Fouille de textes

Reconnaissance d'entités nommées

Solen Quiniou

solen.quiniou@univ-nantes.fr

Master 2 ATAL – Université de Nantes

Année 2020-2021

Plan du cours

Plan du cours

Définitions

- Les entités nommées (*Named Entities*) sont des unités lexicales particulières.
- Elles sont implicitement définies par une énumération de ce qu'elles peuvent représenter : noms de personnes, noms d'organisation, noms de lieux, dates, unités monétaires, pourcentages. . .

→ Il existe diverses typologies selon les besoins applicatifs et selon les campagnes d'évaluation associées.

- La reconnaissance d'entités nommées (REN) consiste généralement en deux ou trois étapes :
  1. L'identification des entités nommées dans un texte ;
  2. La catégorisation de ces entités nommées ;
  3. La normalisation éventuelle des entités nommées.

Exemple

L'ancien premier ministre socialiste Lionel Jospin a confirmé, jeudi 28 septembre, sur RTL, qu'il ne sera pas candidat à l'investiture socialiste pour la présidentielle de 2007.

Exemple

L'ancien premier ministre socialiste Lionel Jospin a confirmé, jeudi 28 septembre, sur RTL, qu'il ne sera pas candidat à l'investiture socialiste pour la présidentielle de 2007.

1. Identification des entités nommées

   Lionel Jospin, jeudi 28 septembre, RTL, 2007.

Exemple

L'ancien premier ministre socialiste Lionel Jospin a confirmé, jeudi 28 septembre, sur RTL, qu'il ne sera pas candidat à l'investiture socialiste pour la présidentielle de 2007.

1. Identification des entités nommées

   Lionel Jospin, jeudi 28 septembre, RTL, 2007.

2. Catégorisation des entités nommées

   L'ancien premier ministre socialiste ***<PERS>Lionel Jospin</PERS>*** *a confirmé,* ***<DATE>jeudi 28 septembre</DATE>****, sur* ***<ORG>RTL</ORG>****, qu'il ne sera pas candidat à l'investiture socialiste pour la présidentielle de* ***<DATE>2007</DATE>****.*

Exemple

L'ancien premier ministre socialiste Lionel Jospin a confirmé, jeudi 28 septembre, sur RTL, qu'il ne sera pas candidat à l'investiture socialiste pour la présidentielle de 2007.

1. Identification des entités nommées

   Lionel Jospin, jeudi 28 septembre, RTL, 2007.

2. Catégorisation des entités nommées

   L'ancien premier ministre socialiste ***<PERS>Lionel Jospin</PERS>*** *a confirmé,* ***<DATE>jeudi 28 septembre</DATE>****, sur* ***<ORG>RTL</ORG>****, qu'il ne sera pas candidat à l'investiture socialiste pour la présidentielle de* ***<DATE>2007</DATE>****.*

3. Normalisation des entités nommées

   L. Jospin $\rightarrow$ *Lionel Jospin.*

Quelques applications des entités nommées (1)

- Aide à l'analyse syntaxique
  - *He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to **<LOC>Egypt</LOC>** and **<LOC>Jordan</LOC>**.*
  - *He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to **<LOC>Egypt</LOC>** and **<ORG>Likud party</ORG>** politicians.*

- Co-référence
  - ***<PERS>John</PERS>** bought a new computer. **He** was able to process XML.*

- Traduction
  - ***<PERS>Jack London</PERS>** was an American writer. → Jack **London** était un écrivain américain.*
  - ***<LOC>London</LOC>** is where I lived my best years. → C'est à **Londres** que j'ai vécu mes meilleures années.*

Quelques applications des entités nommées (2)

- Extraction d'information et veille : les EN peuvent être utilisées pour remplir des bases de données ou pour signaler de nouveaux documents contenant une EN particulière.

- Tâche de question-réponse : les EN permettent d'identifier le type de réponse attendu.

- Anonymisation : les EN peuvent être utilisées pour anonymiser des documents (documents médicaux, par exemple).

Plan du cours

Principales campagnes d'évaluation et catégories

- Les 7 catégories principales les plus courantes sont les suivantes :
  - ENAMEX : `ORGANISATION`, `LIEU`, `PERSONNE`.
  - TIMEX : `DATE`, `TEMPS` (expression temporelle).
  - NUMEX : `MONNAIE`, `MONTANT`, `POURCENTAGE`.

→ Les catégories utilisées pour annoter les entités nommées dépendent des campagnes d'évaluation considérées et du type de texte annoté lors de celles-ci.

Principales campagnes d'évaluation et catégories

- Les 7 catégories principales les plus courantes sont les suivantes :
  - ENAMEX : `ORGANISATION`, `LIEU`, `PERSONNE`.
  - TIMEX : `DATE`, `TEMPS` (expression temporelle).
  - NUMEX : `MONNAIE`, `MONTANT`, `POURCENTAGE`.

→ Les catégories utilisées pour annoter les entités nommées dépendent des campagnes d'évaluation considérées et du type de texte annoté lors de celles-ci.

- Conférences MUC (*Message Understanding Conference*) : 1996-1997
  - Le corpus contient des textes journalistiques en anglais.
  - Les entités nommées considérées sont les 7 catégories définies précédemment.

Principales campagnes d'évaluation et catégories

- Les 7 catégories principales les plus courantes sont les suivantes :
  - ENAMEX : `ORGANISATION`, `LIEU`, `PERSONNE`.
  - TIMEX : `DATE`, `TEMPS` (expression temporelle).
  - NUMEX : `MONNAIE`, `MONTANT`, `POURCENTAGE`.

→ Les catégories utilisées pour annoter les entités nommées dépendent des campagnes d'évaluation considérées et du type de texte annoté lors de celles-ci.

- Conférences MUC (*Message Understanding Conference*) : 1996-1997
  - Le corpus contient des textes journalistiques en anglais.
  - Les entités nommées considérées sont les 7 catégories définies précédemment.

- Campagnes CoNLL (*Conf. on Natural Language Learning*) : 2002-2003
  - Le corpus contient des textes journalistiques en espagnol, hollandais, allemand et anglais.
  - Les entités nommées considérées sont les 3 catégories `ENAMEX` ainsi qu'une catégorie `divers` pour annoter les entités nommées n'appartenant pas aux 3 catégories `ENAMEX`.

Ambiguïtés sémantiques des EN

- Homonymie
  - *Orange* : la ville, la société ;
  - *Vienne* : la ville en France, la ville en Autriche ;
  - *Leclerc* : le Maréchal, l'homme d'affaires.

- Métonymie
  - *Leclerc* : le Maréchal, le char ;
  - *Leclerc* : l'homme d'affaires, le supermarché, le groupe financier.

- Polysémie
  - *Jacques Chirac* : le président de la république, le maire de Paris.
  - *France* : le lieu, le gouvernement.

Ambiguïtés des frontières des EN

- Coordination
  - *Bill and Hillary Clinton flew to Chicago last month.*

  → *<PERS>Bill</PERS> and <PERS>Hillary Clinton</PERS> flew to Chicago last month.*

  → *<PERS type="coll">Bill and Hillary Clinton</PERS> flew to Chicago last month.*

  → *<PERS>Bill Clinton</PERS> and <PERS>Hillary Clinton</PERS> flew to Chicago last month.*

- Imbrication
  - *L'Université de Nantes* : *<ORG>L'Université de Nantes</ORG>* ou *<ORG>L'Université de <LOC>Nantes</LOC> </ORG>* ?

- Prise en compte des modificateurs
  - *Sir Paul Mc Cartney* : *<PERS>Sir Paul Mc Cartney</PERS>* ou *Sir <PERS>Paul Mc Cartney</PERS>* ?
  - *Secretary of State Colin Powell* : *<PERS>Secretary of State Colin Powell</PERS>* ou *Secretary of State <PERS>Colin Powell</PERS>* ?

Mesures d'évaluation

- La précision, le rappel ainsi que la F-mesure peuvent être utilisées pour calculer les entités nommées correctement étiquetées.

Mesures d'évaluation

- La précision, le rappel ainsi que la F-mesure peuvent être utilisées pour calculer les entités nommées correctement étiquetées.

- Une nouvelle mesure pour la reconnaissance d'entités nommées a également été proposée. Il s'agit du *Slot Error Rate* (SER) [MKSW99] :

$$\textbf{SER} = \frac{\lambda_I \times I + \lambda_D \times D + \lambda_{TF} \times TF + \lambda_T \times T + \lambda_F \times F}{\text{Nombre d'EN de la référence}}$$

  - I (insertions) est le nombre d'EN détectées dans l'hypothèse et n'apparaissant pas dans la référence ;
  - D (suppressions) est le nombre d'EN de la référence n'apparaissant pas dans l'hypothèse ;
  - T (erreurs de type) est le nombre d'EN détectées dans l'hypothèse mais avec une catégorie incorrecte ;
  - F (erreurs de frontière) est le nombre d'EN détectées dans l'hypothèse mais avec des frontières incorrectes ;
  - TF (erreurs de type et de frontière) est le nombre d'EN détectées dans l'hypothèse mais avec une catégorie et des frontières incorrectes.

Plan du cours

Indices internes et externes

Des indices permettent d'aider à l'identification et à la catégorisation des entités nommées [McD96].

- Les indices internes s'appuient uniquement sur l'EN elle-même.
  - Majuscule en début de mot ou sur toutes les lettres du mot (acronymes).
  - Prénoms ou marqueurs générationnels : ***Lionel*** *Jospin, Benoît* ***XVI***.
  - Mots ou affixes de type classifiant : *la* ***Banque*** *Populaire, Microsoft* ***Inc.****, le* ***Mont*** *Blanc, l'****avenue*** *des Champs Élysées*.
  - Sigles ou éperluettes : *Crédit Agricole* ***SA***.

→ Des lexiques peuvent être utilisés, notamment pour les noms de lieux, les noms de marques, les prénoms.

- Les indices externes utilisent ce qui apparaît dans le contexte immédiat de l'EN (à gauche et à droite) voire dans le contexte du corpus.
  - Informations supplémentaires ou propriétés spécifiques (titres, grades. . .) : ***Monsieur*** *Jospin,* ***Général*** *Leclerc, l'****entraîneur*** *Aimé Jacquet, le* ***groupe*** *Radiohead, the Coca-Cola* ***company***.

  → Ce type d'indice est généralement précisé lors de la première occurrence de l'EN dans le texte, d'où l'importance de propager ces informations.

Approches symboliques

- Les approches symboliques consiste en l'utilisant de règles contextuelles pour identifier et catégoriser les entités nommées [McD96, Fou02].
- Les règles sont écrites à la main par des experts ; elles prennent la forme de patrons d'extraction.
- Les règles reposent sur l'exploitation de descriptions linguistiques, d'indices, de mots déclencheurs et de lexiques de noms propres.

- Exemple
  - Prénom + Mot avec Majuscule $\rightarrow$ `Personne`.
  - Mot inconnu + « Inc. » $\rightarrow$ `Organisation`.
  - Lieu + verbe d'action $\rightarrow$ `Organisation`.

$\rightarrow$ Les systèmes symboliques obtiennent de bons résultats sur du texte bien formé [GGC09] mais l'écriture des règles est fastidieuse (par exemple, le système Nemesis [Fou02] utilise 93 règles).

Approches à base d'apprentissage supervisé

- Les approches à base d'apprentissage supervisé visent à apprendre les règles d'extraction de manière automatique.

→ L'apprentissage est réalisé à partir d'un corpus annoté en entités nommées.

- Parmi les principales aproches utilisées, on retrouve des approches classiques d'apprentissage supervisé : les SVM [IK02], les modèles de Markov cachés [BMSW97], les modèles à entropie maximale [BSAG98] et les champs conditionnels aléatoires (CRF) [FGM05, BC10].

→ Les modèles à base de CRF sont parmi les plus populaires et les plus performants pour la reconnaissance d'EN.

→ Les performances des systèmes à base d'apprentissage supervisé augmentent proportionnellement avec la qualité et la quantité de données d'apprentissage.

Approches à base d'apprentissage semi-supervisé

- Les approches à base d'apprentissage semi-supervisé utilisent un corpus non annoté ainsi qu'un expert mais avec un faible degré de supervision.

- La technique principalement utilisée pour l'apprentissage semi-supervisé est l'amorçage [CS99]. Cette technique consiste en l'apprentissage automatique de patrons d'extraction en s'appuyant sur un ensemble d'amorces.
  1. Pour extraire des noms de pays ou de villes, le système demande tout d'abord à l'utilisateur d'en donner quelques uns ;
  2. Les phrases du corpus contenant ces noms sont analysées et des indices contextuels communs sont retenus (marqueurs lexicaux, typographiques. . .) ;
  3. Le système essaie ensuite de trouver de nouveaux noms de pays ou de villes apparaissant dans les mêmes contextes.

→ Les systèmes à base d'apprentissage semi-supervisé obtiennent des résultats proches de ceux obtenus par les systèmes à base d'apprentissage supervisé mais nécessitent moins de données annotées.

Approches mixtes

- Les approches mixtes (ou hybrides) tirent parti des avantages respectifs des approches symboliques et des approches à base d'apprentissage.

- Il existe deux types d'approches mixtes :
  1. Soit les règles sont apprises automatiquement sur un corpus annoté en entités nommées puis elles sont révisées par un expert [NRC+12] ;
  2. Soit les règles sont écrites par un expert puis elles sont enrichies automatiquement pour améliorer leur couverture des entités nommées du corpus [MGM98, OS12].

Approches combinées avec d'autres tâches

- L'information recherchée n'est pas obligatoirement de nature textuelle.

→ Les systèmes de REN doivent ainsi être adaptés pour prendre en compte les spécificités de la modalité traitée.

- Modalité textuelle bruitée [LZW+12, RME11]
  - Les textes correspondent à des SMS ou à des messages publiés sur des réseaux sociaux (Twitter, Facebook).
  - → Certains phénomènes rendent difficile la reconnaissance des EN : abréviations, écriture phonétique, rébus typographiques (*2m1* pour *demain*).
- Modalité orale [PO01, ZFS+04, GGC09, BC10]
  - Les données orales sont transformées en données textuelles grâce à un système de transcription de la parole puis la tâche de REN est réalisée.
  - → Différents phénomènes complexifient la reconnaissance des EN : hésitations, faux départs, répétitions, manque de structuration (absence de casse, de ponctuation), erreurs de transcription, erreurs résultant des mots hors-vocabulaire.
- Modalité manuscrite [SPN11, DR12]
  - On retrouve des problématiques similaires à la modalité orale mais avec des difficultés moindres.

Plan du cours

Exercice sur les entités nommées (1)

Soit le texte suivant :

```
Three U.S. Senators said they will propose a temporary ban on imports
of all Toshiba products due to the company 's illegal sales of
sensitive high-technology goods to the Soviet Union . Senator Jake
Garn , John Heinz and Richard Selby said at a hearing of the senate
banking committee on export control , they will offer the proposal
as part of a major trade bill when it is brought before the senate
this summer . Garn , a Utah Republican , said " I am talking about
specific retribution on a company that endangers the security of their
own country and ours . "
```

Questions

1. Listez les entités nommées `ENAMEX` du texte et indiquez leur catégorie.
2. Existe-t-il des ambiguïtés ? Si oui, de quelle(s) nature(s) ?

Exercice sur les entités nommées (2)

Soit l'annotation suivante réalisée par un système automatique :

```
Three <LOCATION>U.S.</LOCATION> Senators said they will propose a
temporary ban on imports of all <ORGANIZATION>Toshiba</ORGANIZATION>
products due to the company 's illegal sales of sensitive
high-technology goods to the <LOCATION>Soviet Union</LOCATION> .
Senator <PERSON>Jake Garn</PERSON> , <PERSON>John Heinz</PERSON>
and <PERSON>Richard Selby</PERSON> said at a hearing of the
<ORGANIZATION>senate</ORGANIZATION> banking committee on export
control , they will offer the proposal as part of a major
trade bill when it is brought before the senate this summer .
<PERSON>Garn</PERSON> , a <ORGANIZATION>Utah Republican</ORGANIZATION>
, said " I am talking about specific retribution on a company that
endangers the security of their own country and ours . "
```

Questions

1. Quelles sont les erreurs de reconnaissance des entités nommées, par rapport à votre référence de la question précédente ? (parmi les 5 types d'erreurs utilisées dans le calcul du SER)
2. Quel est alors le SER obtenu ?

Plan du cours

Conclusion

- Les entités nommées sont des informations lexicales importantes à extraire dans de nombreuses applications.

- Différentes typologies de catégories peuvent être considérées, selon l'application visée, et peuvent inclure des catégories principales et des sous-catégories.

→ Les catégories considérées contiennent généralement les noms de personnes, de lieux et d'organisations (ENAMEX).

- Des ambiguïtés peuvent exister quant à l'annotation des catégories d'EN, à-cause de problèmes d'homonymie, de métonymie et de polysémie mais également en ce qui concerne l'annotation des frontières des EN.

- Les approches pour la REN peuvent être symboliques (écriture manuelle de règles) et/ou à base d'apprentissage (apprentissage automatique des règles).

- La tâche de REN peut être effectuée sur des données de modalité non textuelle. Elle nécessite alors la transformation de ces données en données textuelles et la prise en compte des spécificités de ces données.

→ Les données de modalité orale sont parmi les plus fréquemment considérées.

Plan du cours

Références I

- F. Béchet and E. Charton, *Unsupervised knowledge acquisition for extracting named entities from speech.*, Proc. of ICASSP, 2010, pp. 5338–5341.
- D.M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, *Nymble : a high-performance learning name-finder*, Proc. of ANLP, 1997, pp. 194–201.
- A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, *NYU : Description of the MENE named entity system as used in MUC-7*, Proc. of MUC, 1998, pp. 1–6.
- M. Collins and Y. Singer, *Unsupervised models for named entity classification*, Proc. of the joint conf. EMNLP-WVLC, 1999, pp. 100–110.
- M. Dinarelli and S. Rosset, *Tree-structured named entity recognition on OCR data : Analysis, processing and results*, Proc. of LREC, 2012, pp. 1266–1272.

Références II

- J.R. Finkel, T. Grenager, and C. Manning, *Incorporating non-local information into information extraction systems by gibbs sampling*, Proc. of the ACL, 2005, pp. 363–370.
- N. Fourour, *Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français*, Actes de TALN, 2002, pp. 265–274.
- S. Galliano, G. Gravier, and L. Chaubard, *The ESTER 2 evaluation campaign for the rich transcription of french radio broadcasts*, Proc. of Interspeech, 2009, pp. 2583–2586.
- H. Isozaki and H. Kazawa, *Efficient support vector classifiers for named entity recognition*, Proc. of COLING, 2002, pp. 390–396.
- X. Liu, M. Zhou, F. Wei, Z. Fu, and X. Zhou, *Joint inference of named entity recognition and normalization for tweets*, Proc. of ACL, 2012, pp. 526–535.

Références III

- David D. McDonald, *Corpus processing for lexical acquisition*, ch. Internal and external evidence in the identification and semantic categorization of proper names, pp. 61–76, MIT Press, 1996.
- A. Mikheev, C. Grover, and M. Moens, *Description of the LTG system used for MUC-7*, Proc. of MUC, 1998, pp. 1–12.
- J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, *Performance measures for information extraction*, Proc. of DARPA Broadcast News, 1999, pp. 249–252.
- A. Nagesh, G. Ramakrishnan, L. Chiticariu, R. Krishnamurthy, A. Dharkar, and P. Bhattacharyya, *Towards efficient named-entity rule induction for customizability*, Proc. of the joint conf. EMNLP-CoNLL, 2012, pp. 128–138.
- M. Oudah and K.F. Shaalan, *A pipeline arabic named entity recognition using a hybrid approach*, Proc. of COLING, 2012, pp. 2159–2176.

Références IV

- D.D. Palmer and M. Ostendorf, *Improving information extraction by modeling errors in speech recognizer output*, Proc. of HLT, 2001, pp. 1–5.
- A. Ritter, S.C. Mausam, and O. Etzioni, *Named entity recognition in tweets : An experimental study*, Proc. of EMNLP, 2011, pp. 1524–1534.
- K. Subramanian, R. Prasad, and P. Natarajan, *Robust named entity detection from optical character recognition output*, IJDAR **14** (2011), 189–200.
- L. Zhai, P. Fung, R. Schwartz, M. Carpuat, and D. Wu, *Using n-best lists for named entity recognition from chinese speech*, Proc. of the joint conf. HLT-NAACL, 2004, pp. 37–40.