Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique

Yoann Dupont [1,2]

[1]Lattice (CNRS, ENS, Université Sorbonne Nouvelle, PSL Research University, USPC), UMR 8094, 1 rue Maurice Arnoux, 92120 Montrouge, France

[2]Expert System France, 207 rue de Bercy, 75012 Paris, France

Introduction

Cadre

- Reconnaissance d'entités nommées (NER)

Cadre

- Reconnaissance d'entités nommées (NER)
- trouver personnes, lieux, organisations, etc...

Cadre

- Reconnaissance d'entités nommées (NER)
- trouver personnes, lieux, organisations, etc...
- méthodes par apprentissage

Cadre

- Reconnaissance d'entités nommées (NER)
- trouver personnes, lieux, organisations, etc...
- méthodes par apprentissage
- explorer les traits de la littérature

Cadre

- Reconnaissance d'entités nommées (NER)
- trouver personnes, lieux, organisations, etc...
- méthodes par apprentissage
- explorer les traits de la littérature
- comparaison CRF / réseaux de neurones

Cadre

- Reconnaissance d'entités nommées (NER)
- trouver personnes, lieux, organisations, etc...
- méthodes par apprentissage
- explorer les traits de la littérature
- comparaison CRF / réseaux de neurones
- pas d'optimisation des hyperparamètres.

Évaluation

- F1-score (f-mesure)

$$precision = \frac{VP}{VP + FP}$$

$$rappel = \frac{VP}{VP + FN}$$

$$f\text{-}mesure = 2 * \frac{precision * rappel}{precision + rappel}$$

Évaluation

- F1-score (f-mesure)

$$precision = \frac{VP}{VP + FP}$$

$$rappel = \frac{VP}{VP + FN}$$

$$f\text{-}mesure = 2 * \frac{precision * rappel}{precision + rappel}$$

- p-value
  - *paired bootstrap* (10k échantillons)
  - F-mesure

French TreeBank (Abeillé, 2003) en entités nommées

- 10k phrases du Monde de 1989 à 1995
-
-

French TreeBank (Abeillé, 2003) en entités nommées

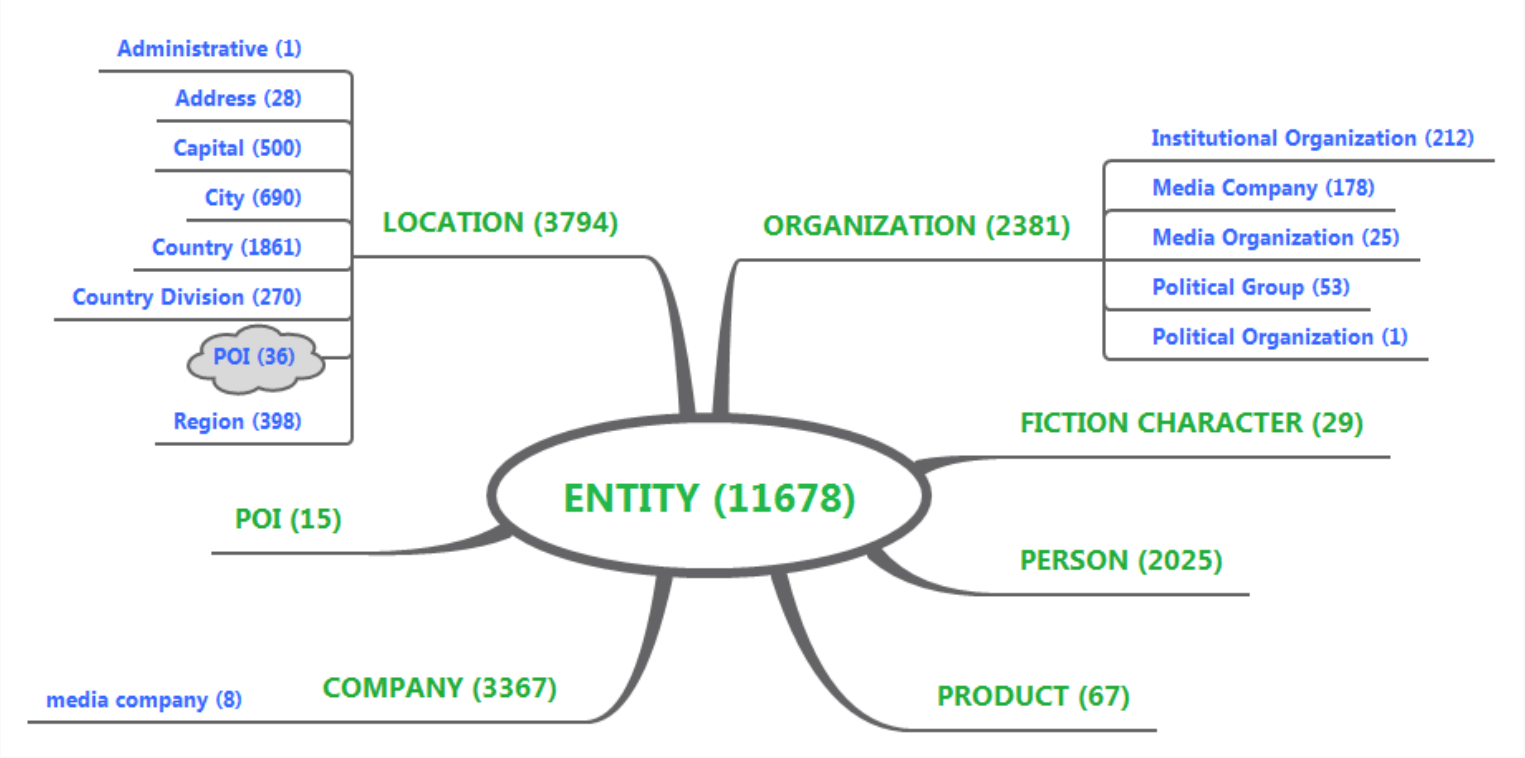
- 10k phrases du Monde de 1989 à 1995
- annotation en entités nommées par Sagot & al. (2012)
-

French TreeBank (Abeillé, 2003) en entités nommées

- 10k phrases du Monde de 1989 à 1995
- annotation en entités nommées par Sagot & al. (2012)
- découpage train / dev / test (Crabbé & Candito, 2008)

French TreeBank (Abeillé, 2003) en entités nommées

- 10k phrases du Monde de 1989 à 1995
- annotation en entités nommées par Sagot & al. (2012)
- découpage train / dev / test (Crabbé & Candito, 2008)



Conditional Random Fields (CRF)

$$p(y|x) = \frac{1}{Z_\lambda(x)} \prod_t \exp\left[\sum_{k=1}^{K} \lambda_k f_k(t, y_t, y_{t-1}, x)\right]$$

Conditional Random Fields (CRF)

$$p(y|x) = \frac{1}{\mathbf{Z_\lambda(x)}} \prod_t \exp\left[\sum_{k=1}^{K} \lambda_k f_k(t, y_t, y_{t-1}, x)\right]$$

Conditional Random Fields (CRF)

$$p(y|x) = \frac{1}{Z_\lambda(x)} \prod_t \exp\left[\sum_{k=1}^{K} \lambda_k \mathbf{f_k}(\mathbf{t}, \mathbf{y_t}, \mathbf{y_{t-1}}, \mathbf{x})\right]$$

Conditional Random Fields (CRF)

$$p(y|x) = \frac{1}{Z_\lambda(x)} \prod_t \exp\left[\sum_{k=1}^{K} \lambda_{\mathbf{k}} f_k(t, y_t, y_{t-1}, x)\right]$$

Conditional Random Fields (CRF)

$$p(y|x) = \frac{1}{Z_\lambda(x)} \prod_t \exp\left[\sum_{k=1}^{K} \lambda_k f_k(t, y_t, y_{t-1}, x)\right]$$

Mots	la	société	Warner	fondée	par	les	frères	Warner
POS	DET	NC	NPP	ADJ	PRP	DET	NC	NPP
sortie	O	O	B-Company	O	O	O	O	B-Person

Conditional Random Fields (CRF)

$$p(y|x) = \frac{1}{Z_\lambda(x)} \prod_t \exp\left[\sum_{k=1}^{K} \lambda_k f_k(t, y_t, y_{t-1}, x)\right]$$

Mots	la	société	Warner	fondée	par	les	frères	Warner
POS	DET	NC	NPP	ADJ	PRP	DET	NC	NPP
sortie	O	**O**	**B-Company**	O	O	O	O	B-Person

Conditional Random Fields (CRF)

$$p(y|x) = \frac{1}{Z_\lambda(x)} \prod_t \exp\left[\sum_{k=1}^{K} \lambda_k f_k(t, y_t, y_{t-1}, x)\right]$$

Mots	la	société	Warner	fondée	par	les	frères	Warner
POS	DET	NC	NPP	ADJ	PRP	DET	NC	NPP
sortie	O	**O**	**B-Company**	O	O	O	O	B-Person

$$f_i(t, y_{t-1}, y_t, x) = \begin{cases} 1 & \textbf{si } mot_{t-1} = \underline{societe} \textbf{ et } POS_t = \underline{NPP} \\ & \quad \textbf{et } y_{t-1} = O \textbf{ et } y_t = B\text{-}Company \\ 0 & \textbf{sinon} \end{cases}$$

Conditional Random Fields (CRF)

$$p(y|x) = \frac{1}{Z_\lambda(x)} \prod_t \exp\left[\sum_{k=1}^{K} \lambda_k f_k(t, y_t, y_{t-1}, x)\right]$$

Mots	la	société	Warner	fondée	par	les	frères	Warner
POS	DET	NC	NPP	ADJ	PRP	DET	NC	NPP
sortie	O	**O**	**B-Company**	O	O	O	O	B-Person

$$f_i(t, y_{t-1}, y_t, x) = \begin{cases} 1 & \textbf{si } mot_{t-1} = \underline{societe} \textbf{ et } POS_t = \underline{NPP} \\ & \quad \textbf{et } y_{t-1} = O \textbf{ et } y_t = B\text{-}Company \\ 0 & \textbf{sinon} \end{cases}$$

besoin de connaissances (Jungermann, 2007) $\rightarrow$ exploration littérature.

Intégrer des connaissances

Traits de Raymond & Fayolle (2010)

Mots	la	société	Warner	fondée	par	les	frères	Warner
traits								
sortie	O	O	B-Company	O	O	O	O	B-Person

Combinaison de plusieurs informations:

-
-
- tout le reste $\longrightarrow$ Partie-du-discours (POS)

Traits légèrement modifiés ici $\longrightarrow$ plus générique

Traits de Raymond & Fayolle (2010)

Mots	la	société	Warner	fondée	par	les	frères	Warner
traits			last-name					last-name
sortie	O	O	B-Company	O	O	O	O	B-Person

Combinaison de plusieurs informations:

- connaissances *a priori* $\Longrightarrow$ ensemble de lexiques
-
- tout le reste $\Longrightarrow$ Partie-du-discours (POS)

Traits légèrement modifiés ici $\Longrightarrow$ plus générique

Traits de Raymond & Fayolle (2010)

Mots	la	société	Warner	fondée	par	les	frères	Warner
traits		société	last-name					last-name
sortie	O	O	B-Company	O	O	O	O	B-Person

Combinaison de plusieurs informations:

- connaissances *a priori* $\Longrightarrow$ ensemble de lexiques
- mots *importants* (forte MI avec sortie) $\Longrightarrow$ laissés tels quels
- tout le reste $\Longrightarrow$ Partie-du-discours (POS)

Traits légèrement modifiés ici $\Longrightarrow$ plus générique

Traits de Raymond & Fayolle (2010)

Mots	la	société	Warner	fondée	par	les	frères	Warner
traits	DET	société	last-name	ADJ	PRP	DET	NC	last-name
sortie	O	O	B-Company	O	O	O	O	B-Person

Combinaison de plusieurs informations:

- connaissances *a priori* $\Longrightarrow$ ensemble de lexiques
- mots *importants* (forte MI avec sortie) $\Longrightarrow$ laissés tels quels
- tout le reste $\Longrightarrow$ Partie-du-discours (POS)

Traits légèrement modifiés ici $\Longrightarrow$ plus générique

Traits de Raymond & Fayolle (2010)

Mots	la	société	Warner	fondée	par	les	frères	Warner
traits	DET	société	last-name	ADJ	PRP	DET	NC	last-name
sortie	O	O	B-Company	O	O	O	O	B-Person

Combinaison de plusieurs informations:

- connaissances *a priori* $\Longrightarrow$ ensemble de lexiques
- mots *importants* (forte MI avec sortie) $\Longrightarrow$ laissés tels quels
- tout le reste $\Longrightarrow$ Partie-du-discours (POS)

Traits légèrement modifiés ici $\Longrightarrow$ plus générique

Traits de Raymond & Fayolle (2010)

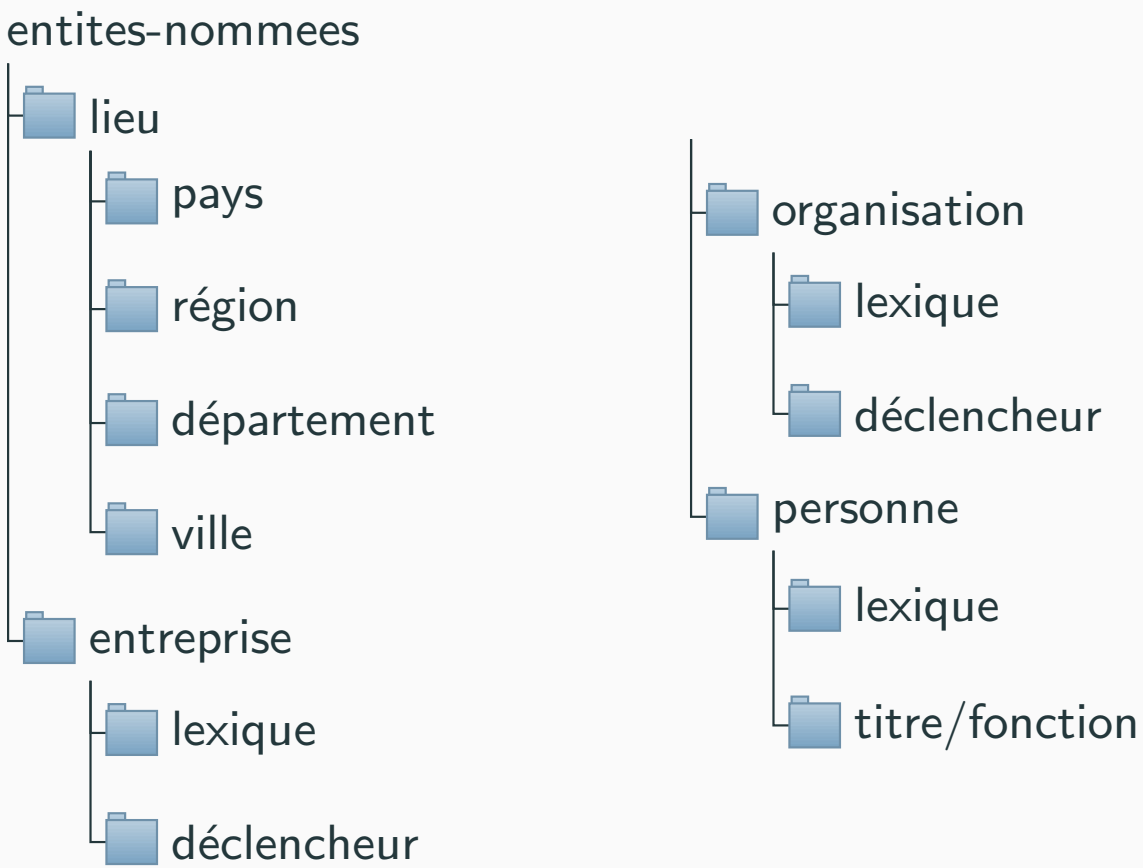
Mots	la	société	Warner	fondée	par	les	frères	Warner
traits	DET	company.trigger	last-name	ADJ	PRP	DET	NC	last-name
sortie	O	O	B-Company	O	O	O	O	B-Person

Combinaison de plusieurs informations:

- connaissances *a priori* $\Longrightarrow$ ensemble de lexiques
- mots *importants* $\Longrightarrow$ lexiques termes déclencheurs
- tout le reste $\Longrightarrow$ Partie-du-discours (POS)

Traits légèrement modifiés ici $\Longrightarrow$ plus générique

Lexiques utilisés



extraits de Wikipedia + corpus d'apprentissage (déclencheurs)

Expériences avec CRF

Généralités

- traits générés sur fenêtre de [-2,2]

Généralités

- traits générés sur fenêtre de [-2,2]
- CRF *étalon*
  - mots
  - préfixes/suffixes (tailles 1 à 5)
  - POS
  - 1 lexique = 1 trait booléen

Généralités

- traits générés sur fenêtre de [-2,2]
- CRF *étalon*
  - mots
  - préfixes/suffixes (tailles 1 à 5)
  - POS
  - 1 lexique = 1 trait booléen
- CRF *article*
  - mots
  - préfixes/suffixes (tailles 1 à 5)
  - traits type Raymond & Fayolle (2010)
    - unigrammes et bigrammes (observation)

Généralités

- traits générés sur fenêtre de [-2,2]
- CRF *étalon*
  - mots
  - préfixes/suffixes (tailles 1 à 5)
  - POS
  - 1 lexique = 1 trait booléen
- CRF *article*
  - mots
  - préfixes/suffixes (tailles 1 à 5)
  - traits type Raymond & Fayolle (2010)
    - unigrammes et bigrammes (observation)
- augmentation du rappel
  - gestion des mots inconnus
  - consistance des annotations

Augmentation du rappel

gestion de l'inconnu:

Mots	la	société	Warner	fondée	par	les	frères	Warner
compte			**<5**					**<5**
inc	la	société	**_unk_**	fondée	par	les	frères	**_unk_**

Augmentation du rappel

gestion de l'inconnu:

Mots	la	société	Warner	fondée	par	les	frères	Warner
compte			**<5**					**<5**
inc	la	société	**_unk_**	fondée	par	les	frères	**_unk_**

consistance des annotations:

- ex: "Calvin Klein"
  - 3 fois en tant que *Person*
  - 2 fois en tant que *Company*
  - Tous "Calvin Klein" non-annotés $\rightarrow$ *Person*
  - *propagation* des annotations

Résultats

Expérience	Précision	Rappel	F-mesure
CRF *étalon*	85.89	76.88	81.13
CRF *article*	86.48	78.58	82.34
$+ \ \mathrm{cct}(\mathrm{inc}_i, \mathrm{trait}_0), \ i \in \{-2,-1,1,2\}$			

p-value $< 0{,}001$

Résultats

Expérience	Précision	Rappel	F-mesure
CRF *étalon*	85.89	76.88	81.13
CRF *article*	86.48	78.58	82.34
+ nom commun à gauche/droite	85.86	78.75	82.15
+ $\mathrm{cct}(\mathrm{inc}_i, \mathrm{trait}_0)$, $i \in \{-2,-1,1,2\}$			

p-value $< 0{,}001$

Résultats

Expérience	Précision	Rappel	F-mesure
CRF *étalon*	85.89	76.88	81.13
CRF *article*	86.48	78.58	82.34
+ nom commun à gauche/droite	85.86	78.75	82.15
+ verbe à droite	86.77	78.92	82.66
+ $\mathrm{cct}(\mathrm{inc}_i, \mathrm{trait}_0)$, $i \in \{-2,-1,1,2\}$			

p-value $< 0{,}001$

Résultats

Expérience	Précision	Rappel	F-mesure
CRF *étalon*	85.89	76.88	81.13
CRF *article*	86.48	78.58	82.34
+ nom commun à gauche/droite	85.86	78.75	82.15
+ verbe à droite	86.77	78.92	82.66
+ cct(inc_i, inc_{i+1}), i $\in \{$-2,1$\}$	88.15	79.95	83.85
+ cct(inc_i, trait_0), i $\in \{$-2,-1,1,2$\}$	**88.41**	80.03	84.05

p-value $<$ 0,001

Résultats

Expérience	Précision	Rappel	F-mesure
CRF *étalon*	85.89	76.88	81.13
CRF *article*	86.48	78.58	82.34
+ nom commun à gauche/droite	85.86	78.75	82.15
+ verbe à droite	86.77	78.92	82.66
+ cct(inc_i, inc_{i+1}), i $\in \{$-2,1$\}$	88.15	79.95	83.85
+ cct(inc_i, trait_0), i $\in \{$-2,-1,1,2$\}$	**88.41**	80.03	84.05
propagation	87.89	**82.34**	**85.02**

p-value $< 0{,}001$

Résultats

Expérience	Précision	Rappel	F-mesure
CRF *étalon*	85.89	76.88	81.13
CRF *article*	86.48	78.58	82.34
+ nom commun à gauche/droite	85.86	78.75	82.15
+ verbe à droite	86.77	78.92	82.66
+ cct(inc_i, inc_{i+1}), i $\in \{$-2,1$\}$	88.15	79.95	83.85
+ cct(inc_i, trait_0), i $\in \{$-2,-1,1,2$\}$	**88.41**	80.03	84.05
propagation	87.89	**82.34**	**85.02**

p-value < 0.001

En somme

Résumé de l'approche:

- gains intéressants (+2 précision, +5.5 rappel, +3.9 F-mesure)
- modèle simple
  - reprendre traits littérature
  - post-traitement simple
- mots inconnus $\rightarrow$ peu d'expériences concluantes
  - test sur autres corpus/langues

En somme

Résumé de l'approche:

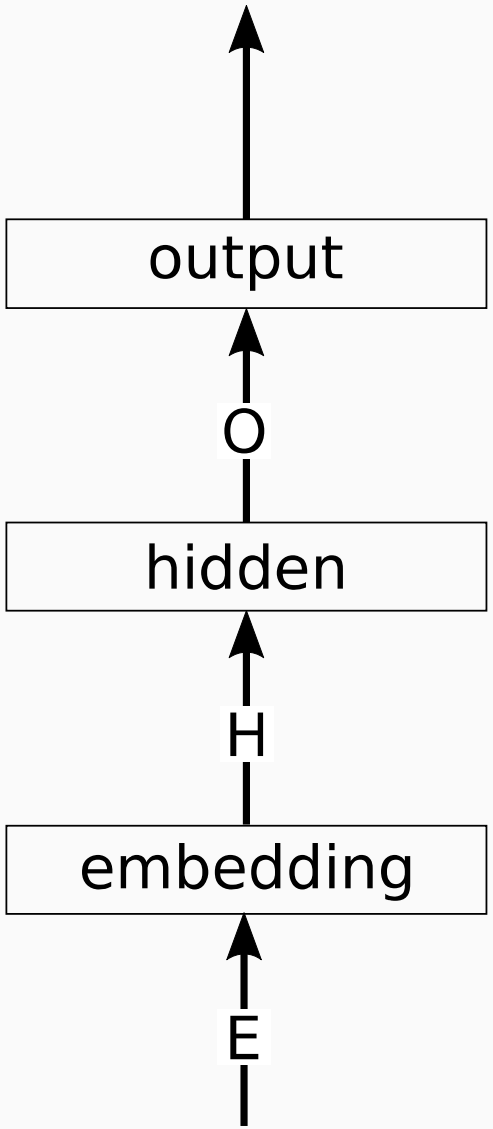
- gains intéressants (+2 précision, +5.5 rappel, +3.9 F-mesure)
- modèle simple
  - reprendre traits littérature
  - post-traitement simple
- mots inconnus $\rightarrow$ peu d'expériences concluantes
  - test sur autres corpus/langues

Comparaison avec état-de-l'art:

- essor réseaux de neurones (NN)
- en particuliers réseaux de neurones récurrents (RNN)
- comparaison CRF / RNN

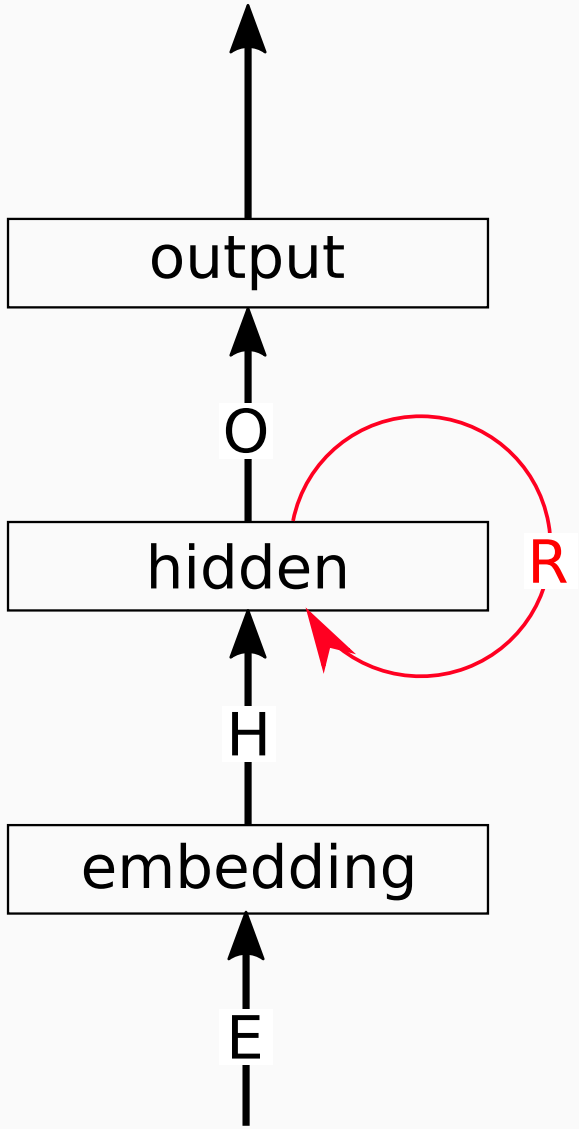
Comparaison avec réseaux de neurones

Réseaux de neurones



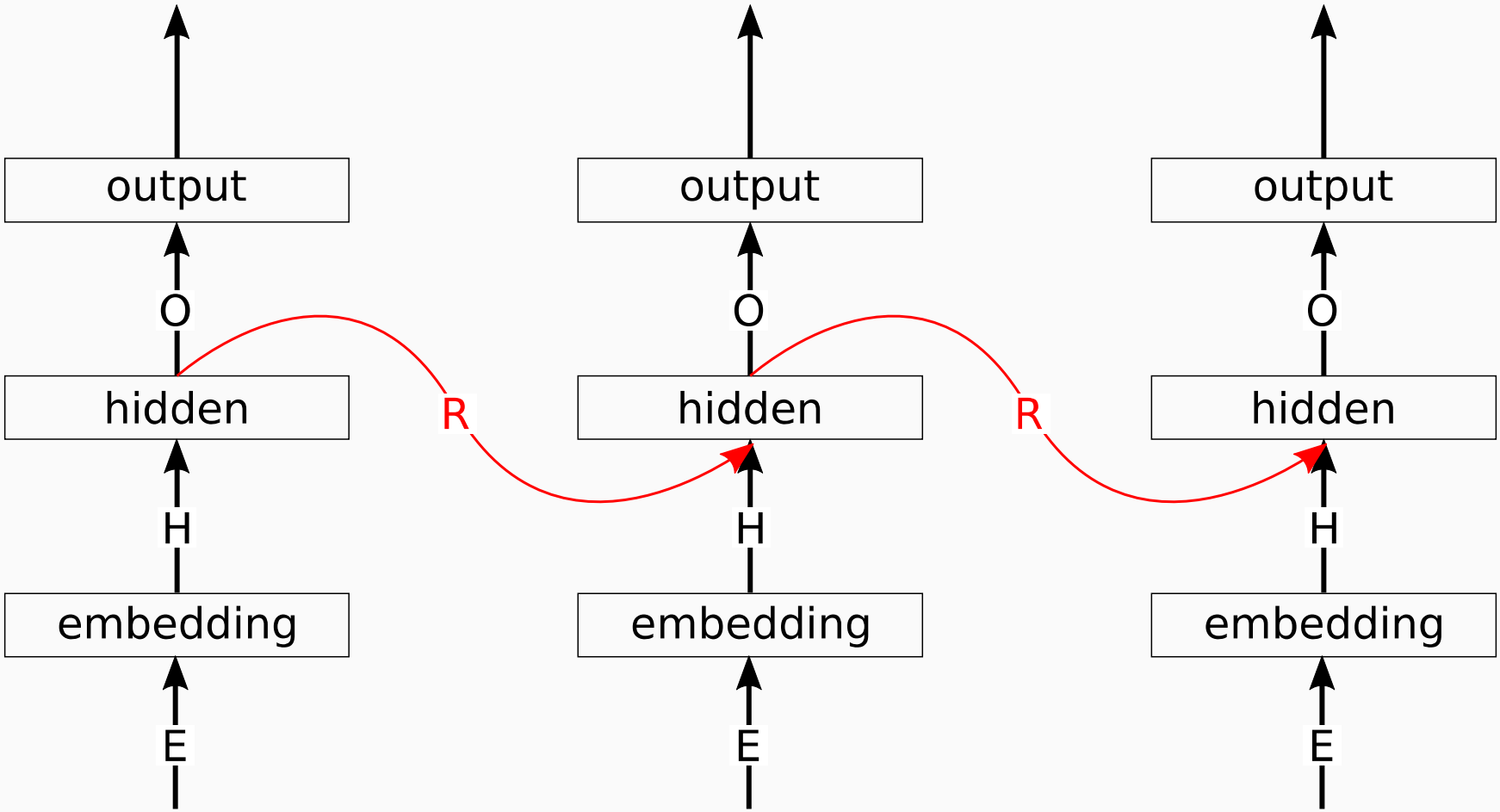
réseau *en avant* classique

Réseaux de neurones



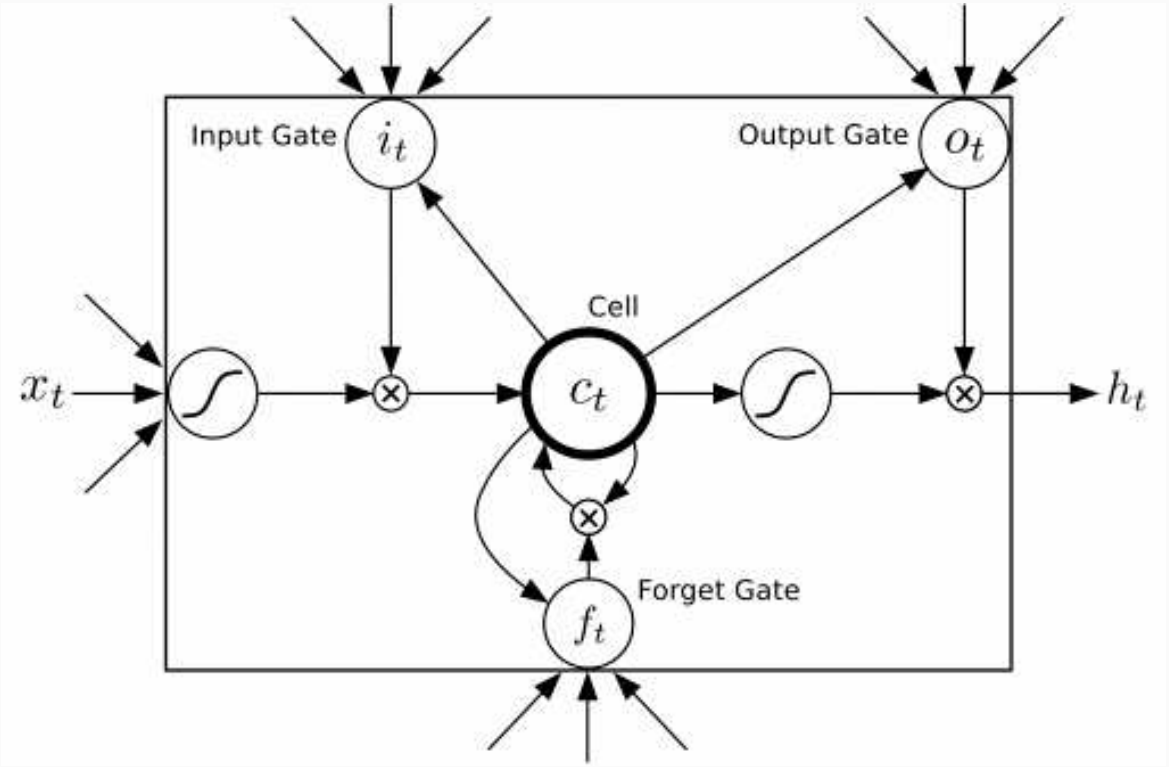
réseau *récurrent* d'Elman (1990)

Réseaux de neurones



réseau *récurrent* déroulé → LSTM (Hochreiter, 1997)

LSTM



$$f_t = \sigma(W_f \times x_t + U_f \times c_{t-1} + b_f)$$

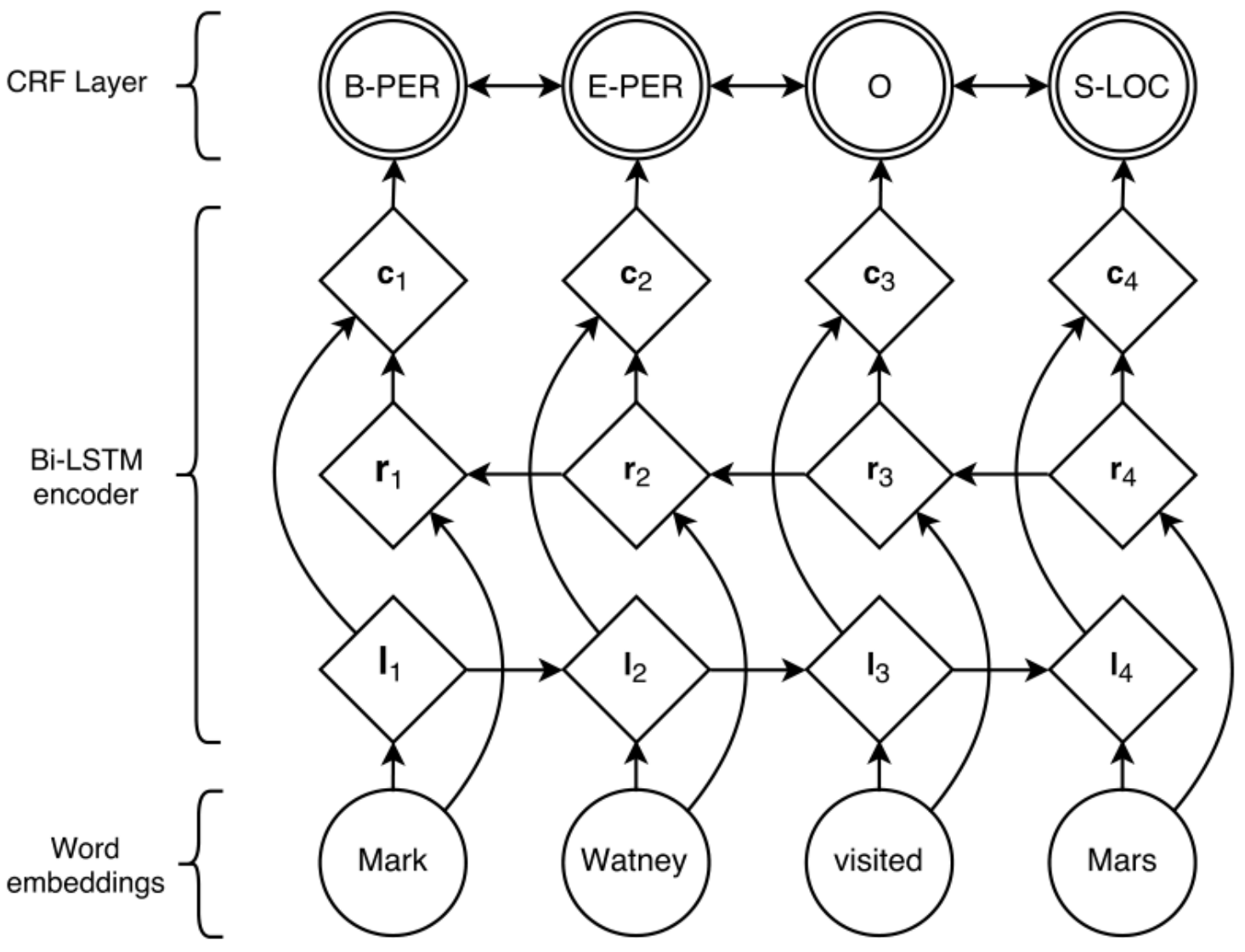
$$i_t = \sigma(W_i \times x_t + U_i \times c_{t-1} + b_i)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c \times x_t + b_c)$$

$$o_t = \sigma(W_o \times x_t + U_o \times c_{t-1} + b_o)$$

$$h_t = o_t \odot \sigma(c_t)$$

Bi-LSTM-CRF (Lample et al., 2016)



Configuration

- base $\rightarrow$ Bi-LSTM-CRF sans information supplémentaire

Configuration

- base $\rightarrow$ Bi-LSTM-CRF sans information supplémentaire
- $+$ traits de Raymond & Fayolle (2010)

Configuration

- base $\rightarrow$ Bi-LSTM-CRF sans information supplémentaire
- $+$ traits de Raymond & Fayolle (2010)
- représentations précalculées
  - corpus FrWac (Baroni et al. 2009) $\rightarrow$ 1.6 milliards mots
  - représentations calculées avec word2vec (Mikolov et al. 2013)
  - représentations prises de : `http://fauconnier.github.io`

Configuration

- base $\rightarrow$ Bi-LSTM-CRF sans information supplémentaire
- $+$ traits de Raymond & Fayolle (2010)
- représentations précalculées
  - corpus FrWac (Baroni et al. 2009) $\rightarrow$ 1.6 milliards mots
  - représentations calculées avec word2vec (Mikolov et al. 2013)
  - représentations prises de : `http://fauconnier.github.io`
- propagation des annotations

Résultats

Système	P	R	F
CRF *étalon*	85.89	76.88	81.13
CRF final	**87.89**	82.34	85.02
LSTM-CRF (*base*)	84.53	78.33	81.31
+ FrWac	86.30	81.14	83.64
+ traits Raymond & Fayolle (2010)	88.16	82.59	85.29
+ propagation	87.23	**83.96**	**85.57**
SEM (Dupont & Tellier, 2014)	86.38	80.30	83.23

Résultats

Système	P	R	F
CRF *étalon*	85.89	76.88	81.13
CRF final	**87.89**	82.34	85.02
LSTM-CRF (*base*)	84.53	78.33	81.31
+ FrWac	86.30	81.14	83.64
+ traits Raymond & Fayolle (2010)	88.16	82.59	85.29
+ propagation	87.23	**83.96**	**85.57**
SEM (Dupont & Tellier, 2014)	86.38	80.30	83.23

p-value > 0.1

Conclusion et perspectives

Conclusion et perspectives

Conclusions:

- traits construits graduellement, pour gains finaux importants
- implémentation disponible[1]
- CRF et NN $\rightarrow$ F-mesures similaires
  - CRF $\rightarrow$ précision
  - NN $\rightarrow$ rappel

[1] logiciel SEM: `https://github.com/YoannDupont/SEM`

Conclusion et perspectives

Conclusions:

- traits construits graduellement, pour gains finaux importants
- implémentation disponible[1]
- CRF et NN $\rightarrow$ F-mesures similaires
  - CRF $\rightarrow$ précision
  - NN $\rightarrow$ rappel

Perspectives:

- enrichissement lexiques + typologie plus fine
- active learning pour annoter de nouvelles données
- intégrer dépendences syntaxiques (Jie & al., 2017)
- autres langues / corpus
- Label-Dependencies aware RNN (Dupont et al. 2017)

[1] logiciel SEM: `https://github.com/YoannDupont/SEM`

Bibliographie


Abeillé A., Clément L. & Toussenel F. (2003).
Building a treebank for french.
In A. Abeillé, Ed., *Treebanks*. Dordrecht: Kluwer.


Dupont Y., Dinarelli M. & Tellier I. (2017).
Label-dependencies aware recurrent neural networks.
In *CICling 2017*.


Jie Z., Muis A. O. & Lu W. (2017).
Efficient dependency-guided named entity recognition.
In *Thirty-First AAAI Conference on Artificial Intelligence*.


Lafferty J., McCallum A. & Pereira F. (2001).
Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
In *Proceedings of ICML 2001*, p. 282–289.


Lample G., Ballesteros M., Subramanian S., Kawakami K. & Dyer C. (2016).
Neural architectures for named entity recognition.
In *Proceedings of NAACL-HLT 2016*.


Raymond C. & Fayolle J. (2010).
Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement.
In *TALN'10*.


Sagot B., Richard M. & Stern R. (2012).
Annotation référentielle du corpus arboré de paris 7 en entités nommées.
In *Traitement Automatique des Langues Naturelles (TALN)*, volume 2.

Qualité : ordre de priorité des lexiques

Expérience	Précision	Rappel	F-mesure
$P > C\&O > L$	85.98	76.79	81.13
$P > L > C\&O$	85.58	76.96	81.04
$L > P > C\&O$	85.47	77.89	81.50
$L > C\&O > P$	85.80	78.49	81.98
$C\&O > P > L$	85.66	76.36	80.74
$C\&O > L > P$	**86.77**	**78.92**	**82.66**
Ambigüe	85.39	76.79	80.86

Table 1: P : Person, L : Location, C&O : Company&Organization. En gras sont marqués les meilleurs scores pour la colonne.

Qualité : par entité

Système (précision)	company	location	organisation	person
CRF *article*+propagation	79.54	**93.86**	**89.54**	**88.15**
LSTM-CRF (FrWac)+ propagation	**83.88**	93.82	82.64	87.04

Système (rappel)	company	location	organisation	person
CRF *article*+propagation	80.07	**90.68**	69.93	90.29
LSTM-CRF (FrWac)+ propagation	**84.72**	90.11	**71.57**	**91.26**

Système (f-mesure)	company	location	organisation	person
CRF *article*+propagation	79.8	**92.24**	**78.53**	**89.21**
LSTM-CRF (FrWac)+ propagation	**84.3**	91.93	76.71	89.1

Qualité : connu vs inconnu

Système	Connues			Inconnues		
	P	R	F	P	R	F
CRF *étalon*	95.04	92.34	93.67	68.68	53.53	60.17
CRF *article*	**97.21**	93.90	95.53	72.63	59.20	65.17
+propagation	96.83	**95.46**	**96.14**	72.46	62.53	67.13
LSTM-CRF (*base*)	96.10	94.33	95.20	64.21	54.18	58.77
+ traits	95.95	94.04	94.99	70.13	59.13	64.27
LSTM-CRF (FrWac)	96.25	94.61	95.42	69.44	60.81	64.84
+ traits	96.11	94.61	95.35	**74.50**	64.45	69.12
+ propagation	95.98	94.89	95.44	73.09	**67.45**	**70.16**

Qualité : répartition des erreurs

mesure	CRF+propagation	LSTM-CRF (FrWac)+propagation
type	19.5%	21%
frontière	11.6%	13%
type+frontière	4.4%	5.0%
bruit	17.5%	21%
silence	47%	39.5%

Lexique des verbes de Dubois & Dubois-Charlier (1997)

- 25 610 entrées pour 12 310 verbes différents
-
-

Lexique des verbes de Dubois & Dubois-Charlier (1997)

- 25 610 entrées pour 12 310 verbes différents
- 14 classes génériques
-

N	munir, démunir
P	verbes psychologiques
R	réalisation, mise en état
S	saisir, serrer, posséder
T	transformation, changement
U	union, réunion
X	verbes auxiliaires

Lexique des verbes de Dubois & Dubois-Charlier (1997)

- 25 610 entrées pour 12 310 verbes différents
- 14 classes génériques
- 54 classes sémantico-syntaxiques

classes E, F, H, L, M, N, R, S, T, U :
1 : humain ou animal propre.
2 : humain figuré.
3 : non-animé propre.
4 : non-animé figuré.

classe C (communication)
1 : humain, animal (*crier, parler*).
2 : humain (*dire qc*).
3 : humain (*montrer*).
4 : figuré.

classe D (donner)
1 : humain.
2 : non-humain propre.
3 : non-humain figuré.

classe X (auxiliaires)
1 : auxiliaires temporels ou aspectuels.
2 : impersonnels.
3 : synonymes de *être* + temps, lieu.
4 : *finir* et *commencer*.

classe P (psychologique)
1 : sujet humain.
2 : objet humain.
3 : objet humain ou non-animé.