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. *INSA Lyon & CNRS, LIRIS UMR 5205, France*
   ludovic.moncla@liris.cnrs.fr
2. *Université de Pau et des Pays de l'Adour & CNRS, LMAP UMR 5142, 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 plus 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 pré-traitement (analyse morpho-syntaxique), 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 sémantique de l'information spatiale de manière automatisée. Nous nous intéresserons en particulier, au sein des textes descriptifs en langue française, aux entités nommées (EN) et aux informations qui leur sont associées. De manière générale, l'étape de détection dans le texte puis de reconnaissance des EN sont des étapes nécessaires pour toute tâche d'extraction d'information, devenant incontournable pour toute tâche de compréhension automatique des textes. Nous regrouperons ici ces deux étapes, repérage et reconnaissance, sous le terme d'annotation sémantique. La notion d'EN, quant à elle, est une notion complexe qui ne peut être définie qu'à un niveau sémantique. Pour cela, il faut au préalable identifier toutes les formes linguistiques, qui, à l'instar des noms propres, désignent de manière univoque une entité (Dupont *et al*., 2002). Comme précisé dans (Vicente, 2005), l'EN est la notion utilisée en TAL pour désigner les éléments discursifs mono-référentiels qui coïncident en partie avec les noms propres et qui suivent des patrons syntaxiques déterminés. Ainsi l'approche ici proposée apporte une réponse à ces préalables.

Dans toute tâche de compréhension automatique de texte, le contenu des textes pour être « compris » par le processus automatique nécessite d'être enrichi par des données externes. Ces données représentent tout ou partie des « connaissances » nécessaires à un humain pour comprendre ce même contenu. L'annotation sémantique ajoute ces informations complémentaires à des textes non-structurés, elle peut permettre en particulier d'identifier et de relier les entités énoncées dans le texte avec les données du Web sémantique. L'annotation sémantique des informations spatiales 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) telles que les expressions de relations spatiales, de déplacement ou de position. L'annotation sémantique des EN et des informations associées peut par exemple être utilisée dans l'objectif d'une interprétation cartographique de tout ou partie de descriptions textuelles (Moncla *et al*., 2016).

Quelque soit le type d'EN (mono-référence à un lieu ou à d'autres catégories d'objets du monde), 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. Ces transducteurs peuvent être exécutés en cascade, de

cette manière les annotations réalisées par un transducteur peuvent être utilisées par les suivants.

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, telles 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 émergé tel que Geonames [1], OpenStreetMap [2], ou Wikimapia [3]. Dans un contexte de données ouvertes et de plates-formes 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 plate-formes, 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.

Enfin, la plupart des approches symboliques s'appuient sur une succession de phases d'analyse. Il est admis que quelque soit les techniques employées ces approches adoptent quatre phases successives : l'analyse morpho-lexicale, qui s'intéresse à la structuration en mots ou groupes de mots ; l'analyse syntaxique, dédiée à l'analyse du rôle structurel des mots dans la phrase ; l'analyse sémantique, qui s'intéresse de tout ou partie du sens de tout ou partie des constituants d'une phrase (considérée individuellement) ; l'analyse pragmatique, qui s'attache à contextualiser autour des phrases.

Dans la pratique, d'une part, les systèmes implémentent très rarement l'ensemble de ces quatre phases. La plupart du temps les uns sont consacrés à l'analyse de la structure des phrases alors que d'autres tentent de comprendre les textes à partir du sens des mots. D'autre part, chacune des phases d'analyse s'appuie à leur tour sur une succession d'étapes. Prenons par exemple l'analyse morpho-lexicale se décompose en trois étapes : la segmentation, dont le but est le découpage du texte en mots distincts ; la lemmatisation, qui tente de déterminer la forme canonique des mots précédemment obtenus ; enfin l'étiquetage, dont le rôle est d'associer à chaque mot la bonne catégorie morpho-syntaxique (verbe, nom commun, adjectif…) selon le contexte de la phrase.

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

Cet ensemble d'observations nous a fait adopter une méthode de conception facilitant la modularité et la décomposition d'une chaîne de traitement d'annotation sémantique.

Nous présentons dans cet article l'instanciation de ce type d'approche sous la forme de services Web dédiés aux différentes phases de l'annotation sémantique de l'information spatiale. L'ensemble des services Web que nous proposons permet de décomposer le problème complexe d'annotation sémantique en sous-problèmes indépendants et complémentaires. La décomposition en services de la chaîne de traitement ainsi que l'ajout à terme de nouveaux services comme par exemple l'appariement entre une trace GPS et les EN spatiales d'un texte sont des tâches réalisées dans le cadre du projet ANR CHOUCAS [4]. La section 2 de cet article présente la problématique liée à l'annotation d'entités nommées. La méthodologie proposée est décrite en section 3. Elle est composée d'une analyse morpho-syntaxique présentée en section 3.1, d'une cascade de transducteurs pour l'annotation des EN et des informations spatiales associées décrite en section 3.2 et d'une utilisation des Linked Data pour la classification des EN décrites en section 3.3 ainsi que de services de post-traitement (géocodage, désambiguïsation et reconstruction d'itinéraire) décrits en section 3.4. La section 4 présente les résultats de l'évaluation de notre approche sur un corpus de description de randonnées. Enfin la section 5 conclut cet article.

2 Problématique

Les premiers travaux de recherche sur la tâche de détection et de reconnaissance des EN ont démarré dès le début des années 80, avec comme cible privilégiée les textes en langue anglaise. Ces travaux ont eu en particulier pour objectif d'identifier « les frontières » ou « les bornes » des EN mais également de définir leur catégorie (personne, lieu, organisation, ...). En effet, en Anglais comme dans la plupart des autres langues, il est nécessaire d'identifier de manière précise les bornes des EN et en particulier les éléments participant à la constitution d'un syntagme pouvant obtenir l'étiquette d'EN.

Dans les années 1990, lors des campagnes MUC (Message Understanding Conference) (Grishman, 1997) d'évaluation de systèmes automatiques de reconnaissance d'informations dans les textes, la notion d'EN n'a été considérée comme telle et donc définie que lors de la sixième campagne (MUC-6) en 1995. R. Grisman et B. Sundheim dans (Grishman, Sundheim, 1995) ont proposé, pour la tâche dédiée à cette notion, de travailler sur cinq catégories : les personnes, les organisations, les noms de lieux, les expressions temporelles et certains types d'expressions numériques. En ce qui concerne les trois premières catégories l'objectif est d'isoler et catégoriser le constituant « nom propre » qui compose l'EN et pour cela une même balise est disponible : ENAMEX (Entity Name Expression) prenant comme attribut l'une des trois

4. Dans le contexte de l'ANR CHOUCAS, il s'agit de proposer un ensemble d'outils pour repérer, extraire, reconnaître et désambiguïser les EN (en particulier de lieux) de sources textuelles décrivant tout type de parcours en montage.
http://choucas.ign.fr/

catégories. Dans les directives qui cadrent la tâche de reconnaissance des ENAMEX les auteurs ont précisé pour chacune des trois catégories des règles d'annotation qui permettent des les différencier les unes des autres. En ce qui concerne les noms de lieux, il est par exemple précisé qu'un nom de lieu évoqué à la suite immédiate d'un nom d'organisation doit selon la position de certains indicateurs textuels, soit être marqué indépendamment de l'organisation, soit s'effacer en la faveur de l'organisation, comme par exemple:

(1) "Hyundai of Korea, Inc."
<ENAMEX TYPE="ORGANIZATION">Hyundai of Korea, Inc.</ENAMEX>

(2) "Hyundai, Inc. of Korea"
<ENAMEX TYPE="ORGANIZATION">Hyundai, Inc.</ENAMEX> of <ENAMEX TYPE="LOCATION">Korea</ENAMEX>
(exemples tirés de (Grishman, Sundheim, 1995))

On voit ici que la décision est prise compte tenu de la position de l'indicateur textuel « Inc. ». Cet ensemble de règles fortement adapté aux méthodes utilisées à l'époque a très rapidement montré ses limites. Dans les deux exemples ci-dessus plusieurs problèmes sont suscités par ce marquage, comme l'obtention d'EN différents pour un même objet initial. Mais, l'un des principaux problèmes vient de la règle interdisant de construire des EN composées, c'est-à-dire des EN constituées d'un groupement d'EN pouvant appartenir à différentes catégories, comme par exemple :

(3) "Hyundai of Korea, Inc."
<ENAMEX TYPE="ORGANIZATION"><ENAMEX TYPE="ORGANIZATION">Hyundai </ENAMEX> of <ENAMEX TYPE="LOCATION">Korea</ENAMEX>, Inc.</ENAMEX>

Dans les décennies suivantes, différents travaux sur les EN ont abouti à en préciser la définition. En particulier dans le cadre du programme de recherche ACE (Automatic Content Extraction) qui a proposé des tâches plus ambitieuses (Doddington *et al.*, 2004) allant de la tâche de détection des événements en passant par celles de détection des relations entre entités. Une des conséquences a été l'évolution de la notion d'EN, celle-ci ne faisant plus uniquement référence strictement aux noms propres mais également aux groupes nominaux sans nom propre et pouvant décrire une entité particulière d'une part, et d'autre part par la prise en compte de la coréférence, comme par exemple celle assurée par les pronoms. Une autre conséquence a été la forte augmentation du nombre de catégories qui a fait émerger la notion de sous-typage. En 2008, dans le cadre du programme de recherche franco-allemand « Quaero », le principe d'entité nommée étendue fait son apparition (Grouin *et al.*, 2011). En plus du sous-typage, cette notion implique la nécessité de détecter la structure des entités. Il s'agit de caractériser les différents éléments composant une entité en s'appuyant sur leur rôle au sein de celle-ci. L'objectif étant de repérer et structurer des informations fines afin de permettre une meilleure constitution de bases de connaissances.

A l'étude des différentes typologies disponibles, qu'elles soient à caractère général ou spécialisé (Sekine *et al.*, 2002), nous pouvons constater qu'il y a une absence de consensus quant à la définition d'une EN. Nous constatons également qu'il n'y a pas non plus de consensus quant à une correspondance unique entre la notion

d'EN et une ou des expressions linguistiques. 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 tels que OpenNLP [5], OpenCalais [6] 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.

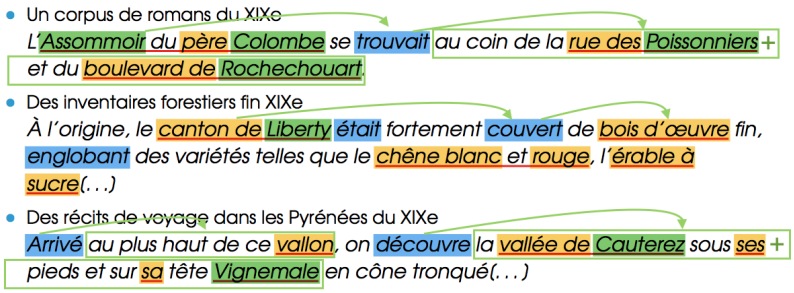
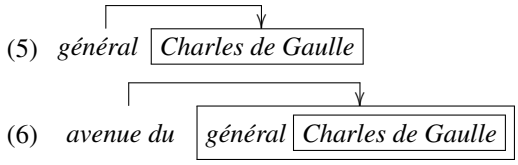


Figure 1. Annotations associant EN spatiales et relations dynamiques

Comme le montre de manière graphique la figure 1, notre problématique pour l'annotation des EN et des informations spatiales est de repérer dans le texte le maximum d'informations ayant un rôle dans la description de l'espace et du déplacement dans la langue. 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 niveaux de description fonction du nombre de termes associés.

(4) *Charles de Gaulle*

5. http://opennlp.apache.org
6. www.opencalais.com



Via cette notion d'ENE, il est possible de représenter à la fois des EN fortes et les EN faibles. Par exemple l'EN (4) est une ENE de niveau 0 car elle est composée uniquement d'un nom propre (EN forte). L'EN (5) est une ENE de niveau 1 car elle associe le terme « général » à une ENE de niveau 0 (EN Faible). De même, l'ENE (6) 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 (6) 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 propres et sera donc très important pour l'étape de classification des EN.

Par ailleurs dans cette proposition nous souhaitons également 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 [7] formalisant les relations entre les verbes de déplacements ou de perception, les relations spatiales et les EN. L'exemple (7) 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.

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

3 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.

7. L'acronyme VT signifiait à l'origine *Verbe Toponyme* (Nguyen *et al*., 2013), il a connu différentes évolutions afin d'étendre la notion de toponyme à celle de lieu.

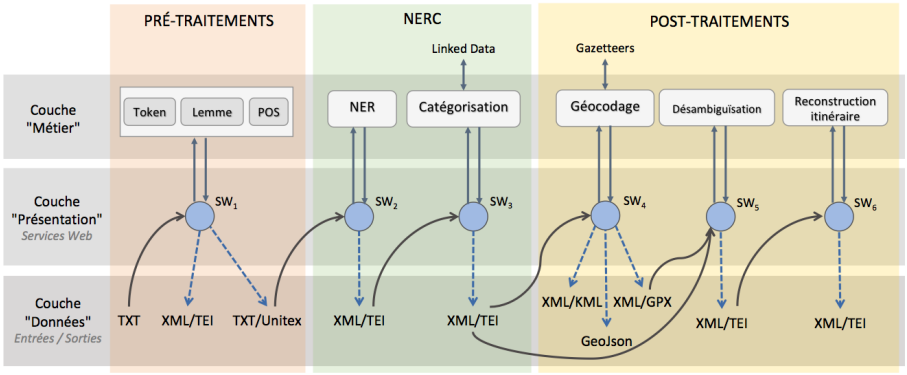


Figure 2. Architecture logicielle de notre proposition

Notre approche modulaire repose sur trois principaux modules, un module de pré-traitement et deux modules pour la reconnaissance et la classification des entités nommées et des informations spatiales. Notre proposition intègre également différents modules de post-traitements tels qu'un module de géocodage ou un module de désambiguïsation utilisant les annotations réalisées par les modules principaux. La conception de l'architecture de cette approche suit les principes de la programmation orientée composant et d'une architecture 3-tiers (figure 2).

La couche « données » fournit un accès aux données d'entrée de la chaîne de traitement. Il s'agit de données textuelles issues de textes bruts (TXT) ou annotés (XML), ainsi que de données géographiques issues de ressources externes (Web des données et gazetiers). La couche « métier » regroupe les différents modules de traitement. Les trois modules principaux sont un module de pré-traitement, 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. Nous présentons également trois modules de post-traitement qui ont pour rôle le géocodage des EN spatiales, la désambiguïsation des résultats produits par le service de géocodage, et la reconstruction d'itinéraire en utilisant les informations issues de la désambiguïsation et les informations spatiales annotées dans le texte. 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 peuvent donc être utilisés directement depuis un programme tiers, tel qu'une chaîne de traitement UIMA [8]. 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.

8. Unstructured Information Management Application : https://uima.apache.org/

3.1 Pré-traitement

L'objectif de cette étape de pré-traitement 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 indiquer : le découpage en phrases, la segmentation en mots, la catégorie morpho-syntaxique des mots et l'ajout de leurs lemmes. Ces quatre traitements sont communément réalisés par des analyseurs morpho-syntaxiques qui sont bien entendu dépendants de la langue.

Les résultats issus de ce module de pré-traitement 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 morpho-syntaxiques 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 [9] et Freeling [10] qui sont compatibles avec le français et l'espagnol et l'italien et d'autre part Talismane [11] 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

```
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
```

(c) Talismane

Figure 3. Exemple de résultat généré par les analyseurs morphosyntaxiques

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

D'après la figure 3, 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 pré-traitement (figure 2) 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 pré-traitement lance l'exécution de l'analyseur spécifié avec en entrée le texte brut (TXT) fourni par la couche « données ». Une fois le pré-traitement terminé, le résultat est transformé grâce à une table de correspondance afin d'être standardisé pour la suite des traitements.

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

Nous avons développé une table de correspondance pour chaque ensemble d'étiquettes dépendant de l'analyseur et de la langue utilisés. 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. Cet ensemble simplifié implique une perte d'information pour certaines spécificités de la langue afin de privilégier une homogénéisation. Comme on peut le remarquer dans le tableau 1 l'ensemble des étiquettes retenues peut être modifié et adapté pour s'adapter à d'autres attentes et permettre la représentation de toutes les catégories grammaticales existantes pour une langue donnée.

Le service Web SW$_1$ offre le choix entre deux formats de sortie différents pour l'étape de pré-traitement. Le premier est au format XML et suit les recommandations TEI [12] (Text Encoding Initiative). Ce service peut être utilisé de manière indépendante par un service tiers nécessitant un pré-traitement réalisé par un analyseur morpho-syntaxique. La figure 4 montre un exemple du résultat XML standardisé retourné par ce service.

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

```
<s>
  <w lemma="marcher" type="V">Marcher</w>
  <w type="NUM">10</w>
  <w lemma="kilomètre" type="ABR">km</w>
  <w lemma="jusque" type="PREP">jusqu'</w>
  <w lemma="au" type="PREPDET">au</w>
  <w lemma="refuge" type="N">refuge</w>
  <w lemma="du" type="PREPDET">des</w>
  <w type="NPr">Barmettes</w>
</s>
```

Figure 4. Sortie au format XML/TEI du service Web de pré-traitement « sW1 »

Le deuxième format de sortie proposé par ce service est compatible avec la plate-forme Unitex/GramLab [13] (figure 5) qui sera utilisée par le module de reconnaissance d'entités nommées détaillé dans la section suivante.

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

Figure 5. Sortie au format TXT/Unitex du service Web de pré-traitement « sW1 »

3.2 Annotation automatique d'entités nommées et d'informations spatiales

Le module d'annotation des EN et des informations spatiales associées se compose d'une cascade de transducteurs qui implémentent les grammaires ENE et VT décrites dans (Gaio, Moncla, 2017). La grammaire ENE décrit la notion d'ENE (présentée dans la section 2) qui étend le concept d'EN par une imbrication de différents niveaux permettant l'inclusion de termes décrivant ou précisant la nature de l'entité. La grammaire VT est une formalisation des relations entre les verbes de déplacements et de perception, les relations spatiales (prépositions spatiales, mesures de distances, orientation, direction, etc) et les entités de lieux servant à exprimer le déplacement ou la localisation dans la langue. Pour développer notre cascade, nous avons suivi les principes introduits par (Maurel *et al.*, 2011) lors du 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 plate-forme Unitex/GramLab. 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/GramLab est une plate-forme logicielle permettant de traiter des textes en langues naturelles en utilisant des ressources linguistiques. Elle 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. Unitex accepte deux types de documents en entrée : un texte brut ou un texte annoté. Lors de l'utilisation de textes bruts Unitex applique

13. http://unitexgramlab.org/

ses propres graphes de pré-traitements 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.

Nous proposons d'encapsuler la cascade d'annotation au sein d'une approche modulaire afin de rendre indépendantes les étapes d'annotation et de classification des ENE mais également afin de pouvoir utiliser le résultat de l'analyse morpho-syntaxique 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 le fait d'être des ressources locales (situées dans des fichiers en local) hors Web et donc difficilement maintenables par une communauté. Afin que nos transducteurs puissent utiliser le résultat de l'analyse morpho-syntaxique, le résultat est retourné par le service de pré-traitement SW_1 en un format compatible avec Unitex (figure 5).

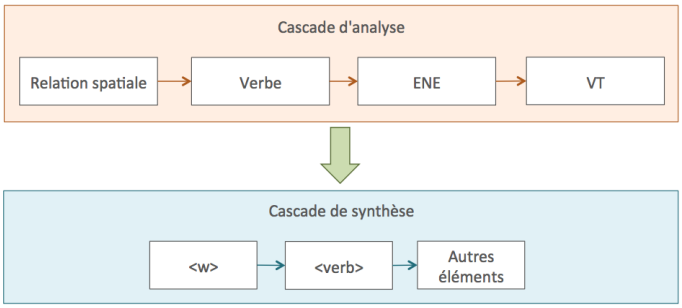


Figure 6. Transducteurs principaux de notre double cascade d'annotation

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 (figure 6). 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 trans-

former le résultat de la première cascade dans le format d'annotation souhaité (XML/-TEI).

La cascade d'analyse exécute quatre transducteurs principaux (figure 6) 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 (adjacence, inclusion) ou les relations cardinales et d'orientation (nord, sud, etc). Le deuxième annote les verbes, plus précisément 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 en fonction des prépositions avec lesquelles ils sont exprimés (Boons, 1987 ; Laur, 1991). Le troisième transducteur annote les ENE composées de noms propres et de termes associés comme défini dans la section 2 (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 tels que les verbes de déplacement ou de perception, les relations spatiales et les ENE. Cette première cascade est dépendante de la langue, elle utilise le résultat de l'analyse morpho-syntaxique ainsi que différents lexiques. 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, tels que la liste des verbes de déplacement ou la liste des prépositions spatiales. En effet, les règles décrites par les principaux transducteurs restent les mêmes pour les trois langues romanes traitées.

```
<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 7. Résultat produit par le module d'annotation

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 dans (Moncla, Gaio, 2015) qui s'appuie sur le standard XML/TEI. La figure 7 montre le résultat simplifié (sans la balise `<w>`) produit par **le service Web SW$_2$** faisant appel au module d'annotation. Le résultat contient les annotations des ENE non catégorisées (balises `<rs>` et `<name>`) et les différentes informations spatiales associées à ces ENE (tels que les balises `<offset>`).

3.3 Classification des ENE

Comme décrit dans la section précédente, le service d'annotation d'entités nommées retourne des ENE non catégorisées. A la différence des méthodes d'annotation d'entités nommées existantes qui catégorisent les entités 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 un module interrogeant des ressources du Web des données. Nous utilisons des ressources génériques telles que DBpedia [14] ou plus spécifiques telles que GeoNames [15], et OpenStreetMap [16]. 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, en particulier avec des ressources à couverture mondiale il est fréquent qu'un même nom soit associé à différents objets du monde. Ces objets pouvant différer par l'une ou plusieurs des métadonnées qui leur sont associées, par exemple leur appartenance à une catégorie différente d'ENE. Dans le cadre des méthodes d'évaluation de ces services ce type d'erreur est généralement appelé un faux positif.

Dans un premier temps, le module de résolution et de désambiguïsation des ENE applique un algorithme générique par défaut s'appuyant sur les méthodes proposées dans (Moncla *et al*., 2014). Après avoir interrogé les ressources géographiques, il supprime les doublons introduits par l'utilisation de plusieurs ressources puis si il n'y a pas de résultats et que l'ENE est de niveau 0 alors elle est considérée comme étant de type non-spatiale. Mais s'il s'agit d'une ENE d'un niveau supérieur à 0 alors il 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.

Par ailleurs, l'annotation de la structure d'imbrication des ENE nous permet de simplifier l'étape de classification grâce à la prise en compte de manière simplifiée du contexte associé aux EN. En effet, nous pouvons déterminer grâce à l'interrogation d'une base de connaissances sémantiques (Linked Data) la nature du terme qui compose une ENE de niveau 1 et ainsi en déduire sa catégorie. Par exemple, comme on peut le voir avec l'ENE (8) le terme « refuge » qui est de nature géographique d'après l'ontologie de GeoNames nous permet de catégoriser l'ENE (8) comme spatiale.

(8) refuge des Barmettes.

Le troisième service Web SW_3 que nous proposons retourne le résultat obtenu par le module de classification des ENE (figure 8b). Il s'agit comme pour le service

14. http://wiki.dbpedia.org/
15. www.geonames.org/export/web-services.html
16. http://nominatim.openstreetmap.org/search

Web précédent d'un résultat au format XML respectant les recommandations TEI et contenant cette fois-ci les annotations sémantiques des ENE spatiales.

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

(a) Avant classification

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

(b) Après classification

Figure 8. Résultat de l'annotation avant et après la classification des ENE

3.4 Post-traitements

Le résultat de l'étape d'annotation des ENE et des informations spatiales associées peut être utilisé et interprété pour différentes tâches dans différents domaines. Nous présentons dans cet article trois services de post-traitements utilisant le résultat produit par les modules présentés précédemment. Il s'agit d'un service de géocodage des ENE spatiales, d'un service de désambiguïsation de ces ENE spatiales et d'un service de reconstruction d'itinéraire.

3.4.1 Geocodage

En plus de l'étape de reconnaissance d'entité nommée, la résolution des toponymes (entité de lieu) se décompose en deux tâches : le géocodage et la désambiguïsation. L'objectif de l'étape de géocodage est de récupérer les coordonnées géographiques associées aux entités de lieu. Pour cela nous interrogeons des bases de données toponymiques disponibles sur le Web, telles que GeoNames et OpenStreetMap, et également des ressources institutionnelles. En effet, en fonction de la localisation territoriale des faits relatés dans le document à traiter nous pouvons être amené à interroger les ressources institutionnelles couvrant le même territoire. Ces ressources proposent, pour la plupart, un accès aux données géographiques grâce au protocole WFS (Web Feature Service) défini par l'OGC. 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 lorsqu'elles existent. Comme relaté précédemment, l'interrogation simultanée de plusieurs ressources améliore l'exhaustivité des données mais peut avoir comme inconvénient la multiplicité des réponses, comme par exemple avoir pour une entité de lieu plusieurs localisations différentes.

Le service Web SW_4 que nous proposons prend en entrée le résultat obtenu par le module de classification des ENE (figure 8b) et pour chaque ENE spatiale interroge les ressources géographiques. Ce service de géocodage produit un résultat au format GeoJson (figure 9a), GPX (figure 9b) ou KML (figure 9c).

```
{"type": "FeatureCollection",
 "features": [{
     "type": "Feature",
     "properties": {
      "name": "refuge des Barmettes" },
     "geometry": {
      "type": "Point",
      "coordinates":
      [6.7527105, 45.3895634]}
   }]
}
```

(a) GeoJson

```
<gpx [...]>
  <wpt lat="45.3895634"
       lon="6.7527105">
     <ele>2006.0</ele>
     <name>refuge des
         Barmettes</name>
  </wpt>
</gpx>
```

(b) GPX

```
<kml xmlns="http://www.opengis.net/kml/2.2">
  <Placemark>
    <name>le refuge des Barmettes</name>
    <description>Refuge des Barmettes, Pralognan-la-Vanoise, Savoie,
        Auvergne-Rhone-Alpes, 73710, France</description>
    <Point>
      <coordinates>6.7527105,45.3895634</coordinates>
    </Point>
  </Placemark>
</kml>
```

(c) KML

Figure 9. Exemple de résultats produits par le service de géocodage

3.4.2 Désambiguïsation

Le fait qu'une ENE de niveau 0 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.

Dans cet article nous présentons un service de désambiguïsation des ENE qui supprime les doublons introduits par l'utilisation de plusieurs ressources et étiquette l'ENE comme étant de type non-spatiale si elle est, ou a atteint, le niveau 0 et qu'aucun résultat n'a été retourné. Le service est implémenté à partir d'algorithmes adaptés à des textes décrivant des itinéraires et qui s'appuient sur les méthodes proposées dans (Moncla *et al.*, 2014). Il s'agit d'une méthode hybride combinant l'utilisation du sous-typage des ENE (Nguyen *et al.*, 2013) et une méthode de clustering par densité spatiale. Concernant l'utilisation du sous-typage des ENE, l'heuristique mise en œuvre est la suivante : lorsqu'il s'agit d'une ENE d'un niveau supérieur à 0 alors le service 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 thésaurus 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 spatiale/non-spatiale 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. La qualité du résultat dépend donc de la richesse des ressources disponibles mais le principe fonctionne à partir d'une seule ressource. Comme le montre la figure 2, ce service prend en entrée le fichier XML produit par l'étape de catégorisation des ENE (détaillé en section 3.3) ainsi que les résultats du service de géocodage. **Le service Web SW**$_5$ retourne le résultat (au format XML/TEI) obtenu par le module de désambiguïsation des ENE spatiales (figure 10). La valeur de l'attribut *ref* de l'élément `<placeName>` contient à présent l'URI correspondante à une ressource référencée sur le Web des données. Le module de désambiguïsation a pour tâche de sélectionner un résultat unique parmi la liste des résultats retournés par le service de géocodage. Pour cela, il utilise les informations spatiales stockées dans le fichier de géocodage (GeoJson, GPX ou KML) ainsi que le contexte issus du fichier XML/TEI produit par les services de reconnaissance et de catégorisation des ENE et des informations spatiales.

```
<placeName ref="https://www.openstreetmap.org/node/451703419">
  <geogName type="S" subtype="RHSE">
    <geogFeat>refuge</geogFeat>
    des
    <name>Barmettes</name>
  </geogName>
</placeName>
```

Figure 10. Résultat de l'annotation avant et après la classification des ENE

3.4.3 Reconstruction d'itinéraires

Comme nous l'avons vu précédemment, un des intérêts majeur de proposer une architecture modulaire implémentée sous forme de services Web est de permettre la réutilisation ou la combinaison de services pour construire une chaîne de traitement adaptée aux besoins spécifiques de chacun.

Le service Web SW$_6$ que nous proposons fait appel à la solution de reconstruction d'itinéraire développée dans (Moncla *et al.*, 2016). Cette solution implémente une approche multi-critère permettant de faire la distinction entre les points de passages et les points de repères visuels décrivant un itinéraire. Elle permet également de déterminer l'ordre dans lequel les points de passages sont atteints. Cette approche permet l'interprétation des informations annotées dans le texte telles que les expressions de déplacement et de localisation (grâce aux verbes de déplacement et de perception et aux relations spatiales). Elle utilise pour cela, d'une part les ENE et les informations spatiales annotées et d'autre part les données géographiques issues des modules de géocodage et de désambiguïsation.

4 Evaluation

Les différents services Web décrits dans cet article ont été regroupés au sein d'une chaîne de traitement nommée Perdido [17]. Cette chaîne de traitement permet l'expérimentation de notre solution avec un enchaînement automatique des différents traitements. Nous avons évalué notre proposition sur un corpus multi-lingue de descriptions de randonnées composé de 90 documents annotés manuellement, 30 documents pour chaque langue (français, espagnol et italien) contenant respectivement 11297, 5549 et 15724 mots. 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 trois 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 (Table 2) sur l'ensemble du corpus multi-lingue sont les suivants : rappel 98%, précision 93%, et SER 21%. Les résultats pour les trois langues traitées sont proches mais ces résultats sont légèrement plus faibles pour l'italien du fait de ressources moins exhaustives. Pour les documents français de notre corpus nous avons comparé les résultats de Perdido avec ceux obtenus

17. http://erig.univ-pau.fr/PERDIDO/

par l'outil d'annotation CasEN (version Quaero). CasEN obtient 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%. Le temps de traitement pour un document d'environ 400 mots est de l'ordre de la seconde pour CasEN et entre deux et trois secondes pour notre chaîne de traitement. La différence de temps de traitement vient du fait de l'interrogation de ressources externes. Ce temps est donc dépendant de la bande passante, du temps de réponse du service interrogé et du nombre d'EN (i.e., nombre de requêtes).

TABLE 2. *Évaluation de l'annotation des EN*

	Perdido			CasEN
	Français	Espagnol	Italien	Français
Rappel	95,9%	98,6%	98,1%	65,3%
Précision	94,9%	97,2%	86,1%	88,1%
Fscore	95,4%	97,9%	91,7	75%
SER	16,7%	15,2%	32,2%	51,1%

La grande différence dans les résultats entre Perdido et CasEN s'expliquent par l'utilisation des données liées par Perdido (par rapport à la seule utilisation des ressources lexicales proposées au sein d'Unitex). Ces résultats encourageants montrent, entre autre, l'intérêt d'avoir adopté une approche modulaire permettant de connecter différents services et différentes sources de données. Par ailleurs, l'utilisation de plusieurs ressources géographiques externes permet à la chaîne de traitement de fonctionné en mode « dégradé ». En effet, si une ressource est indisponible la modularité et la complémentarité de l'approche permet de la remplacer ou de continuer à utiliser les autres ressources disponibles.

5 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. En effet, nous avons proposé de modéliser le processus de traitement de données géospatiales sous la forme d'une chaîne de service modélisée par un Workflow. Cette tâche de décomposition et de mise à disposition de services pour le traitement d'informations géographiques issus de textes est un des objectifs du projet ANR CHOUCAS. La description sémantique de types de données ou de services basée sur l'utilisation de taxonomies (ex : ISO-1911-Information géographique — services) et de métadonnées (ex : ISO-19115- Information géographique — Métadonnées) permet non seulement d'améliorer la découverte de services mais également d'automatiser, au moins partiellement, le processus de composition d'un *workflow*.

La méthodologie présentée dans cet article repose sur la combinaison d'une étape de pré-traitement utilisant des analyseurs morpho-syntaxiques externes, d'une cascade de transducteurs, d'une étape de classification utilisant des ressources du Web des

données et diverses étapes de post-traitements. Les services Web proposés sont spécialement conçus pour répondre à des 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 [18] et le code source des modules de pré-traitement et de classification est disponible sur un dépôt GitHub [19]. Les premiers résultats présentés dans cet article nous encouragent à proposer des évolutions des modules de classification et de désambiguïsation afin d'inclure l'implémentation de différentes approches de désambiguïsation nouvelles ou existantes, adaptées à divers 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.

Doddington G. R., Mitchell A., Przybocki M. A., Ramshaw L. A., Strassel S., Weischedel R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, vol. 2, p. 1.

Dupont M., Vuillaume J.-M., Victorri B., Enjalbert P., Mathet Y., Malandain N. (2002). Nouvelles perspectives en extraction d'information. *Revue des Sciences et Technologies de l'Information - Série TSI : Technique et Science Informatiques*, vol. 1, n° 21, p. 37-63. Consulté sur https://halshs.archives-ouvertes.fr/halshs-00009485

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.

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.

Grishman R. (1997). Information extraction: Techniques and challenges. In *International summer school on information extraction*, p. 10–27.

Grishman R., Sundheim B. (1995). Design of the muc-6 evaluation. In *Proceedings of the 6th conference on message understanding*, p. 1–11. Stroudsburg, PA, USA, Association for Computational Linguistics. Consulté sur https://doi.org/10.3115/1072399.1072401

18. http://erig.univ-pau.fr/PERDIDO/
19. https://github.com/ludal360/Perdido

Grouin C., Rosset S., Zweigenbaum P., Fort K., Galibert O., Quintard L. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th linguistic annotation workshop*, p. 92–100.

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*.

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.

Vicente M. R. (2005). La glose comme outil de désambiguïsation référentielle des noms propres purs. *Corela. Cognition, représentation, langage*, n° HS-2.