle. Map indique les résultats

de la méthode de Smith et Crane (2001). Cette dernière est basée sur la détection des relations spatiales entre les référents des toponymes (voir l'heuristiques H6 dans le chapitre précédent). Les résultats de ces 4 méthodes sont obtenus en utilisant le corpus GeoSemCor. Les linges DC et Map sont prise des articles (Buscaldi et Rosso 2008a) et (Buscaldi and Rosso 2008c) en considérant tous les toponymes du document comme contexte.

Les résultats d'expérimentation montre que la plus grande précision est celle de la méthode SC, cela veut dire que l'occurrence des holonymes d'un toponyme ambigu dans le contexte est le plus précis indicateur de son sens (son référent).

La couverture et le recall en utilisant la densité géographique (qui quantifie le degré de toutes les relations arborescentes) sont plus élevés par rapport à ceux de SC. Cela confirme que la recherche des relations hiérarchiques de type méronymie (quantifiés par SC) n'est pas suffisante pour désambiguïser tous les toponymes du contexte (pourtant elle donne des résultats précis). Il est donc plus performant de détecter tout les types de relations arborescentes pour désambiguïser le plus grand nombre de toponymes.

La couverture et le recall de notre méthode sont considérablement élevés par rapport à ceux des méthodes basées sur la DC (Buscaldi et Rosso 2008a) (+9,9%, +12,8% respectivement) et sur les calculs spatiaux (Smith & Crane, 2001)(+17,2%, +19,1% respectivement). Cependant, pas de différences significatives entre les précisions de ces trois méthodes (-1 ,9% et +0,3 la différence entre la précision de méthode DG et les méthodes DC et MAP respectivement).

De surcroît, les valeurs de toutes les mesures de notre méthode ont dépassé la valeur 80%, ce qui indique de bonnes performances.

4.4 Rapport entre le nombre de toponymes dans le contexte et les performances de la DT

Puisque notre heuristique utilise tous les toponymes du contexte. Nous avons réalisé une autre expérimentation pour étudier le rapport entre le nombre de toponymes du contexte et les performances de la désambiguïsation des toponymes. Mais les résultats ont prouvé qu'il n'y a pas une telle corrélation comme c'est illustré dans la Figure 4-4.

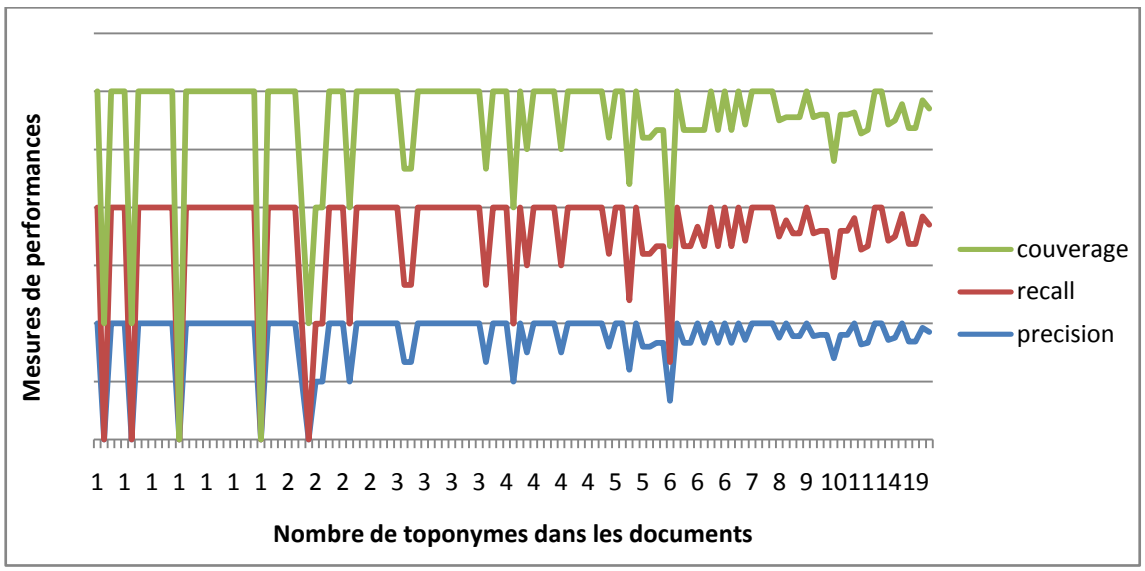

Figure 4-4. Rapport entre le nombre de toponymes et les performances de la DT : pas de corrélation significative

4.5 Conclusion

Nous avons classifié dans ce chapitre les différentes relations géographiques qui peuvent exister entre les toponymes du même contexte en se basant sur notre propre analyse des travaux de la littérature. En plus, nous avons proposé une nouvelle heuristique de désambiguïsation des toponymes. Notre heuristique est basée sur l'hypothèse de l'existence des relations géographiques arborescentes entre les toponymes du même contexte. Donc, elle résout les toponymes ambigus par les référents les plus reliés entre eux dans l'arbre hiérarchique des lieux du monde. Pour quantifier le degré de cette relation nous avons introduit une mesure

de corrélation géographique que nous avons appelé la *densité géographique* (DG), cela est par analogie à la densité conceptuelle (DC) utilisée pour la désambiguïsation des sens des mots, et appliquée par Buscaldi et Rosso (2008) pour la DT.

L'évaluation de notre heuristique en utilisant WordNet et GeoSemCor a montré la validité de notre hypothèse et la performance de notre heuristique. En outre, la comparaison de notre méthode à celle basée sur le SC a montré que la détection des relations de méronymies –qui est une idée utilisée dans quelques méthodes de l'état de l'art– est une heuristique précise mais n'est pas suffisante pour désambiguïser tous les toponymes d'un texte donné.

La comparaison de notre méthode avec celle de Smith et Crane (2001) (voir H6) a montré que la désambiguïsation des toponymes en cherchant une proximité arborescente est plus précise et plus performante que la désambiguïsation en s'appuyant sur la proximité en terme de distance.

Finalement, il faux reconnaître que les ressources GeoSemCor et WordNet nous ont permis d'évaluer notre méthode en la comparant à d'autres mais, à vrai dire ces deux ressources ne sont pas vraiment dédiées à la tâche de désambiguïsation des toponymes. Nul doute que l'utilisation d'autres ressources va nous permettre de mieux évaluer notre méthode.

Conclusion générale

Résumé de 24 mois de recherche

L'ordre des chapitres de ce mémoire reflète l'ordre chronologique des différentes étapes que nous avons connu durant notre chemine de recherche qui a commencé par l'exploration d'un large domaine qui est le data mining spatial et a terminé par une contribution dans un domaine spécifique qui est la désambiguïsation des toponymes.

Notre premiers pas dans cette recherche était de faire une synthèse sur le domaine du data mining spatial. Durant cette phase nous avons découvert que le data mining spatial est un domaine très large et sa largeur a plusieurs aspects.

Premièrement, c'est une extension du data mining sur les données spatiales, ce qui nous a obligé de se documenté dans deux domaine : le data mining d'un côté et les bases de données spatiales d'un autre côté.

Deuxièmement, le data mining –et à fortiori le data mining spatial– est un domine pluridisciplinaire, il se situe dans l'intersection de trois disciplines qui sont la statistique (avec ses trois branche inférencielle, descriptive et mathématique), les bases de données, et l'intelligence artificielle en particulier l'apprentissage machine. Cette nature pluridisciplinaire du data mining nous a fait passer beaucoup de temps pour se familiariser avec son jargon dérivé de plusieurs disciplines en particulier la statistique inférencielle qui était problématique pour nous autant qu'informaticien. Durant cette phase nous avons publié un article (Bensalem & Kholladi, 2008) sur les différents aspects de la relation entre le data mining et la statistique qui est un sujet de débat entre les chercheures informaticiens et statisticiens.

Troisièmement, le data mining y compris le data mining spatial n'est pas une simple technique mais plutôt un processus de plusieurs phases (voir Section 1.5).

Chaque phase est un domaine de recherche qui a ses propres notions, techniques et problèmes.

Le quatrième aspect de la largeur du data mining est la multiplicité de ses domaines d'application, qui varient entre la science, l'environnement, l'économie, la communication, le Web, …, etc.

Parmi cet éventail de sujets, nous avons choisi d'investiguer dans la première phase du data mining spatial qui est la collecte et la préparation de données géographiques. Parmi les sujets de recherche dans cette phrase, il y a l'intégration de données depuis plusieurs sources, et parmi ces sources il y a les documents textuels en langue naturelle. Ces raffinements nous ont conduits finalement vers la problématique de la désambiguïsation des toponymes.

Malgré le fait que la désambiguïsation des toponymes a des relations avec plusieurs autres domaines, nous étions contraints de consacrer le premier chapitre pour discuter précisément sa relation avec le data mining spatial du moment qu'il est le domaine de départ de notre recherche. Une partie de notre article (Bensalem & Kholladi, 2009b) discute cette relation.

Après notre décision de s'investiguer dans le domaine de la désambiguïsation des toponymes, nous avons affronté de nouveau la contrainte de la pluridisciplinarité. En fait, la DT partage plusieurs techniques avec la désambiguïsation des sens des mots et l'extraction des entités nommées qui sont des sous-domaines de la discipline du traitement automatique des langues naturelles, et aussi avec le géocodage et le géoparcing qui sont des sous-domaines des systèmes d'informations géographiques. En plus elle sert au géo-référencement des documents textuels qui permet l'indexation géographique des documents au sein d'un système de recherche d'information. Le chapitre 2 a discuté la position de la DT par rapport à ces domaines. Cela a permis d'un coté de bien exhiber l'utilité de cette tâche dans plusieurs applications, et d'un autre coté, de se familiariser avec son jargon multidisciplinaire.

Pour préparer le chapitre 3, nous avons analysé des dizaines de méthodes de l'état de l'art de la DT. Au début, ces méthodes nous ont apparu complètement différentes, mais par induction nous avons trouvé qu'elles partagent 4 composants, qui sont le contexte, les heuristiques, les connaissances, et les ressources. Cela nous a inspiré l'idée d'articuler l'état de l'art selon ces 4 axes. De plus, nous avons élaboré des classifications des heuristiques[1] et des connaissances, et nous croyions que notre état de l'art est complément de celui de (Leidner, 2007).

L'analyse des méthodes de l'état de l'art nous a permis de remarquer que beaucoup de méthodes sont basées implicitement sur l'idée que les référents des toponymes du même contexte sont proches géographiquement les uns des autres. En outre, nous avons distingué deux types de relations géographiques : les relations *spatiales,* qui résultent des proximités en termes de distance, et les *relations arborescentes* qui résultent des proximités dans l'arbre hiérarchique des lieux du monde. Contrairement aux relations spatiales, les relations arborescentes ne sont pas exploitées explicitement dans les méthodes existantes de DT. Notre contribution consiste à proposer une heuristique de désambiguïsation des toponymes qui est basée sur la quantification de ce type de relations, et ainsi elle porte remède à la dite lacune des méthodes existantes.

L'évaluation de notre heuristique a prouvé la validité de l'idée de désambiguïsation en exploitant les relations arborescentes et en plus elle a montré la performance de notre méthode par rapport d'autres. Notre heuristique ainsi que les résultats de son évaluation seront publié prochainement dans (Bensalem & Kholladi, 2009c)[2].

Il convient de noter que l'évaluation est encore problématique dans ce domaine à cause du manque de corpus standards dédiés à cette tâche. En effet, nous avons contacté une vingtaine d'auteurs pour l'obtention de leurs corpus. Finalement,

[1] L'idée de classifier les heuristiques est inspirée de (Leidner, 2007), mais notre classification est différente de la sienne.

[2] Ce papier est accepté et sera publié dans la conférence ACIT à décembre prochain (si Allah le Veut).

nous avons choisi de travailler sur GeoSemCor qui est gratuitement disponible sur le Web mais il a l'inconvénient de ne pas être vraiment adapté à la tâche de DT.

Perspectives

La désambiguïsation des toponymes est encore un terrien fertile pour la recherche. Dans ce qui suit nous présentons un ensemble de perspectives :

- Étudier les performances de notre heuristique au sein d'un processus de recherche d'information géographique. En effet, beaucoup d'auteurs réalisent ce type d'études en utilisant leurs heuristiques de désambiguïsation comme (Stokes, Li, Moffat, & Rong, 2008), (Overell & Rüger, 2007).
- Étudier l'effet de la taille (nombre de toponymes et nombre de référents pour chaque toponyme) et de la granularité des gazetteers dans la désambiguïsation des toponymes.
- Appliquer la désambiguïsation des toponymes sur des textes en langue arabe, ce qui implique la construction des gazetteers et des corpus d'évaluation en langue arabe. En outre, il est indispensable dans ce cas d'adapter les techniques de l'identification des toponymes dans le texte à la langue arabe. En effet, la reconnaissance des entités nommées (y compris les toponymes) en langue arabe est le sujet de plusieurs articles comme (Nezda, Hickl, Lehmann, & Fayyaz, 2006) dans la littérature du TALN.
- Bénéficier de Wikipedia comme source de données géographiques pour construire automatiquement un gazetteer multilingue.
- Comparer l'ambiguïté des toponymes dans différentes langues: arabe, anglais, français, puis faire des études sur la possibilité de tirer avantage de la différence éventuelle du taux d'ambiguïté des toponymes entre les langues pour proposer d'autres heuristiques de désambiguïsation.

Annexe A : Références de base

Domaine	Références
Data mining	(Han et Kamber 2006)
Désambiguïsation des toponymes	(Leidner 2007)
Désambiguïsation des sens des mots	(Ide et Véronis 1998) (Navigli 2009)
Informations géographiques et système et d'information géographique	(Longley, et al. 2005) (Laurini 1996) (disponible dans la bibliothèque centrale de UMC)
Géo-référencement	(Hill 2006)

Annexe B : Fonction de calcul de la Densité Géographique écrite en Perl

```
sub geographical_density
{   @toponyms = ();
    @toponyms  = @_;
    %tab_frequence = ();
    %topo_file = ();
    foreach $topony (@toponyms)
    {   @topo_hierars = ();
        @topo_hierars = get_topo_hierar($topony);
        @{$topo_file{$topony}}{@topo_hierars} = () x
@topo_hierars;
       #@affich = keys %{$topo_file{$topony}};
        #print "les hirar de $topony sont @affich \n";
        @components = ();
        @components = get_hierars_components(@topo_hierars);
        foreach $r (@components)
        { $tab_frequence{$r}++;  }
    }
    @tab_f = ();
    @tab_f = %tab_frequence;
    foreach $topony (keys %topo_file) #reference to an
hierarchis
    {   $ref_h = $topo_file{$topony};
        @tab_hie =();
        @tab_hie = keys %{$ref_h}; # retrieve hierarchis of a
toponym
        foreach $h (@tab_hie)
        {   @tab_h = ();
            @tab_h = split (/>/,"$h");
            $length_h = @tab_h;
            $topo_file{$topony}{$h}{"frequence"} =
referent_frequence($h,\%tab_frequence);
            $topo_file{$topony}{$h}{"SC"} =
context_inter_score($h,\@toponyms);
            $topo_file{$topony}{$h}{"length_h"} = $length_h;
            $topo_file{$topony}{$h}{"DG"} =
$topo_file{$topony}{$h}{"frequence"} +
$topo_file{$topony}{$h}{"SC"};
        }
    }
    return \%topo_file;}
```

Annexe C : Le toponyme ambigu ‘Georgia’ dans les fichiers de WordNet et le corpus GeoSemCor

Data.noun	08889889 15 n 02 Georgia 1 Sakartvelo 0 008 @i 08578498 n 0000 #p 08401715 n 0000 #m 08181367 n 0000 + 03148509 a 0101 %p 08890235 n 0000 %p 08890396 n 0000 %p 08890614 n 0000 %m 09587708 n 0000 | a republic in Asia Minor on the Black Sea separated from Russia by the Caucasus mountains; formerly an Asian soviet but became independent in 1991 08945623 15 n 04 Georgia 0 Empire_State_of_the_South 0 Peach_State 0 GA 0 018 @i 08534691 n 0000 #p 08915715 n 0000 #m 08920565 n 0000 #m 08921379 n 0000 + 03148282 a 0101 -r 01266860 n 0000 %p 08946257 n 0000 %p 08946399 n 0000 %p 08946706 n 0000 %p 08946835 n 0000 %p 08947280 n 0000 %p 08947398 n 0000 %p 08947538 n 0000 %p 09109867 n 0000 %p 09123267 n 0000 %p 09148673 n 0000 %p 09243465 n 0000 %p 09318270 n 0000 | a state in southeastern United States; one of the Confederate states during the American Civil War 08946145 15 n 01 Georgia 2 001 @i 08918800 n 0000 | one of the British colonies that formed the United States
Index.noun	georgia n 3 6 @ #m #p %m %p - 3 2 08512235 08512738 08459739 georgia n 3 7 @ #m #p %m %p + - 3 2 08945623 08946145 08889889
Index.sense	georgia%1:15:00:: 08945623 1 17 georgia%1:15:01:: 08889889 3 0 georgia%1:15:02:: 08946145 2 1
GeoSemCor	geosemcor2.0/brown1/tagfiles/br-a01:<wf geo=true cmd=done pos=NN lemma=georgia wnsn=1 lexsn=1:15:00::>Georgia</wf> geosemcor2.0/brown1/tagfiles/br-h01:<wf geo=true cmd=done rdf=georgia pos=NN lemma=georgia wnsn=1 lexsn=1:15:00::>Ga.</wf> geosemcor2.0/brown2/tagfiles/br-g17:<wf geo=true cmd=done pos=NN lemma=georgia wnsn=2 lexsn=1:15:02::>Georgia</wf>

Bibliographie

Nombre de références : 66

Agirre, E., & Rigau, G. (1996). Word sense disambiguation using conceptual density. *Proceedings of the 16th conference on computational linguistics (COLING '96)* (pp. 16–22). Copenhaghen: Association for Computational Linguistics.

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: geotagging web content. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 273 - 280). New York: ACM.

Aufaure, M.-A., Yeh, L., & Zeitouni, K. (2000). Fouille de données spatiales. *Le temps, l'espace et l'évolutif en Sciences du Traitement de l'Information* .

Azimi, A., & Delavar, M. (2007). Quality assessment in spatial clustering of data mining. *Proceedings of the 5th International Symposium on Spatial Data Quality (ISSDQ2007).* Enschede, The Netherlands.

Bensalem, I., & Kholladi, M. K. (2009b). L'utilisation des chemins hiérarchiques des lieux pour la désambiguïsation des toponymes. *Quaterième Atelier sur les Systèmes Décisionnels (ASD 2009).* Jijel.

Bensalem, I., & Kholladi, M. K. (2009a). La désambiguïsation des toponymes par la densité géographique. *International Conference on Applied Informatics (ICAI'09).* Bordj Bou Arréridj, Algérie.

Bensalem, I., & Kholladi, M. K. (2009c). Toponym Disambiguation by Arborescent relationships. *International Arab Conference on Information Technology (ACIT'2009).* Yemen.

Bensalem, I., & Kholladi, M. K. (2008). نظرة شاملة: التقنيات في المجالات و علاقته الإحصاء *Premières Journées Scientifiques en Informatique (JSIO'08).* Communication Poster.Oran, Algérie.

Borges, K. A., Laender, A. H., Medeiros, C. B., Silva, A. S., & Davis, C. A. (2003). The web as a data source for spatial databases. *Anais do V Brazilian Symposium on Geoinformatics, Campos do Jordão.* SP, Brazil.

Bunescu, R. C. (2007). *Learning for information extraction: From named entity recognition and disambiguation to relation extraction.* Thèse de doctorat de philosophie, University of Texas, Austin.

Buscaldi, D., & Rosso, P. (2008). A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science , 22* (3), 301-313.

Buscaldi, D., & Rosso, P. (2008a). A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science , 22* (3), 301-313.

Buscaldi, D., & Rosso, P. (2008c). Map-based vs. knowledge-based toponym disambiguation. *Proceeding of the 2nd international workshop on Geographic information retrieval, Napa Valley, California, USA* (pp. 19-22). ACM.

Chinchor, N. (1998). MUC-7 named entity task definition (version 3.5). *Proceedings of the 7th Message Understanding Conference (MUC-7).* Fairfax, Virginia.

Clough, P. (2005). Extracting metadata for spatially-aware information retrieval on the Internet. Dans C. Jones, & R. Purves (Éd.), *Proceedings of the ACM Workshop on Geographic Information Retrieval (GIR) held at the Conference on Information and Knowledge Management (CIKM)* (pp. 25-30). ACM Press.

Collier, N., Doan1, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., et al. (2008). BioCaster: detecting public health rumors with a Web-based text mining system. (J. Wren, Ed.) *Bioinformatics , 24* (24), 2940–2941.

Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM , 1* (39), 80–91.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). knowledge discovery and data mining: Towards a unifying Framework. *Proceedings of the Second International Conference on Knowledge Discovery* (pp. 82-88). California: AAAI Press.

Gaizauskas, R., Humphreys, K., Azzam, S., & Wilks, Y. (1997). Concepticons vs. lexicons: An architecture for multilingual information extraction. Dans P. M. Teresa (Éd.), *Information extraction: A multidisciplinary approach to an emerging information technology , International Summer School, SCIE-97, Frascati, Italy, 14-18, 1997* (Vol. 1299, pp. 28-43). Berlin: Springer-Verlag.

Gale, W., Church, K., & Yarowsky, D. (1992). One sense per discourse. *Proceedings of the Fourth DARPA Speech and Natural Language Workshop* (pp. 233–237). San Mateo, CA: Morgan Kaufmann.

Garbin, E., & Mani, I. (2005). Disambiguating toponyms in news. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)* (pp. 263-270). Vancouver: Association for Computational Linguistics.

Gardarin, G. (1999). *Internet/intranet et bases de données: Data Web, Data Media, Data Warehouse, Data Mining.* Eyrolles.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques.* San Francisco: Morgan Kaufmann.

Hauptmann, A. G., & Olligschlaeger, A. M. (1999). Using location information from speech recognition of television news broadcasts. In T. Robinson, & S. Renals (Ed.), *Proceedings of the ESCA ETRW Workshop on Accessing Information in Spoken Audio* (pp. 102–106). Cambridge, England: University of Cambridge.

Hill, L. L. (2006). *Georeferencing: The geographic associations of information.* Cambridge, MA, USA: The MIT Press.

Ide, N., & Véronis, J. (1998). Word sense disambiguation: Stat of the art. *Computational Linguistics , 24* (1).

Kalashnikov, D. V., Ma, Y., Mehrotra, S., Hariharan, R., & Butts, C. (2006). Modeling and querying uncertain spatial information for situational awareness applications. *Proceedings of the 14th annual ACM international symposium on Advances in Geographic Information Systems (ACM GIS)* (pp. 131 - 138). ACM.

Kalashnikov, D. V., Ma, Y., Mehrotra, S., Hariharan, R., Venkatasubramanian, N., & Ashish, N. (2006). SAT: Spatial Awareness from Textual input. *Proceedings of the 10th International Conference on Extending Database Technology (EDBT), Munich, Germany.*

Larson, R. R. (1996). Geographic information retrieval and spatial browsing. In L. Smith, & M. Gluck (Eds.), *Geographic information systems and libraries: Patrons, maps, and spatial information : [papers presented at the 1995 Clinic on Library Applications of Data Processing, April 10-12, 1995]* (pp. 81–123). Urbana-Champaign, USA: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

Laurini, R. (1996, juin 10). Bases de données géographiques. *Technique de l'ingénieur (Référence H3758)* .

Leidner, J. L. (2006). An evaluation dataset for the toponym resolution task. *Computers,Environment and Urban Systems , 30* (4), 400–417.

Leidner, J. L. (2007). *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names.* PhD dissertation, University of Edinburgh, Institute for Communicating and Collaborative Systems, School of Informatics.

Leidner, J. L. (2004). Towards a reference corpus for automatic toponym resolution evaluation (Extended abstract). *Proceedings of the Workshop on Geographic Information Retrieval held at the 27th Annual International ACM SIGIR Conference (SIGIR)*, (p. pages unnumbered). Sheffield, England, UK.

Leidner, J. L., Sinclair, G., & Webber, B. (2003). Grounding spatial named entities for information extraction and question answering. *Proceedings of the Workshop on the Analysis of Geographic References held at the Joint Conference for Human Language Technology and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics 2003 (HLT/NAACL 2003)*, (pp. 31-38). Edmonton, Alberta, Canada.

Li, H., Srihari, R. K., Niu, C., & Li, W. (2003). InfoXtract location normalization: a hybrid approach to geographic references in information extraction. *Proceedings of the HLT-NAACL 2003 Workshop: Analysis of Geographic References* (pp. 39–44). Edmonton, Alberta, Canada: Association for Computational Linguistics.

Li, H., Srihari, R. K., Niu, C., & Li, W. (2002). Location normalization for information extraction. *Proceedings of the 19th international conference on Computational linguistics . 1*, pp. 1-7. Morristown, NJ, USA: Association for Computational Linguistics .

Li, Y., Moffat, A., Stokes, N., & Cavedon, L. (2006). Exploring Probabilistic Toponym Resolution for Geographical Information Retrieval. *Proceedings of SIGIR Workshop on Geographical Information Retrieval*, (pp. 17-22). Seattle, Washington.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2005). *Geographical information systems and science* (2 ed.). England: John Wiley & Sons Ltd.

MetaCarta, Inc. (2008). *MetaCarta GSRP processing and indexing with the georeferencing engine*. Consulté le Août 27, 2009, sur http://www.metacarta.com/products-platform-indexing.htm

MetaCarta, Inc. (n.d.). Retrieved from Geographic search and referencing solutions - MetaCarta - At the forefront of the GeoWeb: http://www.metacarta.com/

Miller, G. A. (1995). WordNet: a Lexical database for english. *Communication of the ACM , 38* (11), 39-41.

Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T. (1993). A semantic concordance. *Proceedings of the ARPA Workshop on Human Language Technology* (pp. 303-308). Princeton, New Jersey : Association for Computational Linguistics.

Miller, H. J. (2007). Geographic data mining and knowledge discovery. In J. Wilson, & A. S. Fotheringham (Eds.), *Handbook of geographic information science* (pp. 352-366). Wiley-Blackwell.

Miller, H. J., & Han, J. (2001). *Geographic data mining and knowledge discovery.* CRC Press.

Morimoto, Y., Aono, M., Houle, M. E., & McCurley, K. S. (2003). Extracting spatial knowledge from the web. *Proceedings of the 2003 Symposium on Applications and the Internet (SAINT'03)* (p. 326). Los Alamitos, CA, USA: IEEE Computer Society.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys , 41* (2), 69 pages (10:1-10:69).

Nezda, L., Hickl, A., Lehmann, J., & Fayyaz, S. (2006). What in the World is a shahab? Wide coverage named entity recognition for Arabic. *Proccedings of the 5th edition of the international conference on language ressources and evaluation (LERC 2006)*, (pp. 41-46). Genoa, Italy.

Ng, R. T., & Han, J. (1994). Efficient and effective clustering methods for spatial data mining. *Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 144 - 155). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. .

Overell, S. E., & Rüger, S. (2007). Geographic co-occurrence as a tool for GIR. *Proceedings of the 4th ACM workshop on Geographical information retrieval, Lisbon, Portugal* (pp. 71-76). New York, NY, USA: ACM Press.

Overell, S. E., & Rüger, S. (2006a). Identifying and grounding descriptions of places. *Third Workshop on Geographic Information Retrieval, SIGIR 2006.* ACM Press.

Overell, S., Magalhães, J., & Rüger, S. (2006b). Place disambiguation with co-occurrence models.

Pekar, V., Krkoska, M., & Staab, S. (2004). Feature weighting for co-occurrence-based classification of words. *Proceedings of the 20th international conference on Computational Linguistics, Geneva, Switzerland .* Morristown, NJ, USA: Association for Computational Linguistics .

Pouliquen, B., Steinberger, R., Ignat, C., & Groeve, T. D. (2004). Geographical information recognition and visualization in texts written in various languages. *Proceedings of the 2004 ACM Symposium on Applied Computing* (pp. 1051–1058). ACM Press.

Pyle, D. (2003). Data Collection, Preparation, Quality, and Visualization. In N. Ye (Ed.), *The handbook of data mining* (pp. 366-391). Mahwah, New Jersey, USA: Lawrence Erlbaum Associates, Inc.

Rauch, E., Bukatin, M., & Baker, K. (2003). A confidence-based framework for disambiguating geographic terms. *HLTNAACL 2003 Workshop: Analysis of Geographic References* (pp. 50-54). Edmonton, Alberta, Canada: Association for Computational Linguistics.

Rijsberg, C. V. (1979). *Information retrieval.* Oxford: Butterworths.

Rosso, P., Masulli, F., Buscaldi, D., Pla, F., & Molina, A. (2003). Automatic noun sense disambiguation. Dans A. Gelbukh (Éd.), *Computational linguistics and intelligent text processing: 4th International Conference, CICLing 2003 Mexico City, Mexico, February 16–22, 2003 Proceedings , 2588 of Lecture Notes in Computer Science* (pp. 273–276). Berlin: Springer.

Sanderson, M., & Kohler, J. (2004). Analyzing geographic queries. *Proceedings of the Workshop on Geographic Information Retrieval. 27th Annual International ACM SIGIR Conference.* Sheffield, UK.

Schilder, F., Versley, Y., & Habel, C. (2004). Extracting spatial information: grounding, classifying and linking spatial expressions. *InWorkshop on Geographic Information Retrieval held at the Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Sheffield, England, UK: Association for Computing Machinery.

Shekhar, S., & Chawla, S. (2003). *Spatial databases: a tour (Draft copy).* Prentice Hall.

Shekhar, S., Zhang, P., Huang, Y., & Vatsavai, R. R. (2004). Trends in spatial data mining. In H. Kargupta, A. Joshi, K. Sivakumar, & Y. Yesha (Eds.), *Data Mining: Next Generation Challenges and Future Directions.* AAAI Press.

Sinha, R., & Mihalcea, R. (2007). Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. *Proceedings of the International Conference on Semantic Computing (ICSC'07)* (pp. 363-369). Washington, DC, USA: IEEE Computer Society.

Smith, D. A., & Crane, G. (2001). Disambiguating geographic names in a historical digital library. *Research and Advanced Technology for Digital Libraries: Fifth European Conference (ECDL 2001)*, (pp. 127–136).

Smith, D. A., & Crane, G. (2001). Disambiguating geographic names in a historical digital library. *Research and Advanced Technology for Digital Libraries: Fifth European Conference (ECDL 2001)*, (pp. 127–136).

Smith, D. A., & Mann, G. S. (2003). Bootstrapping toponym classifiers. *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References. 1*, pp. 45-46. Morristown, NJ: Association for Computational Linguistics.

Stokes, N., Li, Y., Moffat, A., & Rong, J. (2008). An empirical study of the effects of NLP components on Geographic IR performance. *International Journal of Geographical Information Science , 22* (3), 247-264.

Volz, R., Kleb, J., & Mueller, W. (2007). Towards ontology-based disambiguation of geographical identifiers. *Proceedings of 16th International World Wide Web Conference (WWW2007).* Banff, Canada.

Zheng, D., Zhao, T., Li, S., & Yu, H. (2007). Research on a novel word co-occurrence model and its application. Dans *Knowledge science, engineering and management* (pp. 437-446). Berlin / Heidelberg: Springer.