Cadre d'évaluation de systèmes de reconnaissance d'entités nommées spatiales

Damien Palacio [1], **Christian Sallaberry** [2], **Guillaume Cabanac** [3], **Gilles Hubert** [3]

1. *Department of Geography, University of Zurich, Zurich, Suisse*
   damien.palacio@geo.uzh.ch
2. *LIUPPA ÉA 3000, Université de Pau et des Pays de l'Adour, Pau, France*
   christian.sallaberry@univ-pau.fr
3. *IRIT UMR 5505 CNRS, Université Paul Sabatier, Toulouse, France*
   {guillaume.cabanac,gilles.hubert}@univ-tlse3.fr

RÉSUMÉ. *La reconnaissance d'entités nommées est une tâche de l'activité d'extraction d'information dans des corpus textuels. Des systèmes de reconnaissance d'entités nommées spatiales sont très largement utilisés, mais souvent sans en connaître les forces et faiblesses. C'est pourquoi nous proposons le cadre d'évaluation SNERBM (Spatial Name Entity Recognition BenchMark) comme référentiel commun et nous l'expérimentons sur six systèmes existants de reconnaissance d'entités nommées spatiales. Ce cadre a pour objectif l'évaluation et la comparaison des performances de tels systèmes. Il permet également d'envisager le choix d'un système, ou encore la combinaison de différents systèmes, particulièrement adaptés aux catégories d'entité nommées spatiales (ville, barrage, montagne, par exemple) majoritairement présentes dans un corpus donné.*

ABSTRACT. *Named entity recognition is a task of information extraction from textual corpora. Spatial named entity recognition systems are widely used in this respect, but no one actually knows about their pros and cons. This is why we propose the SNERBM evaluation framework as a benchmark (Spatial Name Entity Recognition BenchMark), which we experimented on six existing systems dedicated to spatial named entity recognition. This benchmark enables the evaluation and the comparison of performances of such systems. In addition, it informs the selection of a system, or a combination of systems, best appropriate to operate on a given textual corpus featuring specific categories of spatial named entities (e.g., cities, mountains).*

MOTS-CLÉS : *reconnaissance d'entité nommée, entité nommée spatiale, cadre d'évaluation de système*

KEYWORDS: *named entity recognition, spatial named entity, system evaluation benchmark*

1. Introduction

La reconnaissance d'entités nommées (REN) dans des textes (Chinchor, 1998), consiste à identifier des syntagmes appelés entités nommées (c'est-à-dire des noms propres, des expressions de temps et des expressions numériques) catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, dates, quantités, distances, valeurs, acronymes, abréviations, etc. Parmi les entités nommées, les noms de lieux, que nous appellerons entités nommées spatiales (ENS), désignent des objets géographiques tels que des entités administratives (commune, par exemple), des éléments du relief, des éléments hydrographiques, etc.

Un système d'analyse d'ENS est généralement composé d'un module de reconnaissance (marquage d'ENS, tel que « Yosemite »), d'un module de classification (typage d'ENS, tel que « Yosemite Park »), d'un module de désambiguïsation (ENS vs. non ENS, tel que « Canari » vs. « canari » ou ENS vs. ENS différente, tel que « Paris » aux U.S. vs. « Paris » en France) et d'un module de géolocalisation (géocodage d'ENS, tel qu'avec les coordonnées du centre de « Yosemite Park » – latitude : 37,75, longitude : $-119{,}50$ ou de la géométrie correspondante). Les premiers modules contribuent à une construction par étape d'une représentation symbolique d'une ENS et le dernier détermine une représentation numérique d'une ENS (coordonnées géolocalisées).

Les systèmes de traitement automatique de la langue (TAL) supportent la reconnaissance et la classification d'entités nommées. Les systèmes d'analyse d'ENS supportent eux la reconnaissance, la classification, la désambiguïsation et la géolocalisation d'ENS (Marrero *et al.*, 2009 ; 2013). Nous pouvons citer Clavin [1] (D'Ignazio, 2013), GeoDict [2], Geolocator [3] (Gelernter, Zhang, 2013), OpenCalais [4] (D'Ignazio, 2013), Unlock [5] (Grover *et al.*, 2010) et Yahoo!PlaceSpotter [6] (Tobin *et al.*, 2010 ; Anastácio *et al.*, 2010 ; D'Ignazio, 2013) comme autant d'exemples de systèmes d'analyse allant de la reconnaissance jusqu'à la géolocalisation d'ENS. Par abus de langage, on parle souvent de systèmes de reconnaissance d'ENS (RENS). L'efficacité de tels systèmes peut être évaluée au travers de campagnes d'évaluation telles que MUC, CoNLL ou ACE (Marrero *et al.*, 2009). Ces campagnes mesurent l'efficacité des systèmes pour la reconnaissance d'entités nommées de type lieu (ENS), nom de personne, nom d'organisation... dans des corpus de documents associés. Peu de systèmes participant à ces compétitions, l'évaluation de systèmes d'analyse d'ENS commerciaux ou libres a été peu étudiée (Marrero *et al.*, 2009).

Les systèmes de recherche d'information géographique (RIG) intègrent des systèmes d'analyse d'entités nommées spatiales (ENS) (Leidner, 2007 ; Andogah, 2010 ;

1. http://clavin.bericotechnologies.com
2. http://www.datasciencetoolkit.org/developerdocs#commandline
3. https://github.com/geoparser/geolocator
4. http://www.opencalais.com
5. http://edina.ac.uk/unlock
6. http://developer.yahoo.com/boss/geo/docs

Sallaberry, 2013). Au delà de la seule RENS, les RIG indexent les ENS et proposent des langages d'interrogation supportant l'appariement des critères spatiaux exprimés dans un besoin d'information avec les données spatiales indexées. Ainsi, le système de RIG SPIRIT (Vaid *et al*., 2005), par exemple, comprend des processus dédiés à la reconnaissance, à la désambiguïsation et à la géolocalisation d'ENS.

Nous proposons le cadre d'évaluation de systèmes de reconnaissance d'entités nommées spatiales, dénommé SNERBM (*Spatial Name Entity Recognition BenchMark*). Ce cadre est ouvert à l'évaluation de tout nouveau système et son corpus de test peut également être étendu par de nouvelles données. La proposition contribue aux travaux de la communauté à plusieurs titres : (*i*) ce *benchmark* est un outil qui permet d'évaluer et d'améliorer un système de RENS ; (*ii*) il permet également à un utilisateur non spécialiste de comparer des systèmes de RENS et de choisir le plus performant pour une catégorie d'entité ; (*iii*) de plus, il propose une évaluation par catégorie d'ENS qui permet d'envisager des combinaisons avantageuses de différents systèmes ; (*iv*) enfin, il est ouvert et, par conséquent, invite la communauté à se l'approprier et à l'enrichir.

Cet article est organisé comme suit. Dans la section 2, nous décrivons la problématique relevant des campagnes d'évaluation de systèmes de recherche d'information et de systèmes de REN. Nous présentons notre cadre d'évaluation nommé « benchmark SNERBM » en section 3. Ensuite, en section 4, nous expérimentons le *benchmark* en comparant six systèmes de RENS en termes de qualité de réponse et de rapidité. Enfin, nous concluons par une discussion et des perspectives.

2. L'évaluation de systèmes de reconnaissance d'entités nommées (REN)

De nombreuses campagnes d'évaluation proposent des plateformes pour évaluer des systèmes de REN ou de RIG (Leidner, 2007 ; Andogah, 2010 ; Nouvel, 2012), comme indiqué dans le tableau 1 :

– MUC : Message Understanding Conference (Chinchor, 1998),

– MET : Multilingual Extraction Task (Chinchor, 1998),

– IREX : Information Retrieval and Extraction eXercise (Sekine, Eriguchi, 2000),

– CoNLL : Computational Natural Language Learning conference (Tjong Kim Sang, De Meulder, 2003),

– HAREM : Avaliaçao de sistemas de Reconhecimento de Entidades Mencionadas (Santos *et al*., 2006),

– GeoClef : Geographic Cross Language Evaluation Forum (Mandl *et al*., 2009),

– ACE : Automatic Content Extraction program (Strassel *et al*., 2008),

– EVALITA : Evaluation of NLP and Speech Tools for Italian (Lenzi *et al*., 2013),

– TREC-CS : Text Retrieval Conference – Contextual Suggestion (Voorhees, 2001 ; Dean-Hall *et al*., 2013).

Ainsi, différents types de *benchmarks* (référentiels) ont été expérimentés (tableau 2). Les ressources disponibles comprenant des ENS pré-marquées sont composées de

Tableau 1. Campagnes d'évaluation de systèmes de REN

Campagne	Année	Corpus	Système	Cible	Représentation	
					symbolique	numérique
MUC-7	1997	Reportages	NER	Personne, lieux, organisation, temps, mesure	X	
MUC-7/MET-2	1998	Journaux	NER	Personne, lieux, organisation, temps	X	
IREX	1999	Journaux	NER	Personne, lieux, organisation, temps	X	
CoNLL	2003	Journaux	NER	Personne, lieux, organisation, divers	X	
HAREM	2006	Journaux	NER	Personne, lieux, organisation, temps, mesure	X	
GeoCLEF	2008	Journaux	GIR	Lieux, thème	X	X
ACE	2008	Journaux	NER	Personne, lieux, organisation, géopolitique	X	
EVALITA	2011	Journaux	NER	Personne, lieux, organisation, géopolitique	X	
TREC-CS	2013	Open web, ClueWeb12	GIR	Lieux, thème	X	X

petits échantillons de documents (par ex., la campagne CoNLL propose un jeu de 231 documents en anglais avec 1 668 ENS annotées manuellement et la campagne ACE propose 400 documents en anglais avec aucune ENS annotée). Par ailleurs, les ressources dédiées à l'évaluation de systèmes de RIG proposent des jeux de données regroupant sans distinction les documents à la fois thématiquement et spatialement pertinents (par ex., la campagne GeoCLEF propose un jeu de documents associé à 100 *topics* géographiques (besoins d'information comportant des critères spatiaux et thématiques) et aux jugements de pertinence correspondants (*Qrels*), mais aucune ENS pré-annotée n'est proposée).

D'autres approches, généralement dédiées à l'évaluation d'un système particulier, utilisent des corpus plus petits dont les ressources annotées ne sont que rarement mises à disposition (tableau 3). Bucher *et al*. (2005) ont ainsi proposé d'adapter des techniques d'évaluation existantes pour évaluer le système de RIG SPIRIT. Marrero *et al*. (2009) présentent une plateforme qui a permis l'évaluation des systèmes de REN Supersense, Afner, Annie, Freeling, TextPro, YooName, ClearForest et Lingpipe. Tobin *et al*. (2010) décrivent une approche dédiée à l'évaluation de système de RENS et plus particulièrement des modules de désambiguïsation et de géolocalisation des systèmes Unlock et Yahoo!PlaceMaker. Anastácio *et al*. (2010) ciblent également l'évaluation de systèmes de RENS. Les auteurs comparent les méthodes de calcul de portée spatiale supportées respectivement par les systèmes Yahoo!PlaceMaker, GIPSY, Web-a-Where et GraphRank. D'Ignazio (2013) compare les services d'analyse spatiale des systèmes de RENS OpenCalais, Clavin et Yahoo!PlaceSpotter. Enfin, Gelernter et Zhang (2013) évaluent le système de RENS Geolocator. Les auteurs mesurent la qualité du module de reconnaissance de toponymes de Geolocator ainsi que de son module d'analyse et de géolocalisation.

Notre objectif est d'évaluer différents systèmes de RENS suivant un même référentiel. Les *benchmarks* cités précédemment sont rarement mis à disposition. Or, nous avons pu nous procurer et travailler sur les *benchmarks* TREC-CS (Dean-Hall *et al*.,

Tableau 2. Compléments relatifs aux campagnes d'évaluation de systèmes de REN

Campagne	Année	Documents	Langues	Entités nommées	Bibliographie	Ressources annotées
MUC-7	1997	158 000	anglais	-	[Chinchor'98]	jeu de test
MUC-7/MET-2	1998	500	chinois, japonnais	-	[Chinchor'98b]	jeu de test
IREX	1999	1 371	japonnais	-	[Sekine'00]	jeu de test
CoNLL	2003	1 499	anglais, allemand	11 503	[Tjong'03]	jeu de test
HAREM	2006	1 202	portugais	5 132	[Santos'06]	jeu de test
GeoCLEF	2008	200 000	portugais, allemand, anglais	-	[Mandl'08]	jeu : thématique, géographique
ACE	2008	10 000	anglais, arabe	-	[Strassel'08]	jeu de test
EVALITA	2011	42 595	italien	1 924	[Bartalesietal'11]	jeu de test
TREC-CS	2013	30 144	anglais	-	[Dean-Hall'13]	jeu : géographique

Tableau 3. Campagnes d'évaluation ad hoc dédiées aux systèmes de REN spécifiques

Bibliographie	Documents	Langues	Entités nommées	Systèmes évalués	Ressources annotées
[Bucher'05]	21 094	anglais	-	RIG : SPIRIT	-
[Marrero'09]	1	anglais	100 EN	REN : Supersense, Afner, Annie, Freeling, TextPro, YooName, ClearForest, Lingpipe	-
[Tobin'10]	1 032	anglais	13 077 ENS	RENS : Unlock, Yahoo PlaceMaker	-
[Anastacio'10]	6 000	anglais	1 100 ENS	RENS : Yahoo PlaceMaker, GIPSY, Web-a-Where, GraphRank	-
[Ignazio'13]	75	-	-	RENS : OpenCalais, Clavin, Yahoo PlaceSpotter	-
[Gelernter'13]	1 306	espagnol, anglais	799 ENS	RENS : Geolocator	-

2013) et GeoparsingQT (Gelernter, Zhang, 2013). Le premier ne propose que des ENS de type nom de grandes villes d'Amérique du nord. Le second est plus intéressant car il traite de quinze catégories différentes d'ENS.

Aussi, nous proposons de construire et d'expérimenter un *benchmark* ouvert, basé sur GeoparsingQT. En effet, GeoparsingQT comporte un jeu d'ENS constitué par des géographes et utilisé par l'équipe de développeurs du système Geolocator (Gelernter, Zhang, 2013) pour mesurer les éventuels effets de bord engendrés par chaque passage à une version supérieure du système de RENS. À l'image de *PABench*, pour *Points Of Interest (POI) Alignment Benchmark* (Morana *et al*., 2014 ; Berjawi *et al*., 2015), dédié à l'évaluation de systèmes d'appariement de POI issus de différents services de description et de géolocalisation (Geonames, Google Maps, Bing Maps, par exemple), nous proposons un référentiel ouvert et évolutif, validé par des utilisateurs.

Nous prenons comme cas d'application un ensemble de systèmes de RENS existants : Clavin, Geodict, Geolocator, OpenCalais, Unlock et Yahoo!PlaceSpotter. À ce jour, ces systèmes n'ont pas été évalués ni confrontés au sein d'un même *benchmark*.

3. Le cadre d'évaluation SNERBM

Plus d'une dizaine de systèmes de reconnaissance d'ENS ont été proposés dans la littérature lors des vingt dernières années (Lieberman *et al*., 2010 ; Lingad *et al*., 2013). Certains ont été évalués sur des critères de qualité de réponse, de temps de réponse, ou

les deux (tableau 4). Cependant, les cadres d'évaluation mis en œuvre reposent sur des corpus et des métriques différents. Comme illustré au tableau 3, peu d'études ont visé la comparaison de systèmes et aucune n'a mis des ressources annotées à disposition. Par conséquent, il est impossible à ce jour de connaître les performances relatives des systèmes suivant un même référentiel.

Tableau 4. Compléments relatifs aux campagnes d'évaluation ad hoc dédiées aux systèmes de REN

		Evaluation	
Système	**Cible**	**Effectiveness (qualité)**	**Efficiency (temps)**
Afner	REN	[Marrero'09]	
Annie	REN	[Marrero'09]	
Clavin	RENS	[Ignazio'13]	
ClearForest	REN	[Marrero'09]	
Freeling	REN	[Marrero'09]	
Geolocator	RENS	[Gelernter'13]	
GIPSY	RENS	[Anastacio'10]	
GraphRank	RENS	[Anastacio'10]	
LingPipe	REN	[Marrero'09]	
OpenCalais	RENS	[Ignazio'13]	
SPIRIT	RIG	[Bucher'05], [Vaid'05]	[Vaid'05]
Supersense	REN	[Marrero'09]	
TextPro	REN	[Marrero'09]	
Unlock	RENS	[Tobin'10]	
Web-a-Where	RENS	[Anastacio'10]	
Yahoo!PlaceMaker	RENS	[Tobin'10], [Anastacio'10]	
Yahoo!PlaceSpotter	RENS	[Ignazio'13]	
YooName	REN	[Marrero'09]	

Ce type de problème a trouvé des réponses en recherche d'information (RI). La RI s'appuie sur une longue tradition d'évaluation, notamment, via des campagnes d'évaluation de systèmes de recherche d'information (SRI). Ces campagnes définissent et implémentent des *benchmarks* pour évaluer et confronter les performances de SRI (Voorhees, 2002 ; 2007). Par exemple, la campagne TREC implémente le *benchmark* TREC-CS visant à évaluer la qualité (*effectiveness*) des recommandations de lieux, sans toutefois évaluer le temps de réponse (*efficiency*) des systèmes.

En ce qui concerne les systèmes de RENS, aucun *benchmark* existant ne s'est imposé comme référentiel commun. Le principal frein à l'adoption d'un *benchmark* est certainement le verrouillage (la non diffusion) des corpus et du code du logiciel d'évaluation. Dans cet article, nous proposons un *benchmark* d'évaluation des systèmes de RENS, basé sur un corpus ouvert. Nous avons appelé ce cadre d'évaluation SNERBM. Les principaux apports de cette proposition sont :

– *la couverture.* La qualité des systèmes est doublement évaluée : l'*effectiveness* mesure la qualité de la RENS tandis que l'*efficiency* mesure la réactivité des systèmes. Ces deux indications permettent d'identifier le système offrant le meilleur rapport *effectiveness–efficiency*.

– *la neutralité.* Le *benchmark* proposé est ouvert et peut être étendu par des jeux de tests complémentaires afin de prendre en compte des cas de figure non étudiés jusqu'alors. Il s'agit de traiter le maximum de cas possibles et de ne favoriser aucun système au regard du corpus employé ou des métriques d'évaluation.

– *la réutilisabilité.* Nous comparons dans cet article N systèmes. Le *benchmark* permet de reproduire nos résultats et de positionner une variante de système ou même un nouveau système par rapport à ces N performances de référence (*baselines*).

Le *benchmark* SNERBM s'appuie sur la collection de test GeoparsingQT (Gelernter, Zhang, 2013) et comprend :

– des catégories d'ENS : 15 catégories de type *Nom unique aux États-Unis*, *Nom et contexte associé*, *Abréviation administrative*, *Nom sans typologie associée*, *Nom et niveau administratif associé*, *Nom avec diacritique*, *Séquence de noms avec caractéristiques communes*, *Nom avec typologie associée*, *Nom officiel, court ou dérivé*, *Nom retors avec caractères spéciaux*, *Abréviation*, *Nom retors avec caractères numériques*, *Surnom*, *Nom historique*, *Autre type de référence* (tableau 5).

– 244 phrases : selon le modèle *sentence case*. Chaque phrase est associée à une seule catégorie (tableau 5).

– des jugements de pertinence (appelés *qrels*, pour *query relevance judgments* dans le vocabulaire TREC) : pour chaque phrase, les ENS sont géolocalisées par un ponctuel et/ou une géométrie déterminés par des experts (figures 1 et 2), respectivement à l'aide des ressources Geonames [7] et Google Map API V3 Tool [8].

Par exemple, pour la phrase « Rhodesia and Northern Rhodesia were renamed Zimbabwe and Zambia » de la catégorie « Historical Places / Nom historique », le fichier des Qrels contient 2 polygones (figure 1, éditée sous *kml.appspot* [9]). De même, pour la phrase « He traveled to a cape named Big Island in North Carolina » de la catégorie « Names Which Are The Same for Different Feature Types / Nom avec typologie associée », le fichier des Qrels contient 1 polygone (figure 2).

Ainsi, SNERBM permet de mesurer la performance d'un système de RENS suivant deux volets : la qualité des résultats (*effectiveness*) et la rapidité du système (*efficiency*).

3.1. Critères d'efficacité (effectiveness)

Pour mesurer la qualité des résultats fournis par un système de RENS, nous lui soumettons chaque phrase du jeu de test et évaluons les ENS détectées selon une démarche TREC classique. Nous capitalisons sur ce cadre de référence pour calculer des indicateurs de précision, de rappel et de F-mesure (Manning *et al.*, 2008, chap. 8). Ainsi, comme le montre la figure 3, une phrase est soumise au système qui restitue

7. http://www.geonames.org
8. http://www.birdtheme.org/useful/v3tool.html
9. http://display-kml.appspot.com/

Figure 1. Deux géométries sont associées à la phrase « Rhodesia and Northern Rhodesia were renamed Zimbabwe and Zambia » dans les Qrels

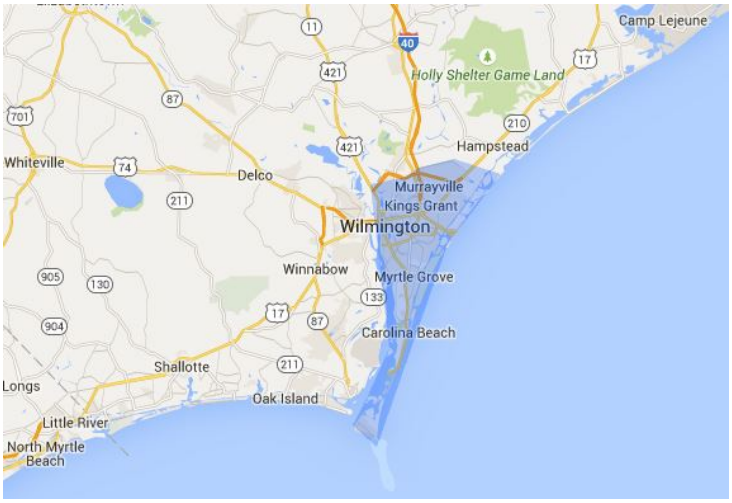

Figure 2. Une géométrie est associée à la phrase « He traveled to a cape named Big Island in North Carolina » dans les Qrels

alors aucune, une ou plusieurs ENS détectées et géolocalisées (points ou polygones correspondants). Une seconde phase de traitement, nommée « Intersections » sur la figure 3, compare ce résultat d'annotation avec les géométries définies dans les Qrels. Cette phase permet, pour une phrase donnée, de construire un résultat avec les caractéristiques suivantes :

– s'il existe une intersection entre une géométrie du résultat et une géométrie des Qrels alors : création d'un n-uplet (numéro phrase i, ..., numéro document d correspondant dans Qrels, ..., nom du système de RENS), par exemple, [phrase_31, ..., doc_34, système_Clavin] qui signifie que ce ponctuel (ou polygone) renvoyé par le système Clavin est en intersection avec un polygone associé dans les Qrels;

– s'il n'existe pas d'intersection entre une géométrie du résultat et une géométrie des Qrels alors : création d'un n-uplet (numéro phrase i, ..., numéro document d fictif (numéro phrase $i \times 1000$) inexistant dans les Qrels, ..., nom du système de RENS), par exemple, [phrase_31, ..., doc_31000, système_Clavin] qui signifie que ce ponctuel

Tableau 5. Exemples associés aux catégories du benchmark

Catégorie	Phrase	Commentaire
Abréviation	He climbed Mt. McKinley in Alaska.	Mount McKinley, Alaska
Abréviation administrative	He went to San Francisco, CA.	California
Autre type de référence	There was an accident at -120.9762, 41.25.	Longitude/Latitude Adin, CA
Divers	The French value their freedom.	People of France
Nom avec diacritique	Biên Hòa is a city in Vietnam.	City of Bien Hoa, Vietnam
Nom avec typologie associée	He went to the town of Big Island.	Town named Big Island
Nom et contexte associé	She was born in Paris in Idaho.	City of Paris, Idaho
Nom et niveau administratif associé	He went to Georgia, Kansas.	City of Georgia, KS
Nom et niveau administratif associé	He went to Tblisi, Georgia.	Country of Georgia
Nom historique	Ceylon became Sri Lanka.	Ceylon was renamed Sri Lanki
Nom officiel, court ou dérivé	The Republic of Korea is on a peninsula.	Country of South Korea
Nom retor avec caractères numériques	Green Creek Dam Nr. 5 was in need of repair.	Dam in Erath, Texas
Nom retor avec caractères spéciaux	They were vacationing in Trinidad and Tobago.	Country Trinidad and Tobago
Nom sans typologie associée	He climbed Rainier.	Mount Rainier
Nom unique aux États-Unis	She vacationed on the shores of Metacomet Lake.	Lake
Noms avec caractéristiques communes	The bicylce race went through Paris, Clarksville, and Hugo.	Cities in Texas
Surnom	There is a baseball team in the Big Apple.	New York

(ou polygone) renvoyé par le système Clavin n'est en intersection avec aucun des polygones associés à la phrase 31 dans les Qrels.

Enfin, ces n-uplets et les Qrels sont donnés dans le format idoine en entrées du programme `trec_eval` [10] de TREC qui calcule un ensemble de mesures d'évaluation (figure 3). Il en résulte les mesures de Précision, de Rappel et de F-mesure déterminées par TREC. Ces résultats nous permettent ensuite de construire des tableaux synthétiques de présentation des mesures par catégorie et par système de RENS.

Nous présentons ces mesures d'*effectiveness* de façon synthétique : toutes catégories confondues, d'une part, et pour chacune des quinze catégories, d'autre part.

3.2. Critères de performance (efficiency)

Nous proposons de mesurer le temps de traitement de l'ensemble des phrases toutes catégories confondues. Il est également intéressant de travailler sur le temps moyen de traitement d'une phrase, tout comme sur le temps moyen de traitement des phrases d'une catégorie donnée. Cet indicateur renseigne les utilisateurs de systèmes de RENS pour trouver l'équilibre adéquat entre qualité des résultats et temps de réponse acceptables pour les utilisateurs finals.

3.3. Mise à disposition du benchmark

Nous proposons la démarche d'utilisation totalement ouverte suivante :

– Le participant télécharge la liste de phrases associée au *benchmark* SNERBM ;

10. http://trec.nist.gov/trec_eval

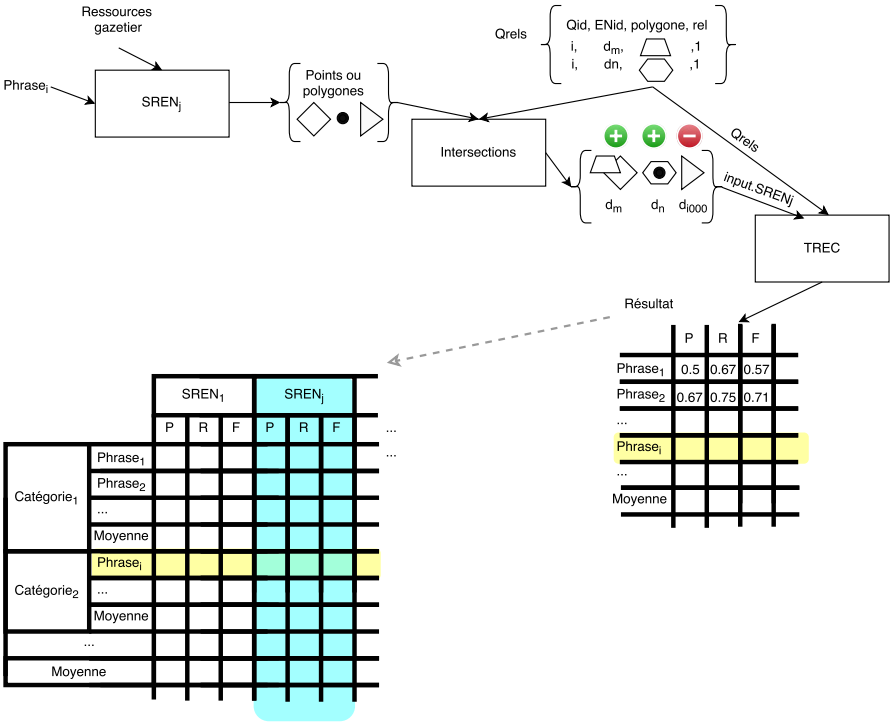

Figure 3. Processus d'évaluation des systèmes de RENS selon le benchmark SNERBM

– Pour chaque phrase, le participant calcule et produit, à l'aide de son système, des triplets (numéro de phrase, ENS, coordonnées géocodées) décrivant les ENS retrouvées et nous les communique afin d'obtenir un tableau synthétique des résultats ;

– À l'issue de chaque utilisation du *benchmark*, le participant est invité à proposer des catégories ou phrases supplémentaires qu'un groupe de modérateurs de SNERBM validera afin d'enrichir le jeu de données ou d'en créer de nouveaux.

Notons que, s'il le souhaite, le participant pourra télécharger les *qrels* et calculer ses résultats directement avec le programme `trec_eval`.

4. Évaluation de systèmes RENS avec SNERBM

Pour valider le cadre d'évaluation SNERBM et son utilisation comme référentiel commun, nous l'avons mis à l'épreuve en évaluant plusieurs systèmes existants : Clavin, Geodict, Geolocator, OpenCalais, Unlock et Yahoo!PlaceSpotter.

4.1. Mesure de l'efficacité (effectiveness)

Les résultats de l'évaluation du point de vue de l'*effectiveness* sont présentés dans le tableau 6. Pour un système donné, sont représentés la *Précision*, le *Rappel* et la *F1-mesure*. Le dernier jeu de mesures correspond à la meilleure combinaison de systèmes qui, pour chaque phrase, retient le système ayant proposé le meilleur résultat. Cette analyse de la performance globale (c.-à-d., sur l'ensemble des phrases du *benchmark*), au regard de l'*effectiveness*, classe les systèmes Yahoo!PlaceSpotter, Opencalais et Geolocator aux trois premières places respectivement.

Tableau 6. Analyse de la précision (P), du rappel (R) et de la F1-mesure (F) pour chaque système évalué

	Clavin			Geodict			Geolocator			Opencalais			Unlock			**Yps**			Meilleure combinaison		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Mesure	0,29	0,31	0,29	0,19	0,18	0,19	0,42	0,48	0,44	0,42	0,49	0,44	0,33	0,39	0,35	**0,66**	**0,66**	**0,65**	0,86	0,85	0,87
Classement	4	5	5	5	6	6	2	3	2	2	2	2	3	4	4	1	1	1			

Tableau 7. Analyse de la F1-mesure par système et par catégorie d'ENS

F1-Mesure	Clavin	Geodict	Geolocator	Opencalais	Unlock	Yps	Meilleure combinaison
Catégories d'ENS							
Abréviation	0,24	0,14	0,37	0,57	0,57	**0,91**	0,94
Abréviation administrative	0,87	0,60	0,93	0,73	0,00	**1,00**	1,00
Autre type de référence	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Divers	0,21	0,19	0,33	0,25	0,26	**0,50**	0,87
Nom avec diacritique	0,33	0,00	0,58	0,24	0,24	**0,80**	1,00
Nom avec typologie associée	**0,40**	0,20	0,33	0,00	0,13	0,20	0,75
Nom et contexte associé	0,16	0,30	0,12	0,29	0,17	**0,65**	0,66
Nom et niveau administratif associé	0,19	0,53	0,42	0,44	0,37	**0,84**	0,95
Nom historique	0,55	0,22	0,69	0,67	**0,88**	0,69	0,96
Nom officiel, court ou dérivé	0,82	0,51	0,82	0,83	0,62	**0,86**	0,97
Nom retor avec caractères numériques	0,00	0,00	0,17	0,00	0,00	**0,33**	1,00
Nom retor avec caractères spéciaux	0,20	0,05	0,60	0,67	0,33	**0,71**	1,00
Nom sans typologie associée	0,40	0,40	**0,42**	0,00	0,00	0,33	1,00
Nom unique aux États-Unis	0,57	0,00	0,69	**1,00**	0,92	0,50	1,00
Séquence de noms avec caractéristiques communes	0,31	0,17	0,40	0,41	**0,61**	0,46	0,56
Surnom	0,50	0,00	0,00	0,00	0,00	**0,67**	1,00

L'analyse des résultats par catégorie montre des catégories d'ENS correctement analysées et d'autres générant un fort taux d'échec. Le tableau 7 et la figure 4 présentent les meilleurs résultats obtenus pour chaque catégorie. Les différentes dénominations administratives, les abréviations et les noms officiels, par exemple, sont des catégories analysées avec une F1-mesure supérieure à 80 %. Sous la barre des 50 %, les références autres (de type GPS, par exemple) et les homonymes désignant des lieux différents sont difficiles à détecter et à analyser, y compris en présence de la typologie associée (nom avec typologie associée). De même, les noms avec des caractères numériques ou sans typologie associée sont difficiles à reconnaître et à analyser.

Pour un corpus hétérogène en termes de catégories, des combinaisons de systèmes sont envisageables afin d'améliorer la qualité des résultats de RENS. En effet, il existe presque toujours une complémentarité des systèmes qui permet d'envisager des combinaisons avantageuses (figure 4). Pour chaque phrase d'une catégorie donnée, nous avons observé les résultats des différents systèmes. Ainsi, la combinaison de

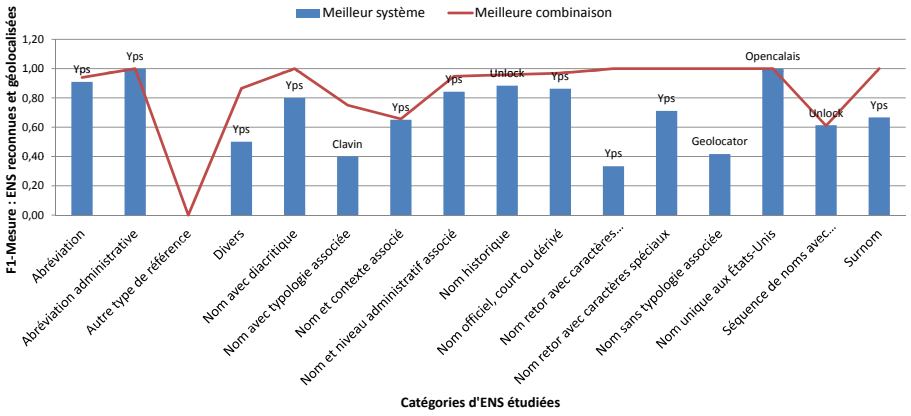

Figure 4. Meilleurs résultats par catégorie d'ENS

systèmes, pour une phrase, consiste à considérer le système donnant le meilleur résultat et, pour une catégorie, la moyenne des meilleurs résultats ainsi obtenus.

L'analyse qualitative des erreurs montre que tous les systèmes rencontrent des difficultés sur les mêmes phrases des catégories « Autre type de référence » et « Nom retors avec caractères numériques ». La catégorie « Nom retors avec caractères numériques » est singulière car chaque phrase comporte une énumération de noms de lieux. Une combinaison de systèmes permet dans ce cas d'améliorer considérablement la qualité des résultats : les systèmes sont bien complémentaires pour cette catégorie.

Pour un corpus donné, il sera donc intéressant d'analyser les résultats relatifs à différentes catégories (voir exemple figure 4) afin de préconiser un système de RENS particulier ou une combinaison de systèmes.

4.2. Mesure de la performance (efficiency)

Nous comparons les systèmes en termes de temps de traitement de la collection. Comme le montre la figure 5, d'importantes différences de temps de réponse sont observées : de 8 secondes à 51 minutes. Tous les traitements ont été effectués sur une machine Ubuntu 12.04 64 bits dotée d'un CPU simple cœur, de 4 Go de RAM et de 100 Go de disque dur.

Unlock est le système le plus lent. Ceci est dû au fait que le service web ne retourne pas directement les résultats mais propose un fonctionnement par lot (*batch*). Yahoo!Placespotter est le service web le plus rapide avec moins de 143 secondes pour l'ensemble de la collection. Clavin est plus rapide avec seulement 8 secondes de traitement. Contrairement à Yahoo!Placespotter, il est installé en local sur la machine qui réalise les traitements. Toutefois, Geolocator, qui est aussi installé sous la forme d'une application locale, nécessite 837 secondes pour la même collection.

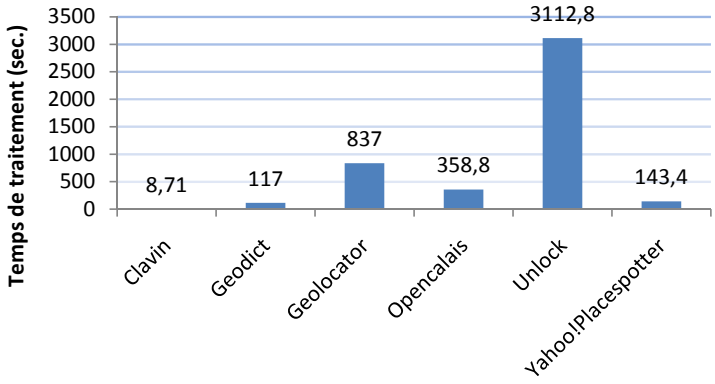

Figure 5. Temps de traitement de la collection

Le tableau 8 reprend des caractéristiques de ces systèmes. Nous avons choisi ces prototypes afin de comparer des systèmes avec des architectures et des stratégies de mises en œuvre différentes.

Tableau 8. Avantages et limites de chaque système

Système	Avantages	Limites
Clavin	Open source, local, très rapide, peu d'espace requis : RAM 1GB, DD 4GB	
Geodict	Open source (server image disponible) pas de limitation en nombre de requête	Service web limité pour trop de requêtes simultanées
Geolocator	Open source, local, espace requis : DD 7GB	>3GB de mémoire RAM, en cours de développement (bugs)
Opencalais		Clé limitée à 50K requêtes par jour (4 par seconde)
Unlock	Nombre de requêtes illimité	Traitements asynchrones
Yahoo!Placespotter	Rapide, peut traiter des URL directement	2K par jour et par IP ou clé limitée à 100K par jour

4.3. Synthèse

Cette synthèse reprend les mesures de temps de réponse et de F1-mesure sous la forme d'un diagramme composé de barres horizontales (figure 6). Nous avons défini la T-Mesure (équation 1), normalisée entre 0 et 1, en étendant la formule proposée dans (Lee, 1997). Ainsi, l'équation 1 calcule la performance (*efficiency*) T_i d'un système i relativement aux durées d'exécution $t_{1 \leqslant j \leqslant n}$ des n systèmes testés. Étant donné l'écart très important des temps de réponse observé pour les différents systèmes (figure 5), nous avons intégré un seuil pour mieux discriminer les performances desdits systèmes : $\forall i \in [1, n] \quad t_i \leftarrow \min(t_i, \text{seuil})$. Dans notre cas, ce seuil a été fixé à 500 secondes.

$$T_i = 1 - \frac{t_i - \min_{1 \leqslant j \leqslant n}(t_j)}{\max_{1 \leqslant j \leqslant n}(t_j) - \min_{1 \leqslant j \leqslant n}(t_j)} \in [0, 1] \qquad (1)$$

Les systèmes Clavin, Yps et Geodict maximisent les deux critères (figure 6) : le plus grand cumul correspond au meilleur résultat. Aussi, le système Clavin, qui offre

des traitements très rapides, malgré des résultats moyens en termes de F1-mesure, est-il au coude à coude avec le système Yps. Ces résultats globaux sont à considérer avec prudence toutefois. En effet, il est clair que mélanger *effectiveness* et *efficiency* présente des biais : l'un ne « rattrape » pas vraiment l'autre. Le critère *effectiveness* devrait sans doute être privilégié par une pondération plus importante.

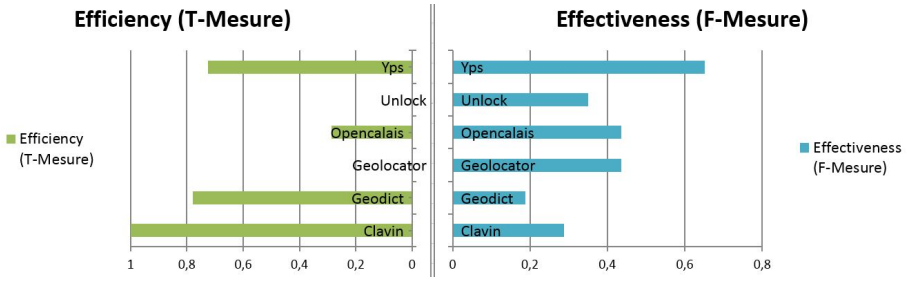

Figure 6. Synthèse des mesures de performance (efficiency) et d'efficacité (effectiveness) – plus le pourcentage est élevé, meilleure est la performance du système

5. Conclusion

Le cadre d'évaluation de système de RENS que nous avons appelé *benchmark* SNERBM est ouvert. Comme nous l'avons souligné, tout nouveau système peut être évalué selon le même protocole et comparé aux précédentes *baselines*. SNERBM est extensible, c'est-à-dire ouvert à la contribution : les contributeurs peuvent proposer de nouvelles phrases ou catégories d'ENS. La mise en ligne du *benchmark* SNERBM est en cours. De même, il nous paraît intéressant, à moyen terme, d'en proposer une version multilingue en conservant les mêmes *qrels* comme éléments de départ.

Enfin, ce travail présente un premier résultat de comparaison de systèmes de RENS disponibles en ligne. Au delà du comparatif global de systèmes, il pointe les catégories d'ENS correctement reconnues et analysées ainsi que celles encore mal gérées par les différents systèmes de RENS.

Remerciements. Nous remercions tout particulièrement Judith Gelernter de l'Université Carnegie Mellon et Doug Caldwell du U.S. Army Topographic Engineering Center (USATEC) pour avoir mis à notre disposition le jeu de test GeoparsingQT.

Bibliographie

Anastácio I., Martins B., Calado P. (2010). Using the geographic scopes of web documents for contextual advertising. In *GIR'10: Proceedings of the 6th Workshop on Geographic Information Retrieval*, p. 18:1–18:8. ACM.

Andogah G. (2010). *Geographically Constrained Information Retrieval*. Thèse de doctorat, University of Groningen, Netherlands.

Berjawi B., Duchateau F., Favetta F., Miquel M., Laurini R. (2015). PABench: Designing a Taxonomy and Implementing a Benchmark for Spatial Entity Matching. *GEOProcessing'2015: The 7th International Conference on Advanced Geographic Information Systems, Applications, and Services*, p. 7-16.

Bucher B., Clough P., Joho H., Purves R., Syed A. K. (2005). Geographic IR Systems: Requirements and Evaluation. In *ICC'05: Proceedings of the 22nd International Cartographic Conference.* Global Congressos. (CDROM)

Chinchor N. A. (1998). MUC/MET evaluation trends. In *Proceedings of the TIPSTER text program: Phase III*, p. 235–239. Association for Computational Linguistics.

Chinchor N. A. (1998). Overview of MUC-7. In *MUC-7: Proceedings of the 7th Message Understanding Conference.*

Dean-Hall A., Clarke C. L. A., Kamps J., Thomas P., Simone N., Voorhees E. (2013). Overview of the TREC 2013 Contextual Suggestion Track. In *TREC'13: Proceedings of the 22nd text retrieval conference.* NIST.

D'Ignazio C. (2013). *Big data, news and geography: Research update.* Consulté sur https://civic.mit.edu/blog/kanarinka/big-data-news-and-geography (MIT Center for Civic Media)

Gelernter J., Zhang W. (2013). Cross-lingual geo-parsing for non-structured data. In *GIR'13: Proceedings of the 7th Workshop on Geographic Information Retrieval*, p. 64–71. ACM.

Grover C., Tobin R., Byrne K., Woollard M., Reid J., Dunn S. *et al.* (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 368, n° 1925, p. 3875–3889.

Lee J. H. (1997). Analyses of Multiple Evidence Combination. In *SIGIR'97: Proceedings of the 20th annual international ACM SIGIR conference*, p. 267–276. ACM Press.

Leidner J. L. (2007). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names.* Thèse de doctorat, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh, Scotland.

Lenzi V. B., Speranza M., Sprugnoli R. (2013). Named entity recognition on transcribed broadcast news at EVALITA 2011. In *Revised Papers from EVALITA'11: International Workshop on the Evaluation of Natural Language and Speech Tools for Italian*, vol. 7689, p. 86-97. Springer.

Lieberman M. D., Samet H., Sankaranarayanan J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proceedings of the 26th International Conference on Data Engineering, ICDE*, p. 201–212. IEEE.

Lingad J., Karimi S., Yin J. (2013). Location extraction from disaster-related microblogs. In *Proceedings of the 22nd International Conference on World Wide Web*, p. 1017–1020. ACM.

Mandl T., Carvalho P., Nunzio G. M. D., Gey F. C., Larson R. R., Santos D. *et al.* (2009). GeoCLEF 2008: The CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. In *Revised Selected Papers of CLEF'08: Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum*, vol. 5706, p. 808–821.

Manning C. D., Raghavan P., Schütze H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Marrero M., Sánchez-Cuadrado S., Lara J. M., Andreadakis G. (2009). Evaluation of named entity extraction systems. *Advances in Computational Linguistics. Research in Computing Science*, vol. 41, p. 47–58.

Marrero M., Urbano J., Sánchez-Cuadrado S., Morato J., Berbís J. M. G. (2013). Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, vol. 35, nº 5, p. 482-489.

Morana A., Morel T., Berjawi B., Duchateau F. (2014). Geobench: a geospatial integration tool for building a spatial entity matching benchmark. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 533–536. ACM.

Nouvel D. (2012). *Reconnaissance des entités nommées par exploration de règles d'annotation*. Thèse de doctorat, Université François Rabelais de Tours, France.

Sallaberry C. (2013). *Geographical information retrieval in textual corpora*. Wiley-ISTE.

Santos D., Seco N., Cardoso N., Vilela R. (2006). HAREM: An Advanced NER Evaluation Contest for Portuguese. In *LREC'06: Proceedings of the 5th International Conference on Language Resources and Evaluation*, p. 1986–1991.

Sekine S., Eriguchi Y. (2000). Japanese Named Entity Extraction Evaluation – Analysis of Results. In *COLING'00: Proceedings of the 18th conference on Computational linguistics*, p. 1106-1110. Association for Computational Linguistics.

Strassel S., Przybocki M. A., Peterson K., Song Z., Maeda K. (2008). Linguistic Resources and Evaluation Techniques for Evaluation of Cross-Document Automatic Content Extraction. In *LREC'08: Proceedings of the International Conference on Language Resources and Evaluation*, p. 2706–2709.

Tjong Kim Sang E. F., De Meulder F. (2003). Introduction to the CoNLL-2003 Shared Task Language-Independent Named Entity Recognition. In *CoNLL-2003: Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003*, p. 142–147. Association for Computational Linguistics.

Tobin R., Grover C., Byrne K., Reid J., Walsh J. (2010). Evaluation of georeferencing. In *GIR'10: Proceedings of the 6th Workshop on Geographic Information Retrieval.* ACM.

Vaid S., Jones C. B., Joho H., Sanderson M. (2005). Spatio-textual Indexing for Geographical Search on the Web. In *SSTD'05: Proceedings of the 9th international Symposium on Spatial and Temporal Databases*, vol. 3633, p. 218–235. Springer.

Voorhees E. M. (2001). Overview of TREC 2001. In *TREC'01: Proceedings of the 9th Text REtrieval Conference.* NIST.

Voorhees E. M. (2002). The philosophy of information retrieval evaluation. In *CLEF'01: Proceedings of the Second Workshop of the Cross-Language Evaluation Forum*, vol. 2406, p. 355–370. Springer.

Voorhees E. M. (2007). TREC: Continuing information retrieval's tradition of experimentation. *Communications of the ACM*, vol. 50, nº 11, p. 51–54.