CONCEPTION D'UNE MÉTHODE HYBRIDE D'EXTRACTIONS D'INFORMATIONS GÉOGRAPHIQUES À PARTIR DE DONNÉES TEXTUELLES

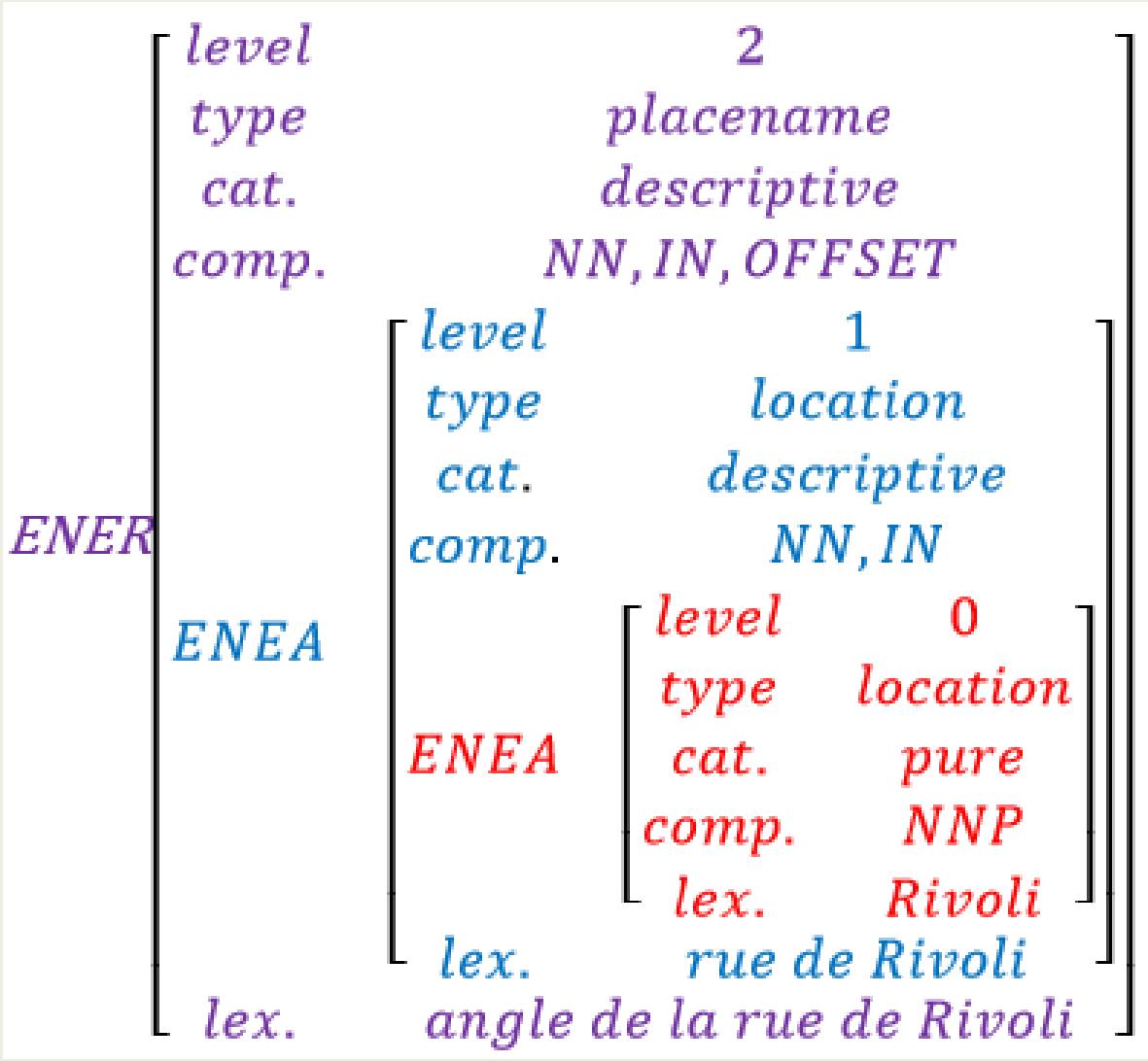
Appliqué à un corpus littéraire

I. Introduction
II. Géocodage
III. Désambiguïsation
IV. Détection des ENE par apprentissage supervisé
V. Conclusion

INTRODUCTION

Entité Nommée Etendue

- Entité Nommée :
  - *Niveau 0 : Rivoli*
- Entité Nommée Etendue Absolue :
  - *Niveau 1 : rue de Rivoli*
- Entité Nommée Etendue Relative :
  - *Niveau 2 : angle de la rue de Rivoli*



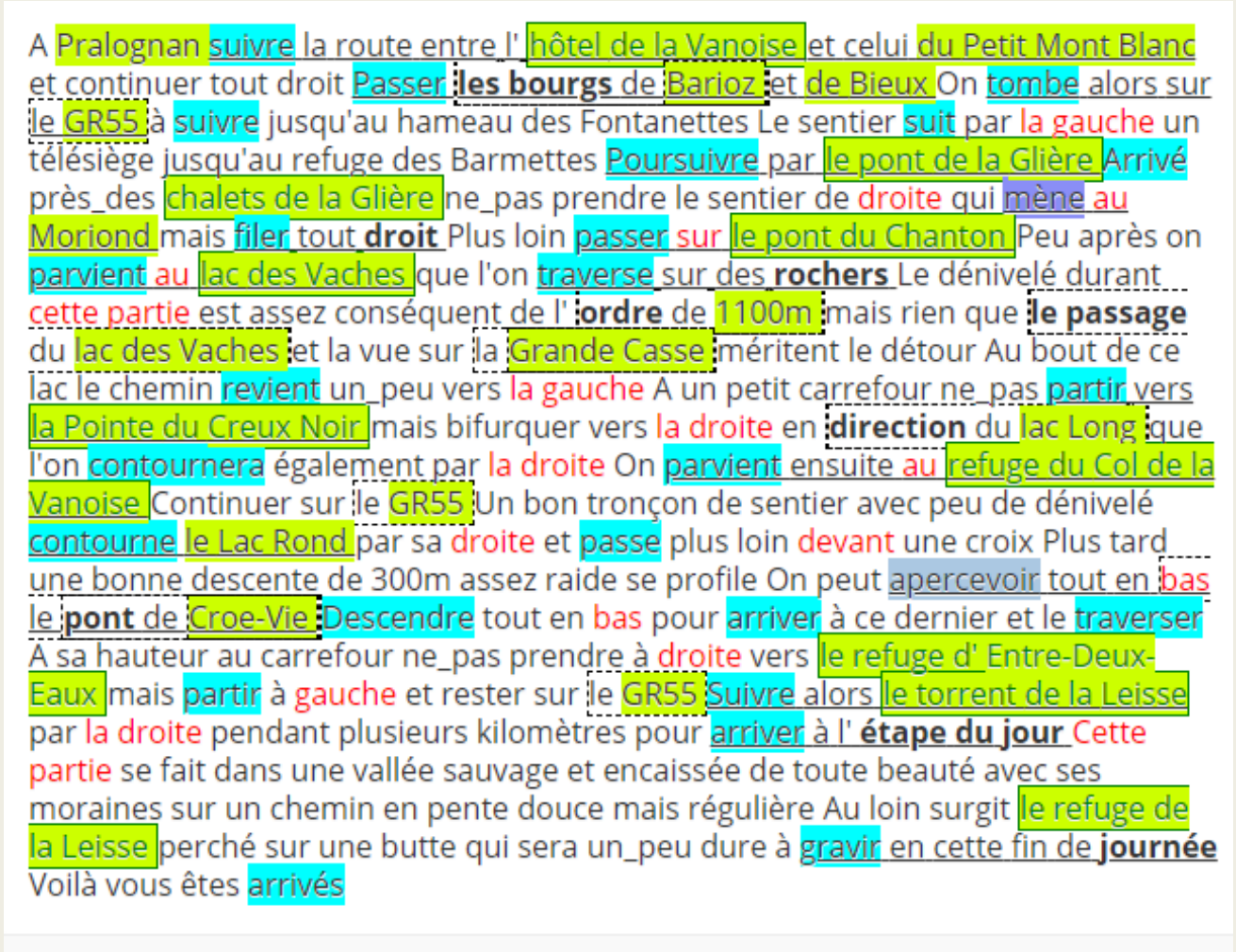
Geoparsing

- **Geotagging** : identification des références géographiques
- **Geocoding** : association d'une référence géographique à sa géolocalisation

```
<w lemma="le" type="DET" xml:id="w60">le</w>
<w lemma="long" type="A" xml:id="w61">long</w>
<w lemma="de" type="PREPDET" xml:id="w62">du</w>
<placeName n="1" xml:id="esne1">
  <geogName type="R" subtype="ST">
    <geogFeat>
      <w lemma="quai" type="N" xml:id="w63">quai</w>
    </geogFeat>
    <w lemma="de" type="PREP" xml:id="w64">de</w>
    <w lemma="la" type="DET" xml:id="w65">la</w>
    <rs type="unknown">
      <name type="unknown">
        <w type="NPr" lemma="" xml:id="w66">Grève</w>
      </name>
    </rs>
  </geogName>
</placeName> .
```

Repérage des ENE avec Perdido

- Chaine de traitement pour l'annotation d'entités nommés étendus et d 'informations géographiques
- Système basé sur un ensemble de règle



Informations extraites avec Perdido

- L’ESNE complète : devant le quai Saint Bernard
- Le nom court : quai Saint Bernard
- « street » ou «other » : street
- Le feature : quai
- L ’entité nommée : Saint Bernard
- La position dans le texte : 25eme mot

Corpus

Titre	Auteur	Date	Titre	Auteur	Date
La maison du Chat-qui-pelote	Balzac	1830	L'Assommoir	Zola	1877
Ferragus	Balzac	1833	Nana	Zola	1880
La fille aux yeux d'or	Balzac	1835	Une belle journée	Céard	1881
Le père Goriot	Balzac	1835	Pot-Bouille	Zola	1882
Grandeur et décadence de César Birotteau	Balzac	1837	Au Bonheur des dames	Zola	1883
Les mystères de Paris	Sue	1842	Le vingtième siècle	Robida	1883
Sans Cravate ou les Commissionnaires	Kock	1844	Sapho	Daudet	1884
L'envers de l'histoire contemporaine	Balzac	1848	Bel-ami	Maupassant	1885
M. Choublanc à la recherche de sa femme	Kock	1856	L'oeuvre	Zola	1876
Les misérables	Hugo	1862	La vie électrique	Robida	1892
Les demoiselles de magasin	Kock	1863	Paris	Zola	1897
Paris au XXème siècle	Verne	1863	La charpente	Rosny jeune	1900
L'éducation sentimentale	Flaubert	1869	Mr. Bergeret à Paris	France	1901
La Curée	Zola	1871	La Maternelle	Frapié	1904
Le ventre de Paris	Zola	1873	La Vague rouge	Rosny aîné	1910
Jack	Daudet	1876	Dans les rues	Rosny aîné	1913

GÉOCODAGE

Problématique

- Comment retrouver une position GPS à partir d'une entité nommée étendue ?
  - ➔Index géographiques

Index géographiques

Index	Zone	Type de lieux	Points forts	Utilisation
Géohistorical	Région parisienne	Rues	Rues historiques de Paris	Street & Others
IGN	France	Régions, villes, rues	Réponses complètes	Street
Nominatim	Monde	Lieux connus et rues	Précis, complet et noms alternatifs	Street & Others
Geonames	Monde	Lieux connus et rues	Requêtable en local et noms alternatifs	Others
Wikipédia	Monde	Lieux connus	Entrées Wikipédia	Street & Others

DÉSAMBIGUÏSATION

Problématiques

- Comment supprimer les résultats non pertinents renvoyés par les index?

  ➔Filtrage des réponses

- Comment choisir parmi plusieurs réponses celle qui est la plus pertinente?

  ➔ Heuristiques

Filtrage des réponses

- Supprimer le plus de réponses non pertinentes avant d’utiliser des heuristiques
- Supprimer toutes les réponses si l’entité n’est pas un lieu

Filtre distance orthographique

Réponse conservée

Rue Saint-Jacques

Réponse supprimée

Saint-Mandé

Filtre doublons

Boulevard Montmartre

Réponse conservée

Boulevard Montmartre

[48.87;2,34]

Réponse supprimée

Boulevard Montmartre

[48.87;2,34]

Filtre de feature

Boulevard Montmartre

Réponse conservée

Boulevard Montmartre

Réponse supprimée

Rue Montmartre

Filtre de ville

Boulevard de la Chapelle (Paris)

Réponse conservée

Boulevard de la Chapelle

Paris

Réponse supprimée

Boulevard de la Chapelle

Bordeaux

Filtre de nom propre

Réponse conservée

Cimetière de Montmartre

Réponse supprimée

Cimetière de Montparnasse

Filtre inter-index

Rue de la Michodière

Réponse 1 (GeoHistorical)

Rue de la Michodière

[48.86;2.33]

Paris

Type : rue

Réponse 2 (IGN)

Rue de la Michodière

[48.86;2.33]

Gaillon, 2e, Paris, Île-de-France, France

Type : unknown

Filtre inter-index

 Rue de la Michodière

Réponse finale

Rue de la Michodière

[48.86;2.33]

Gaillon, 2e, Paris, Île-de-France, France

Type : rue

Filtre correspondance parfaite

Rue d’Alger

Réponse conservée

Rue d’Alger

Similarité : 1,0

Réponse supprimée

Rue Auger

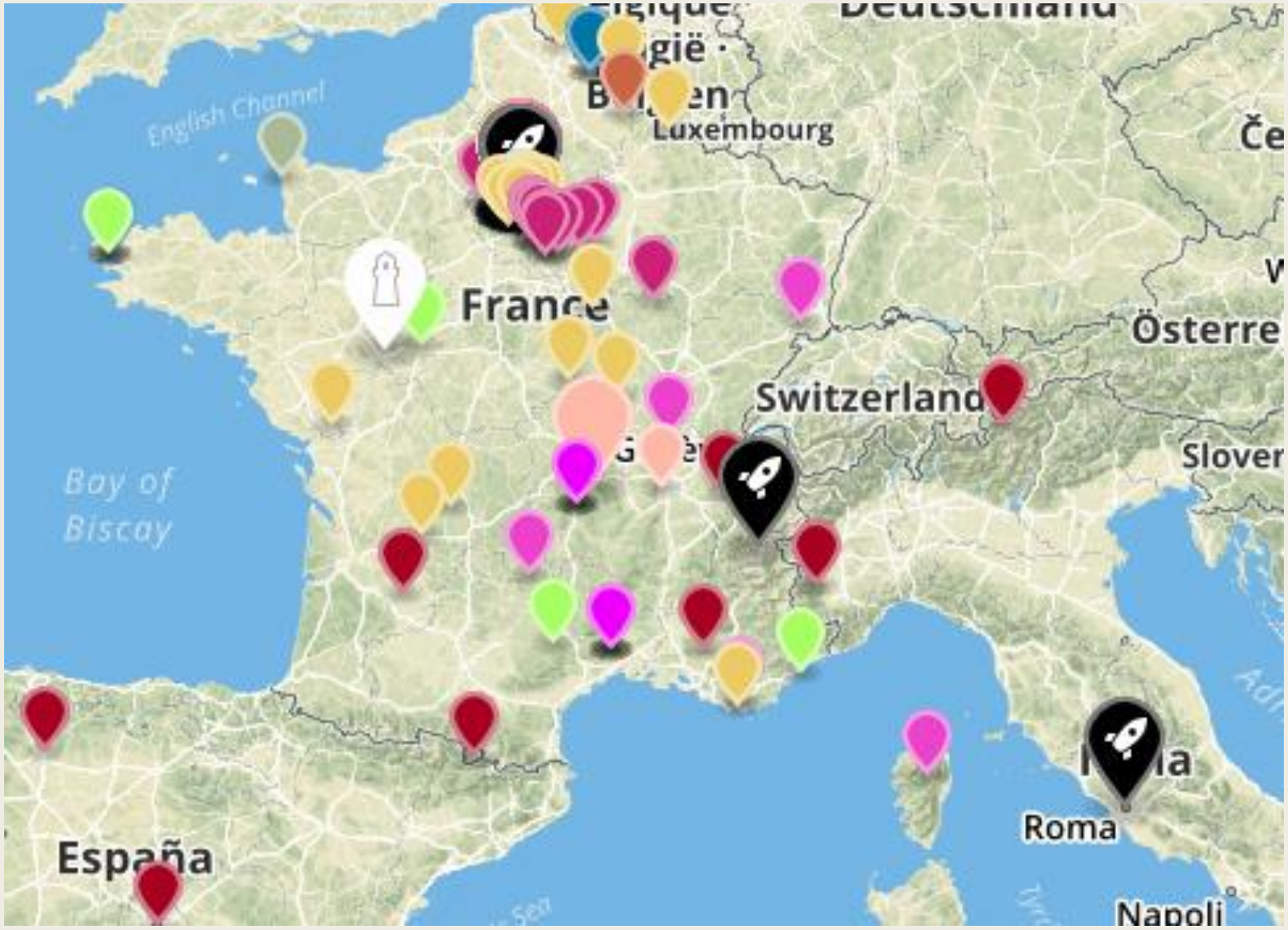
Similarité : 0,7

Evaluation filtrage

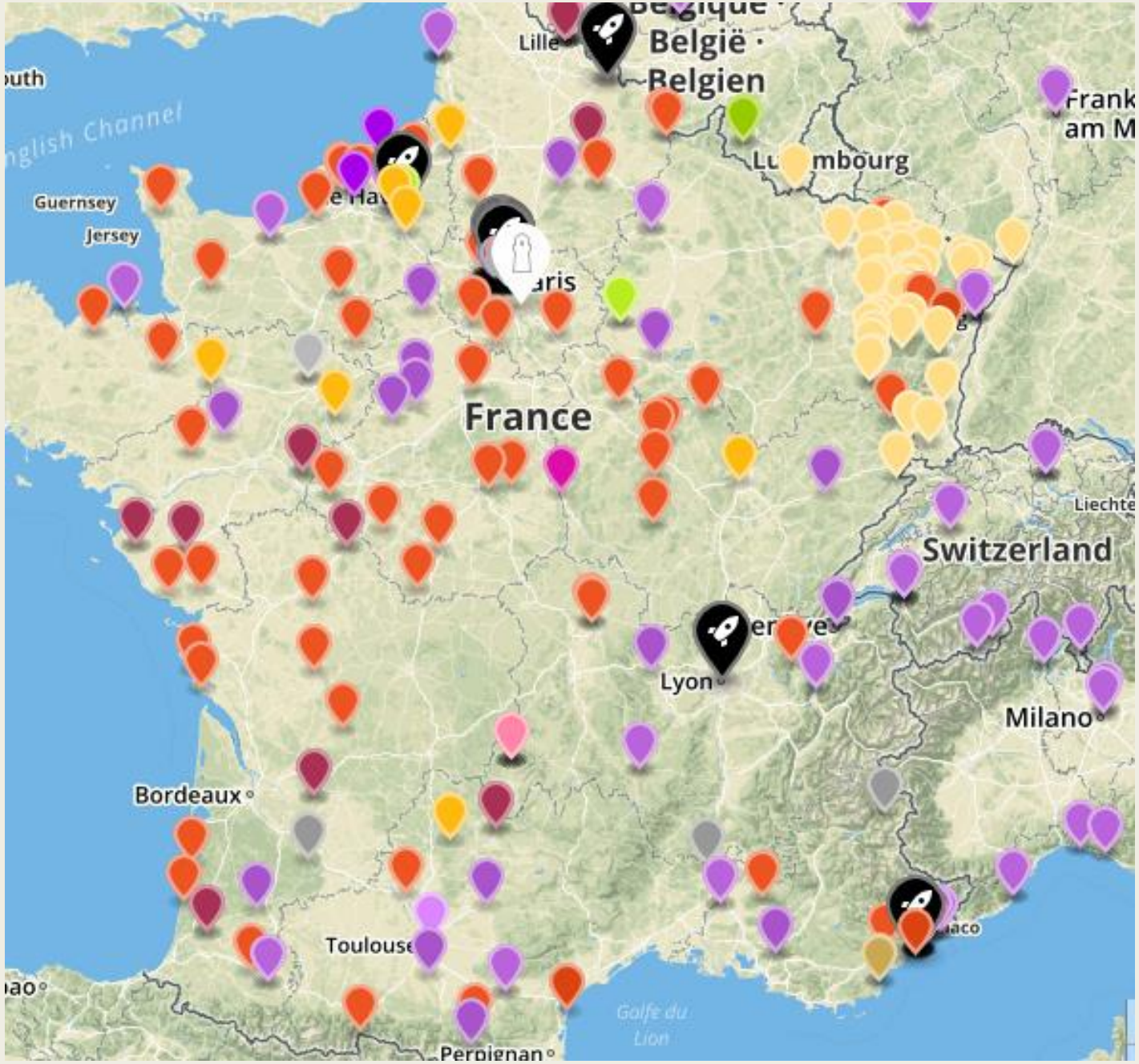
	Avant filtrage	Après filtrage
Nombre réponses par entité	83,4	4
Proportion d'entité valide avec 1 réponse correcte(%)	0	49,3
Proportion d'entité valide sans réponse(%)	0,2	8,3
Proportion entité valide ambiguë(%)	99,5	39,5
Proportion fausse entité sans réponse(%)	13,4	83,9

Heuristiques

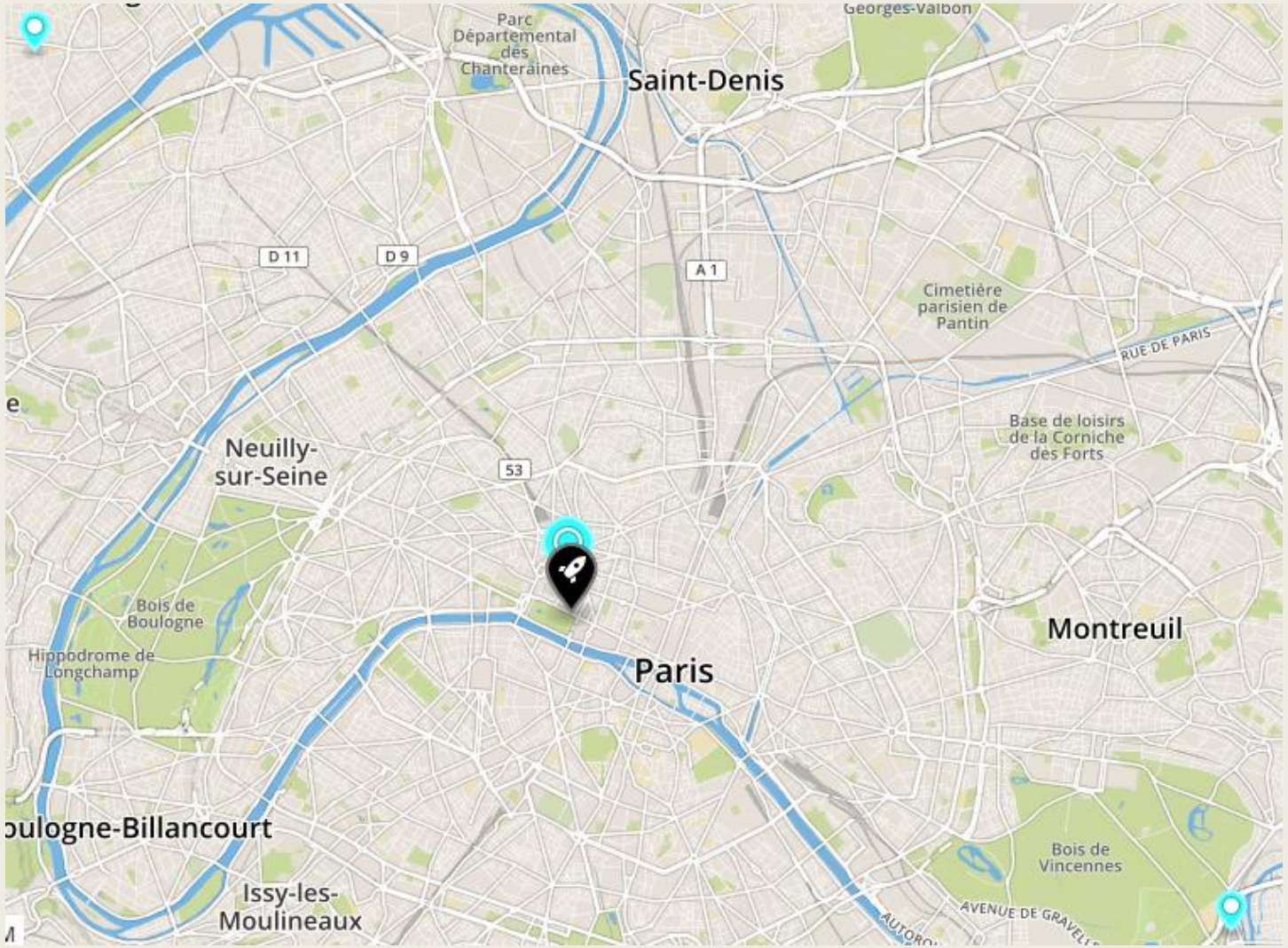
- Centroïde
- Voisin textuel
- Notre méthode



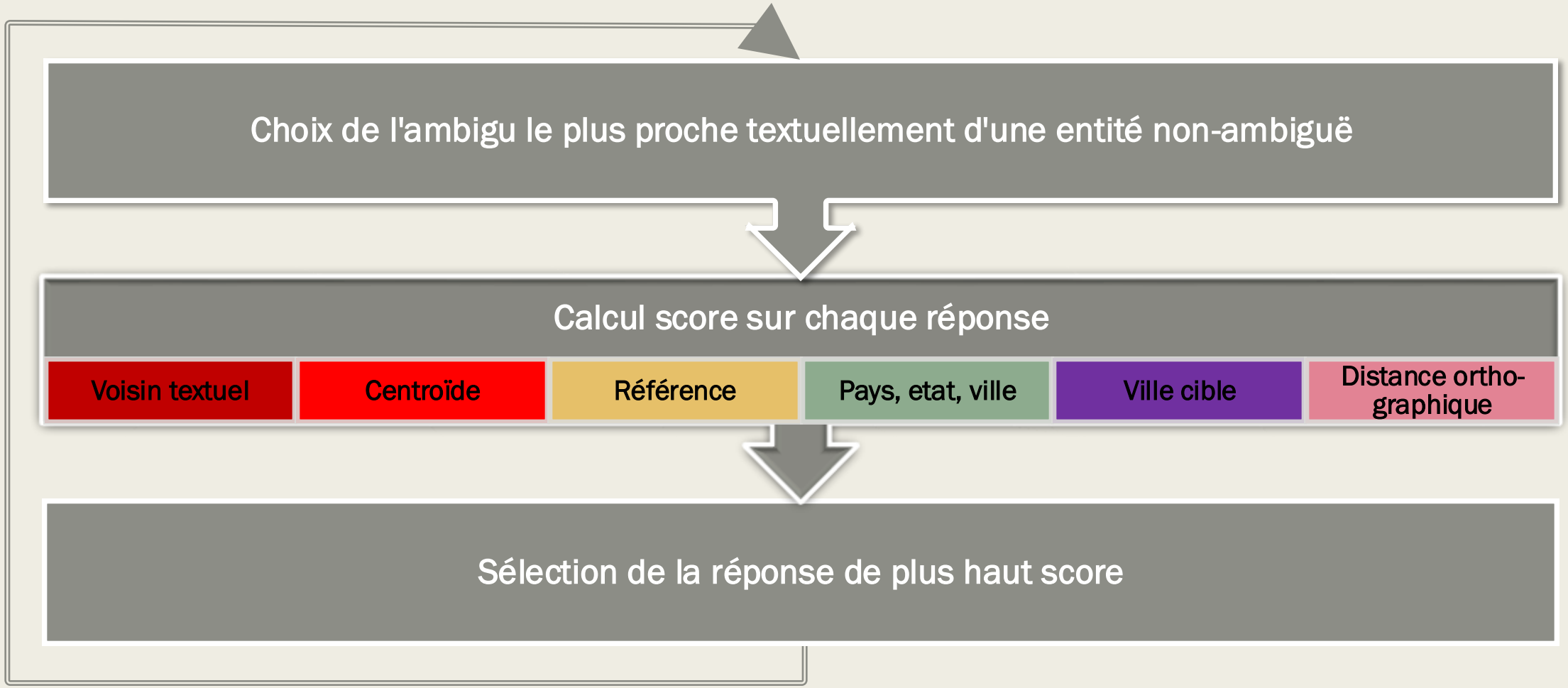
Centroïde


België
Belgien
Guernsey
Jersey
France
Switzerland
Lyon
Milano
Bordeaux
Toulouse
Perpignan
Golfe du Lion

Voisin textuel



Méthode par score



Evaluation heuristiques

	STREET	OTHERS	ESNE
CENTROIDE	64,2	55,5	59,7
VOISINS TEXTUELS	**77,9**	55,5	66,4
METHODE PAR SCORE	77,0	**61,0**	**68,8**

DÉTECTION DES ENE PAR APPRENTISSAGE SUPERVISÉ

Problématiques

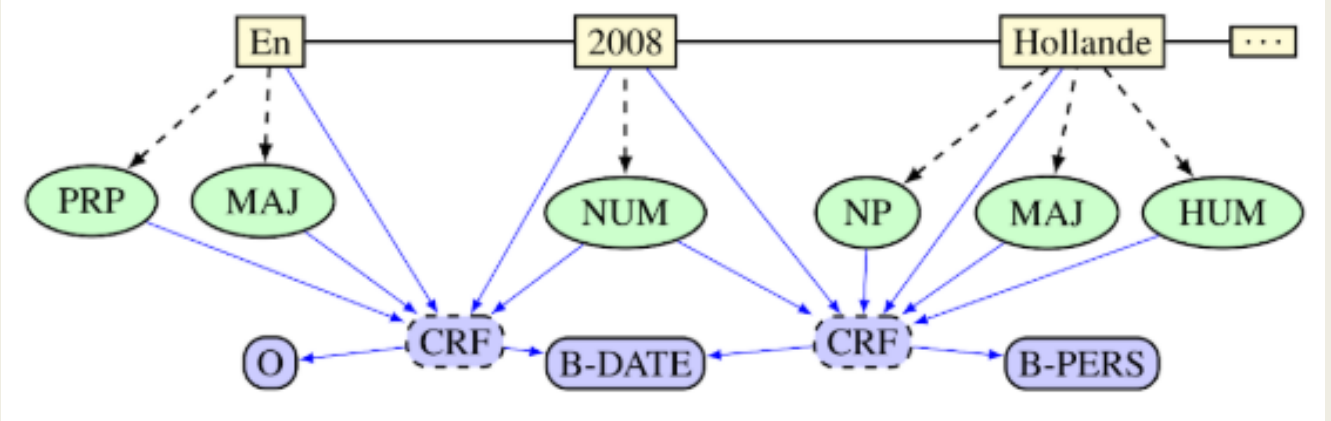
- Peut-on entraîner un modèle d'apprentissage supervisé à reconnaître les ESNE?
- Peut-il généraliser à partir d'une entrée bruitée?

Reconnaissance d’entités nommées

- Stanford-NLP
- CogComp
- MITIE

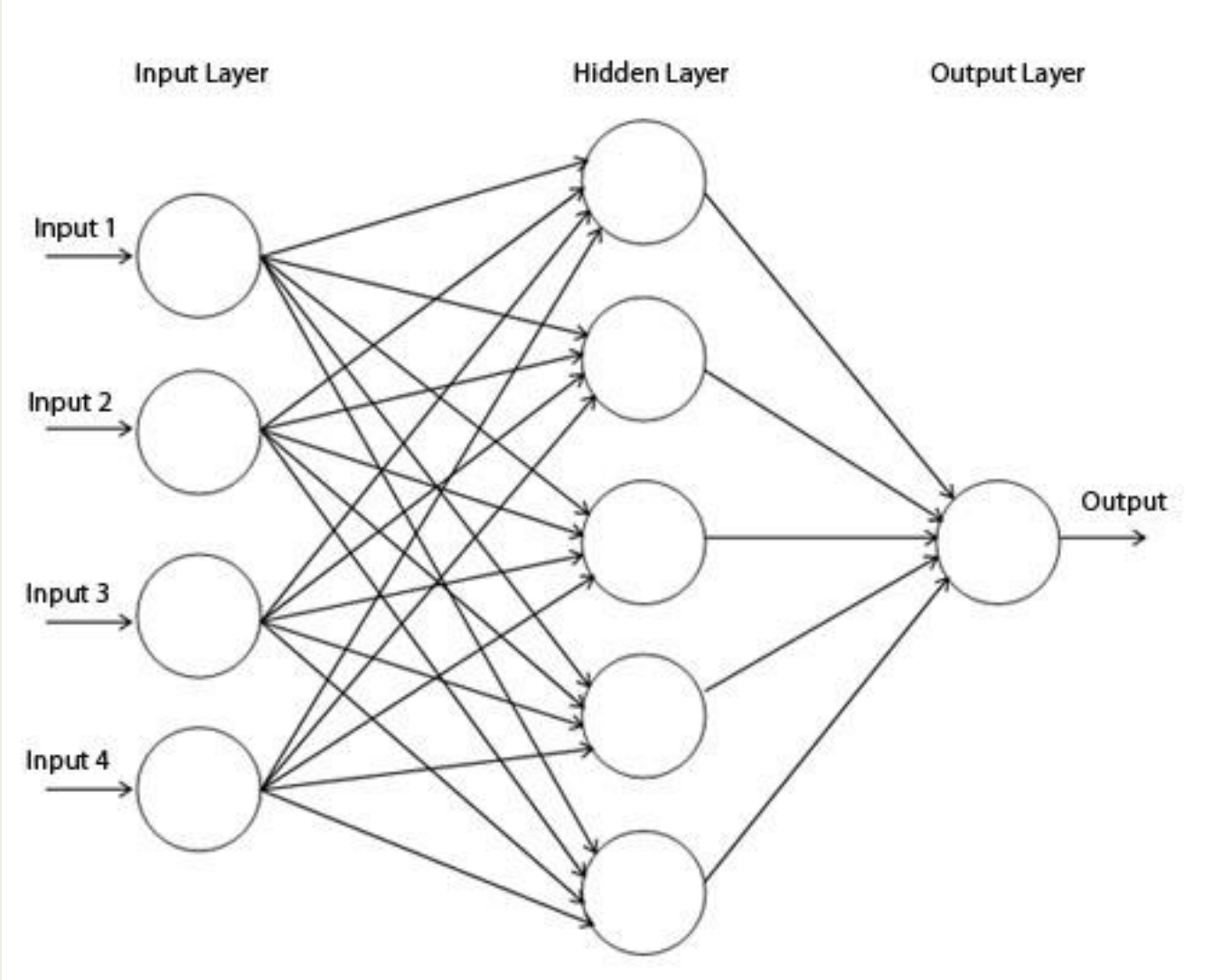
Stanford-NLP

- Groupe de traitement du langage naturel de l'Université de Stanford
- Champ aléatoire conditionnel linéaire



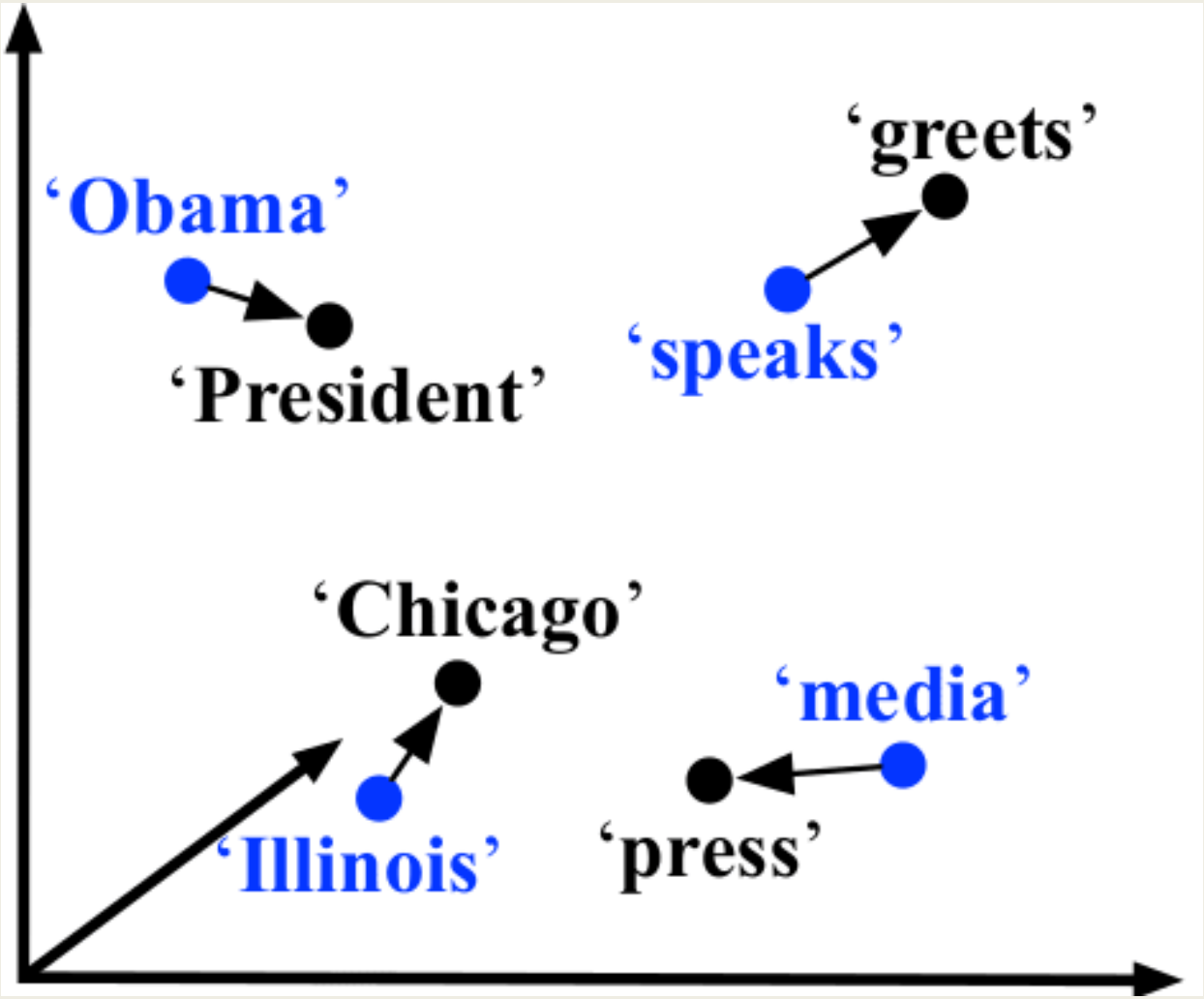
CogComp

- Groupe de calcul cognitif de l'Université de Pennsylvanie
- « Averaged » Perceptron



MITIE

- Institut de Technologie du Massachussets
- Word Embedding + Machine à Vecteur de Support (SVM)



Evaluation

- Ajout : une entité est annotée par le NER mais pas par Perdido
- Suppression : une entité est annotée par Perdido mais pas par le NER
- Erreur balisage : une entité est annotée par Perdido mais les balises de l'annotation du NER sont mal placées

Mesures

- Le Slot Error Rate (SER) : le taux d'erreur total
- Le rappel : la proportion d'entités correctement annotés par rapport à toutes les entités qu'il fallait annoter
- La précision : la proportion d'entités correctement annotées par rapport à celles qu'on a annotées
- La précision du balisage : les entités sont elles correctement balisées?

Résultats

Taux apprentissage	Slot Error Rate			Rappel			Précision			Précision du balisage		
	Stanford	CogComp	MITIE	Stanford	CogComp	MITIE	Stanford	CogComp	MITIE	Stanford	CogComp	MITIE
30	**29,88**	44,46	73,07	**77,23**	69,42	49,58	**97,58**	90,74	75,05	**84,44**	72,98	56,38
40	**25,21**	41,04	64,13	**81,19**	72,10	52,57	**98,18**	92,97	82,55	**86,33**	73,14	65
50	**19,64**	38,05	75,29	**86,38**	73,95	45,78	**97,57**	93,76	74,27	**88,83**	75,82	57,38
60	**17,26**	35,80	57,24	**88,38**	77,27	52,36	**97,46**	92,48	93,39	**90,10**	76,23	72,37
70	**16,47**	35,39	56,33	**89,00**	78,53	53,02	**97,74**	92,36	93,57	**90,25**	74,90	73,45

Généralisation

NER	Slot Error Rate	
	Perdido	Corrigé
Stanford	12,50	18,41
CogComp	27,90	33,44

Généralisation

NER	Nombre ESNE « quai » détectées	Nombre ESNE « quai » détectées jamais vu
Stanford	42	26
CogComp	43	27

CONCLUSION