Services Web pour l'annotation sémantique d'information spatiale à partir de corpus textuels

Ludovic Moncla, Mauro Gaio

Services Web pour l'annotation sémantique d'information spatiale à partir de corpus textuels

Ludovic Moncla[1], **Mauro Gaio**[2]

1. *Institut de Recherche de l'Ecole Navale (IRENav), CC 600, 29240 Brest Cedex 9, France*
   ludovic.moncla@ecole-navale.fr
2. *Laboratoire d'informatique, Université de Pau et des Pays de l'Adour (LIUPPA), 64000 Pau, France*
   mauro.gaio@univ-pau.fr

RÉSUMÉ. *L'annotation sémantique d'information spatiale a pour objectif de repérer des mots ou des syntagmes décrivant des références géographiques (noms de lieux) ainsi que diverses expressions spatiales associées. L'une des importantes difficultés pour concevoir un système automatique d'annotation d'un tel type d'information est due aux ambiguïtés liées aux entités spatiales. Une approche modulaire basée sur des services Web a été choisie. La méthodologie proposée repose sur la combinaison d'une étape de prétraitement (analyse morphosyntaxique), d'une cascade de transducteurs, et d'une étape de classification utilisant des ressources du Web des données. Un avantage de cette approche est la possibilité d'obtenir des traitements partiels ou encore de mettre en concurrence certains modules réalisant la même tâche.*

ABSTRACT. *The semantic annotation of spatial information aims to identify words or phrases describing geographical references (place names) as well as various associated spatial expressions. One of the major difficulties in designing an automatic annotation system for such information is due to ambiguities related to spatial entities. A modular approach based on Web services was chosen. The proposed methodology is based on the combination of a pre-processing step using external morphosyntactic analysers, a cascade of transducers, and a classification step using linked data. An advantage of this approach is the possibility to obtain partial processing or to evaluate competing methods doing the same task.*

MOTS-CLÉS : *Annotation sémantique, Services Web, Reconnaissance d'entités nommées*

KEYWORDS: *Semantic annotation, Web services, Named entity recognition*

1. Introduction

Dans cet article nous présentons une approche modulaire basée sur différents services Web pour l'annotation automatique des entités nommées (EN) et des informations spatiales associées, dans des textes descriptifs. Le terme « approche modulaire » peut faire référence aux travaux de Roulet sur l'analyse du discours (Roulet, 1991). Bien que le parallèle avec ces travaux soit intéressants sur certains aspects et concepts, nous faisons ici référence, à la théorie des systèmes complexes et à la conception de modules logiciels. Nous appréhendons l'approche modulaire comme permettant de décomposer un problème complexe en sous-problèmes indépendants et liés entre eux.

Deux types d'approches existent pour l'annotation automatique des EN : les approches linguistiques ou symboliques à base de règles et les approches probabilistes centrées sur les données et les techniques d'apprentissage (Poibeau, 2011). Ces deux types d'approches initialement présentées comme concurrentes, coexistent de plus en plus dans des systèmes hybrides. L'approche symbolique repose sur la description lexicale et syntaxique des syntagmes recherchés. Les EN sont repérées grâce à la construction de patrons lexico-syntaxiques utilisant des marqueurs lexicaux, et des dictionnaires. De nombreuses méthodes d'annotation développées selon l'approche symbolique (Maurel *et al.*, 2011), utilisent des transducteurs à états finis pour modéliser et implémenter les patrons lexico-syntaxiques (Poibeau, 2003). Un transducteur est un automate à états finis qui agit sur un texte par des insertions, des remplacements ou des suppressions. Ils peuvent être exécutés en cascade, de cette manière les annotations réalisées par un transducteur peuvent être utilisées par les transducteurs suivants.

L'annotation sémantique ajoute des informations complémentaires à des textes non-structurés, elle peut permettre en particulier d'identifier et de relier les entités du texte avec les données du Web sémantique. L'annotation sémantique d'information spatiale a pour objectif de repérer des mots ou des syntagmes décrivant des références géographiques (noms de lieux) ainsi que diverses expressions associées faisant référence à l'espace dans la langue (Aurnague *et al.*, 1997) tel que les expressions de relations spatiales, de déplacement ou de position. L'annotation sémantique des EN et des expressions de relations spatiales, de déplacement ou de position peut par exemple être utilisée dans l'objectif d'une interprétation cartographique de tout ou partie de descriptions textuelles (Moncla *et al.*, 2016). Les principales difficultés pour concevoir un système d'annotation automatique sont les ambiguïtés inhérentes au langage naturel, en particulier ici à celles liées aux entités spatiales. Un nombre important de types d'entités spatiales existe, tel que les entités géopolitiques (pays, divisions administratives), les lieux habités (villes, adresses, codes postaux) et les entités de nature géographique (parcs, vallées, montagnes, rivières). Sans être la seule, cette diversité typologique est l'une des sources provoquant des situations d'ambiguïté.

La désambiguïsation des EN spatiales est considérée comme une sous-tâche de la résolution des toponymes (Leidner, 2007), elle consiste à associer une localisation non-ambiguë à un nom de lieu. Des ressources géographiques de type « gazetiers » (index géographique) contribuent fréquemment à la réalisation de cette tâche. Depuis quelques années, de nombreuses ressources ont émergées tel que Geonames [1], OpenStreetMap [2], ou Wikimapia [3]. Dans un contexte de données ouvertes et de plateformes participatives, ces ressources ont connu une très forte croissance et sont accessibles grâce à des services Web et aux technologies du Web des données (Linked Data). Mais ce nombre et cette diversité de plateformes, tout en rendant des services d'une qualité croissante, compliquent l'utilisation de ces données et le choix des ressources qu'il faut au préalable sélectionner.

Nous présentons dans cet article une approche modulaire instanciée sous la forme de services Web dédiés aux différentes phases de l'annotation sémantique d'informations spatiales. La méthodologie proposée est décrite en section 2. Elle est composée, d'une analyse morphosyntaxique présentée en section 2.1, d'une cascade de transducteurs pour l'annotation des EN et des informations spatiales associées décrite en section 2.2 et, d'une utilisation des Linked Data pour la classification des EN ainsi que d'une désambiguïsation des EN spatiales décrites en section 2.3. La section 3 présente les résultats de l'évaluation de notre approche sur un corpus de description de randonnées. Enfin la section 4 conclut cet article.

2. Méthodologie

Une chaîne de traitement est définie comme une séquence de traitements connectés par des données d'entrée/sortie. Une approche modulaire permet une adaptation plus simple et rapide de la chaîne à de nouvelles contraintes. Par exemple, cela permet d'adapter la chaîne pour le traitement d'une nouvelle langue en ne modifiant que les modules nécessaires. Cette approche nous a permis de concevoir une première version pour l'analyse de corpus en français puis avec un minimum de modifications d'obtenir des versions adaptées pour des corpus en espagnol et en italien. Par ailleurs, les différents services Web proposés peuvent être appelés indépendamment les uns des autres dans d'autres chaînes de traitement existantes, ou dans le cadre d'un traitement particulier. Cela permet également d'exécuter un traitement additionnel sur des données pré-annotées par un autre traitement automatique ou manuel. Enfin, un dernier avantage de cette approche est la possibilité de mettre en concurrence certaines solutions réalisant la même tâche afin de choisir celle obtenant les meilleurs résultats.

1. www.geonames.org
2. www.openstreetmap.org
3. www.wikimapia.org

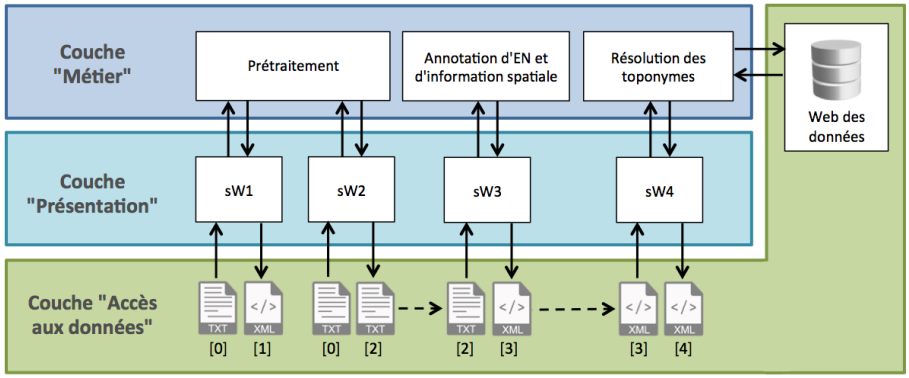

FIGURE 1 – Architecture logicielle de notre proposition

Notre approche modulaire repose sur trois principaux modules de traitement et propose quatre services Web pour l'annotation sémantique d'information spatiale à partir d'un corpus textuel. La conception de l'architecture de cette approche suit les principes de la programmation orientée composant et d'une architecture 3-tiers (figure 1).

La couche « accès aux données » fournit un accès aux données d'entrée de la chaîne de traitement. Il s'agit de données textuelles issues de textes bruts [0] ou annotés [2] et [3], ainsi que de données géographiques issues de ressources externes (Web des données). La couche « métier » regroupe les différents modules de traitement. Les trois modules principaux sont un module de prétraitement, un module d'annotation des EN et des informations spatiales associées et un module de classification des EN utilisant des ressources du Web des données. Enfin la couche « présentation » propose à l'utilisateur l'accès aux services Web. Les services Web que nous proposons permettent d'interroger en ligne les différents modules de traitement. Ils acceptent des requêtes de type POST et GET et peuvent donc être utilisés directement depuis un navigateur internet ou un programme tiers, tel qu'une chaîne de traitement UIMA [4]. Ils prennent par défaut trois paramètres en entrée, la clé d'API attribuée lors de l'enregistrement, la langue du document et enfin le contenu textuel à analyser. Ces services Web sont décrits dans les sections suivantes.

2.1. Prétraitement

L'objectif de cette étape de prétraitement est de préparer les textes bruts pour les étapes suivantes. Il simplifie le fonctionnement des autres modules en apposant dans les textes, grâce à des traitements, des marques pouvant

4. uima.apache.org

indiquer: le découpage en phrases, la segmentation en mots, la catégorie morphosyntaxique des mots et l'ajout de leurs lemmes. Ces quatre traitements sont communément réalisés par des analyseurs morphosyntaxiques qui sont bien entendu dépendants de la langue.

Les résultats issus de ce module de prétraitement ont une influence non négligeable sur la qualité des traitements qui seront effectués par les modules suivants. Nous avons choisi de nous appuyer sur des analyseurs morphosyntaxiques existants afin de pouvoir sélectionner celui obtenant les meilleurs résultats en fonction de nos attentes et de la langue analysée. Pour cela, dans le cadre de nos expérimentations nous avons sélectionné trois analyseurs différents, Treetagger [5] et Freeling [6] qui sont compatibles avec le français et l'espagnol et d'autre part Talismane [7] disponible uniquement pour le français mais ayant de meilleurs résultats eu égard à nos attentes.

```
Poursuivre  VER:infi  poursuivre
par         PRP       par
le          DET:ART   le
pont        NOM       pont
de          PRP       de
la          DET:ART   le
Glière      NAM       <unknown>
.           SENT      .
```

(a) Treetagger

```
Poursuivre poursuivre VMN0000 1
par par DA0FS0 0.972269
le le DA0MS0 1
pont pont NCMS000 1
de de SPS00 1
la le DA0FS0 1
Glière NP00000 1
. . Fp 1
```

(b) Freeling

```
0   Poursuivre   poursuivre   VINF   v       _
1   par          par      P   P      _
2   le           le       DET DET    g=m|n=s
3   pont         pont     NC  nc     g=m|n=s
4   de           de           P      P       _
5   la           la       DET DET    g=f|n=s
6   Glière       _        NPP _      _
7   .            .        PONCT PONCT _
```

(c) Talismane

FIGURE 2 – Exemple de résultat généré par les analyseurs morphosyntaxiques

D'après la figure 2, on remarque que les trois analyseurs utilisent chacun un format de sortie différent. De la même manière les étiquettes permettant d'identifier les catégories grammaticales des mots diffèrent d'un analyseur à l'autre ainsi que d'une langue à une autre pour un même analyseur afin de tenir compte des spécificités de chaque langue. Par conséquent, le module de prétraitement

5. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
6. http://nlp.lsi.upc.edu/freeling/
7. http://redac.univ-tlse2.fr/applications/talismane.html

(figure 1) est spécialement conçu pour s'adapter aux sorties fournies par différents analyseurs, et permet de produire en sortie un étiquetage standardisé grâce à une transformation générique (conçue pour être adaptable en fonction des attentes). Un fichier de configuration indique l'analyseur utilisé, en fonction de la langue du document et des attentes de l'utilisateur. Le module de prétraitement lance l'exécution de l'analyseur spécifié avec en entrée le texte brut [0] fourni par la couche « données d'entrée ». Dans le cadre de nos travaux, nous avons défini, *via* la table de correspondance, un ensemble d'étiquettes unique permettant d'homogénéiser les résultats des analyseurs. Comme on peut le remarquer dans la (tableau 1) l'ensemble des étiquettes retenues peut être modifié et adapté pour s'adapter à d'autres attentes.

TABLE 1 – Etiquettes grammaticales utilisées par le système

Etiquette	Description	Etiquette	Description
A	adjectif	PREP	préposition
ABR	abréviation	PREPDET	préposition + déterminant
ADV	adverbe	PUN	ponctuation
CONJC	conjonction	PRO	pronom
DET	déterminant	PRO+POS	pronom possessif
N	nom	PRO+REL	pronom relatif
NPr	nom propre	SYM	symbole
NUM	numérique	V	verbe

Le service Web « sW1 » (figure 1) transforme la sortie du module de prétraitement dans un format XML [1] suivant les recommandations TEI [8] (Text Encoding Initiative). Ce résultat peut être utilisé par un service tier nécessitant un prétraitement réalisé par un analyseur morphosyntaxique. La figure 3 montre un exemple de résultat standardisé retourné par ce service. Le second service permettant d'interroger le module de prétraitement sera décrit dans la section suivante. Il propose un résultat au format texte adapté au module d'annotation automatique des EN et des informations spatiales associées.

```
<s>
  <w lemma="poursuivre" type="V">Poursuivre</w>
  <w lemma="par" type="PREP">par</w>
  ...
  <w lemma="le" type="DET">la</w>
  <w type="NPr">Glière</w>
</s>
```

FIGURE 3 – Sortie du service Web de prétraitement « sW1 »

8. TEI-C http://www.tei-c.org

2.2. Annotation automatique d'entités nommées et d'informations spatiales

Du point de vue de la structure syntaxique, (Rangel Vicente, 2005) différencie deux catégories d'EN. Les EN dites fortes composées exclusivement de noms propres et les EN dites faibles constituées par un nom propre et une forme catégorisante. Cette notion d'EN fortes ou faibles s'appuie sur l'opposition entre noms propres purs et noms propres descriptifs introduite par (Jonasson, 1994). Les outils de reconnaissance d'EN classiques tel que OpenNLP [9], OpenCalais [10] ou CasEN (Friburger, Maurel, 2004), considèrent généralement les EN comme étant des EN fortes. Certains outils, comme CasEN, utilisent le contexte d'apparition des noms propres pour construire des grammaires locales et catégoriser les EN mais n'incluent pas la nature du référent du nom propre au sein de l'EN. (Sekine *et al.*, 2002) introduisent la notion d'entité nommée étendue qui propose une classification typologique complexe contenant plusieurs centaines de types d'EN. Il s'agit d'une classification enrichie par rapport à celle proposée lors des compagnes d'évaluation MUC mais qui ne permet pas d'intégrer le contexte immédiat des noms propres au sein des EN.

Ainsi, nous avons redéfini la notion d'entité nommée étendue (ENE) afin de structurer et hiérarchiser les EN en tenant compte des formes catégorisantes (Gaio, Moncla, 2017); l'objectif étant de les intégrer directement au sein de l'EN. Une ENE est une entité construite à partir d'un nom propre associé à un ou plusieurs termes exprimant sa nature (forme catégorisante). Les ENE sont définies comme une imbrication de niveau en fonction du nombre de termes associés.

(1) Charles de Gaulle

(2) général Charles de Gaulle

(3) avenue du général Charles de Gaulle

Via cette notion d'ENE, il est possible de représenter à la fois des EN fortes et les EN faibles. Par exemple l'EN (1) est une ENE de niveau 0 car elle est composée uniquement d'un nom propre (EN forte). L'EN (2) est une ENE de niveau 1 car elle associe le terme « général » à une ENE de niveau 0 (EN Faible). De même, l'ENE (3) est une ENE de niveau 2 car elle associe le terme « avenue » à une ENE de niveau 1. On note ici l'importance du concept d'imbrication des ENE, où chaque nouvelle imbrication est utilisée pour préciser la nature de l'ENE. En effet l'ENE (3) et de nature géographique bien qu'elle soit composée d'une EN faisant référence à une personnalité. Ce concept permet de structurer le contexte d'apparition immédiat des noms propre et sera donc très important pour l'étape de classification des EN.

9. http://opennlp.apache.org
10. http://www.opencalais.com

Par ailleurs il est également nécessaire d'annoter les informations spatiales associées aux EN, telles que les relations spatiales ou les évènements de déplacement. Nous nous appuyons sur les travaux de (Nguyen *et al.*, 2013) qui ont introduit les structures VT formalisant les relations entre les verbes de déplacements ou de perception, les relations spatiales et les EN. L'exemple (4) montre une structure VT composée du verbe de déplacement « Marcher », d'une mesure de distance « 10 km », d'une préposition spatiale « jusqu'au » et d'une ENE « refuge des Barmettes ». Notre objectif est d'annoter la sémantique et le rôle de chacun de ces éléments dans la phrase.

(4) Marcher 10 km jusqu'au refuge des Barmettes.

Le module d'annotation des EN et des informations spatiales associées se compose d'une cascade de transducteurs. Pour développer notre cascade, nous avons suivi des principes introduits par (Maurel *et al.*, 2011) pour le développement de l'outil d'annotation CasEN. CasEN est un système conçu pour la reconnaissance des EN qui s'exécute entièrement au sein de la plateforme Unitex [11]. Il utilise des ressources lexicales telles que le dictionnaire de la langue cible (ici le français) ou le dictionnaire des noms propres et des descriptions locales de motifs (transducteurs). Unitex est une plateforme logiciel permettant de traiter des textes en langues naturelles en utilisant des ressources linguistiques. Il présente les transducteurs sous la forme de graphes ce qui permet une prise en main simple pour l'écriture et la maintenance des différentes règles d'annotation.

Nous proposons d'encapsuler la cascade d'annotation au sein d'une approche modulaire afin de séparer les étapes d'annotation et de classification mais également afin de pouvoir utiliser le résultat de l'analyse morphosyntaxique en entrée de la cascade. Cette proposition permet de réduire les ambiguïtés dues au fait que plusieurs catégories grammaticales peuvent être associées à un même mot. En effet, les ressources lexicales utilisées par défaut par Unitex associent à chaque mot toutes ses catégories grammaticales possibles, cela ne permet donc pas de connaître la catégorie grammaticale du mot dans le contexte particulier d'un texte donné. Un autre inconvénient de l'utilisation de ces ressources lexicales, est leur non-exhaustivité (en particulier le dictionnaire des noms propres utilisé pour la classification) mais également par le fait d'être des ressources locales (situées dans des fichiers en local) hors Web et donc difficilement maintenables par une communauté.

Unitex accepte deux types de documents en entrée : un texte brut ou un texte annoté. Lors de l'utilisation de textes bruts Unitex applique ses propres graphes de prétraitements ainsi que différentes ressources lexicales. Les textes pré-annotés permettent par exemple d'utiliser en entrée d'une cascade le résultat d'une autre cascade. Afin que nos transducteurs puissent utiliser le résultat

11. http://www-igm.univ-mlv.fr/~unitex/

de l'analyse morphosyntaxique, le résultat est retourné par le module de prétraitement en un format compatible avec Unitex.

Ce traitement est réalisé par le service Web « sW2 » comme le montre la figure 4.

```
{Marcher,marcher.V} {10,.NUM} {km,kilomètre.ABR}
    {jusqu',jusque.PREP} {au,au.PREPDET} {refuge,refuge.N}
    {des,du.PREPDET} {Barmettes,.NPr}
```

FIGURE 4 – Format de sortie du module de prétraitement adapté à Unitex

Le module d'annotation est composé de 80 transducteurs pour lesquels nous distinguons deux catégories : les transducteurs principaux qui ont pour rôle d'annoter les éléments en ajoutant de l'information directement au contenu textuel, et les sous-graphes qui peuvent être utilisés par les transducteurs principaux ou par d'autres sous-graphes. Ces sous-graphes contiennent des lexiques tels que la liste des verbes de déplacement ou des règles génériques telles que des expressions régulières mais ils n'ajoutent aucune information directement au contenu textuel.

Concernant cette étape d'annotation, de la même manière que CasEN, nous proposons une combinaison de deux cascades. La première, appelée *cascade d'analyse*, est la cascade principale. Elle exécute une séquence de transducteurs qui annotent les éléments dans un ordre spécifique. En effet, les éléments annotés par les transducteurs précédents peuvent être imbriqués dans des annotations d'éléments plus grands réalisées par les transducteurs suivants. La deuxième, appelée *cascade de synthèse* a pour rôle de transformer le résultat de la première cascade dans le format d'annotation souhaité.

La cascade d'analyse exécute cinq transducteurs principaux et produit un résultat au format XML défini par le programme CasSys d'Unitex. Le premier transducteur annote les relations spatiales exprimées dans le texte telles que les distances, les relations topologiques ou les directions. Le deuxième annote les verbes, plus précisemment les verbes de déplacement et de perception. Ce transducteur permet également d'associer une sémantique à l'annotation réalisée, en particulier pour les verbes de déplacement leur polarité (Boons, 1987 ; Laur, 1991). Le troisième transducteur annote les ENE composées de noms propres et de termes associés (Gaio, Moncla, 2017). Enfin, le dernier transducteur annote les expressions de déplacement ou de perception, qui sont repérées à l'aide de la formalisation des structures VT (Nguyen *et al.*, 2013) mettant en relation les différents éléments annotés par les transducteurs précédents.

Cette première cascade est dépendante de la langue, nous avons donc développé une version adaptée pour chacune des langues traitées par notre système (le français, l'espagnol et l'italien). Les différences entre les trois versions de cette cascade d'analyse se limitent à la traduction des lexiques utilisés dans les sous-graphes, tel que la liste des verbes de déplacement ou la liste des prépo-

sitions spatiales. En effet, les règles décrites par les principaux transducteurs restent les mêmes pour les trois langues traitées, toutes étant des langues romanes.

Une fois le résultat de la première cascade obtenu, la cascade de synthèse a pour rôle de transformer ce résultat dans un format plus interopérable. Nous utilisons le langage d'annotation proposé par (Moncla, Gaio, 2015) qui s'appuie sur le standard TEI. La figure 5 montre le résultat simplifié (sans la balise `<w>`) produit par **le service Web « sW3 »** faisant appel au module d'annotation.

```
<phr type="verb_phrase" subtype="motion">
  Marcher
  <measure type="distance">10 km<\offset>
  <offset type="direction" subtype="final">jusqu'au<\offset>
  <rs n="1">
    <term type="N">refuge</term>
    des
    <rs n="0">
      <name>Barmettes</name>
    </rs>
  </rs>
</phr>
```

FIGURE 5 – Résultat produit par le module d'annotation

Le résultat contient les annotations des ENE non catégorisées (balise `<rs>`) et les différentes informations spatiales associées à ces ENE [3].

2.3. Classification des ENE et résolution des toponymes

Comme nous l'avons vu, à la fin du processus d'annotation automatique nous obtenons des ENE non catégorisées. A la différence des méthodes d'annotation d'EN existantes qui catégorisent les EN selon une typologie prédéfinie (Sekine *et al.*, 2002 ; Ehrmann, 2008), notre objectif est uniquement de distinguer les entités qui font référence à un lieu. Pour réaliser cet objectif nous proposons, grâce à un module complémentaire, une solution interrogeant des ressources du Web des données. Cette solution a deux objectifs, le premier est de classer les entités selon qu'elles soient spatiales ou non spatiales, le second est de récupérer les coordonnées géographiques associées aux entités spatiales. Nous utilisons les services web REST fournis par GeoNames [12] et l'API nominatim [13] proposée par OpenStreetMap afin d'interroger ces ressources. Concernant les ressources institutionnelles, les ressources officielles espagnoleet italienneproposent un accès aux données géographiques grâce au protocole WFS (Web Feature Service) défini par l'OGC. Les données fournies par l'IGN pour la

12. http://www.geonames.org/export/web-services.html
13. http://nominatim.openstreetmap.org/search

France ne sont pas directement accessibles en ligne, nous avons donc téléchargé et installé ces données dans une base de données locale PostGIS. L'interrogation de plusieurs ressources ayant des caractéristiques différentes a des avantages et des inconvénients. Un avantage important est d'avoir accès à une couverture, une granularité et une exhaustivité des données plus importantes. En revanche cela peut augmenter le risque d'erreurs (faux positifs) et les doublons.

Le fait qu'une ENE soit répertoriée dans une base de données géographiques ne permet pas de savoir avec certitude si l'entité énoncée dans le texte fait référence à un lieu ou non. En effet, il peut s'agir d'un cas de métonymie ou bien d'un cas où le nom de lieu est utilisé dans un contexte non géographique. Ce problème d'ambiguïté spatial/non spatial a été défini sous le terme *referent class ambiguity* par (Smith, Mann, 2003). Il existe également différentes formes d'ambiguïtés liées au problème de classification des EN tel qu'un lieu ayant plusieurs noms (*reference ambiguity*) ou un même nom pouvant désigner plusieurs lieux (*referent ambiguity*). Il existe un nombre important de méthodes (Buscaldi, 2011) proposées pour la désambiguïsation des toponymes (EN spatiales). Comme à notre connaissance il n'existe pas de solution unique, les méthodes sont la plupart du temps adaptées à une catégorie de texte en particulier, tel que les textes historiques (Smith, Crane, 2001), les *news* (Garbin, Mani, 2005) ou les descriptions de randonnées (Moncla *et al.*, 2014). Notre approche modulaire et l'utilisation de services Web simplifient l'intégration de modules complémentaires adaptés aux spécificités du corpus analysé. Ici l'objectif est donc de pouvoir intégrer des modules de désambiguïsation qui soient adaptés à la catégorie des documents et aux spécifications de la tâche à accomplir.

Le module de résolution et de désambïguation des ENE implante un algorithme générique par défaut s'appuyant sur les méthodes proposées dans (Moncla *et al.*, 2014). Après avoir interrogé les ressources géographiques, celuici supprime les doublons introduits par l'utilisation de plusieurs ressources puis s'il n'y a pas de résultats et que l'ENE est de niveau 0 alors il considère que l'ENE est de type non-spatiale. Mais s'il s'agit d'une ENE d'un niveau supérieur à 0 alors le module interroge une nouvelle fois les ressources avec l'ENE de niveau n-1 imbriquée dans la précédente. Si le terme appartenant à une ENE correspond à la nature géographique exprimée dans les métadonnées fournies par la ressource géographique pour l'ENE de niveau inférieur, alors l'ENE est considérée comme spatiale. Le module récupère également les autres métadonnées, telles que les coordonnées géographiques. Par ailleurs lorsque l'ENE de niveau inférieur n'existe pas dans les ressources géographiques le module recherche le terme dans un thesaurus voire dans une ontologie (si disponible) pour déterminer s'il s'agit d'un concept géographique. Dans ce dernier cas, l'ENE est classée selon la typologie spatial/non-spatial mais sans avoir de coordonnées géographiques associées. A la fin du processus, chaque ENE spatiale ayant des référents dans les ressources interrogées est associée à l'URI qui permet de faire le lien avec les ressources géographiques du Web des données.

Le quatrième service Web « sW4 » que nous proposons retourne le résultat obtenu par le module de résolution et de désambiguïsation des ENE (figure 6b). Il s'agit comme pour le service Web précédent d'un résultat au format XML respectant les recommandations TEI et contenant cette fois-ci les annotations sémantiques des ENE spatiales. La valeur de l'attribut *ref* de l'élément `<placeName>` (ici tronqué dans l'exemple) contient l'URI correspondante à la ressource du Web des données.

```
<rs n="1">
  <term>refuge</term>
  des
  <rs n="0">
    <name>Barmettes</name>
  </rs>
</rs>
```

(a) Avant classification

```
<placeName ref="#451703419">
  <geogName type="S" subtype="RHSE">
    <geogFeat>refuge</geogFeat>
    des
      <name>Barmettes</name>
  </geogName>
</placeName>
```

(b) Après classification

FIGURE 6 – Résultat de l'annotation avant et après la classification des ENE

3. Evaluation

Nous avons évalué notre proposition sur un corpus multilingue de descriptions de randonnées composé de 90 documents annotés manuellement, 30 documents pour chaque langue (français, espagnol et italien). Ce corpus contient 1556 ENE, dont 1525 de type spatiale, 47% des ENE spatiales sont associées à un verbe de déplacement et font référence à un évènement de déplacement. Par ailleurs, 47% des ENE spatiales sont composées d'une forme catégorisante (ENE de niveau >0).

Pour l'évaluation de l'annotation (reconnaissance et classification) des ENE nous utilisons 3 métriques. Le rappel qui mesure le ratio entre le nombre de réponses pertinentes données et le nombre de réponses pertinentes existantes, la précision qui mesure le ratio entre le nombre de réponses pertinentes données et le nombre total de réponses et enfin le Slot Error Rate (SER) qui permet de tenir compte des différents cas d'erreurs comme l'insertion, la suppression, la classification et les limites du balisage. À la différence du rappel et de la précision, plus le SER est bas, meilleure est la mesure. Les résultats pour l'annotation des ENE sur l'ensemble du corpus multilingue sont les suivants : rappel 98%, précision 93%, et SER 21%. Les résultats pour les trois langues traitées sont proches mais légèrement plus faible pour l'italien du fait de ressources moins exhaustives. Pour les documents français de notre corpus nous avons comparé les résultats de notre méthode avec ceux obtenus par l'outil d'annotation CasEN (version Quaero). CasEN obtiens les scores suivants : rappel 65%, précision 88% et SER 51%. Les scores pour les documents français avec notre méthode sont les suivants : rappel 96%, précision 95% et SER 17%. La grande différence dans les résultats s'expliquent par l'utilisation des données

liées (par rapport à la seule utilisation des ressources lexicales proposées au sein d'Unitex). Ces résultats encourageants montrent, entre autre, l'intérêt d'avoir adoptée une approche modulaire permettant de connecter différents services et différentes sources de données.

4. Conclusion

L'annotation fine des entités nommées de lieu et des informations spatiales associées nécessite des traitements spécifiques et des ressources particulières. Ce constat nous a conduit à proposer une approche modulaire basée sur plusieurs services Web. La méthodologie proposée repose sur la combinaison d'une étape de prétraitement utilisant des analyseurs morphosyntaxiques externes, d'une cascade de transducteurs, et d'une étape de classification utilisant des ressources du Web des données. Les services Web proposés sont spécialement conçus pour répondre aux besoins spécifiques ou encapsulent des outils préexistants. Ils peuvent par exemple être appelés indépendamment dans une chaîne de traitement tiers, tel qu'une chaîne UIMA. La cascade d'annotation est accessible au téléchargement [14] et le code source des modules de prétraitement et de classification est disponible sur un dépôt GitHub [15]. Les premiers résultats présentés dans cet article nous encouragent à proposer des évolutions du module de classification afin d'inclure une implémentation d'autres approches de désambiguïsation existantes, adaptées à d'autres types de textes.

Bibliographie

Aurnague M., Vieu L., Borillo A. (1997). Langage et cognition spatiale, sciences cognitives. In, p. 69–102. Denis, M.

Boons J.-P. (1987). La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs. *Langue Française*, vol. 76, n° 76, p. 5–40.

Buscaldi D. (2011, juillet). Approaches to disambiguating toponyms. *SIGSPATIAL Special*, vol. 3, n° 2, p. 16–19.

Ehrmann M. (2008). *Les entités nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation.* Thèse de doctorat non publiée, Paris 7 - Denis Diderot.

Friburger N., Maurel D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, vol. 313, n° 1, p. 93–104.

Gaio M., Moncla L. (2017). Extended named entity recognition using finite-state transducers: An application to place names. In *9th international conference on advanced geographic information systems, applications, and services.* Nice, France.

14. http://erig.univ-pau.fr/PERDIDO/
15. https://github.com/ludal360/Perdido

Garbin E., Mani I. (2005). Disambiguating toponyms in news. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, p. 363–370. Stroudsburg, PA, USA, ACL.

Jonasson K. (1994). *Le nom propre*. Duculot, Belgique, Louvain-la-Neuve.

Laur D. (1991). *Sémantique du déplacement et de la localisation en français: une étude des verbes, des prépositions et de leurs relations dans la phrase simple*. Thèse de doctorat non publiée, Toulouse 2.

Leidner J. L. (2007). *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Universal-Publishers.

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

Moncla L., Gaio M. (2015). A multi-layer markup language for geospatial semantic annotations. In *9th workshop on geographic information retrieval (gir'15)*. Paris, France.

Moncla L., Gaio M., Nogueras-Iso J., Mustière S. (2016). Reconstruction of itineraries from annotated text with an informed spanning tree algorithm. *International Journal of Geographicac Information Science*, vol. 30, n° 2.

Moncla L., Renteria-Agualimpia W., Nogueras-Iso J., Gaio M. (2014). Geocoding for texts with fine-grain toponyms: An experiment on a geoparsed hiking descriptions corpus. In *22nd ACM SIGSPATIAL international conference on advances in geographic information systems*, p. 183–192. Dallas,TX, USA, ACM.

Nguyen V. T., Gaio M., Moncla L. (2013). Topographic subtyping of place named entities: a linguistic approach. In *The 15th AGILE international conference on geographic information science*, p. 1–5. Louvain, Springer.

Poibeau T. (2003). Extraction automatique d'information: du texte brut au web sémantique. In *Extraction automatique d'information: du texte brut au web sémantique*. Hermès Lavoisier.

Poibeau T. (2011). *Traitement automatique du contenu textuel*. Lavoisier.

Rangel Vicente M. (2005). La glose comme outil de désambiguïsation référentielle des noms propres purs. *Corela, Numéros Spéciaux le traitement lexicographique des noms propres*.

Roulet E. (1991). *Vers une approche modulaire de l'analyse du discours*. Cahiers de linguistique française.

Sekine S., Sudo K., Nobata C. (2002). Extended named entity hierarchy. In *Proceedings of the third international conference on language resources and evaluation, LREC 2002, may 29-31, 2002, las palmas, canary islands, spain.*

Smith D. A., Crane G. (2001). Disambiguating Geographic Names in a Historical Digital Library. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, p. 127–136. London, UK, Springer.

Smith D. A., Mann G. S. (2003). Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003 workshop on analysis of geographic references - volume 1*, p. 45–49. Stroudsburg, PA, USA, ACL.