Coreference Resolution

Marco Dinarelli

Laboratoire d'Informatique de Grenoble (LIG), Getalp
Chargé de recherche (CRCN) CNRS
marco.dinarelli@univ-grenoble-alpes.fr

November 22nd 2022

Introduction : coreference resolution

Example

- *um and* ***[I]***$_1$ *think that is what's*
- *Go ahead* ***[Linda]***$_2$.

...

- *Well and uh thanks goes to* ***[you]***$_1$ *and to* ***[the media]***$_3$ *to help* ***[us]***$_4$, *so* ***[our]***$_4$ *hat is off to all of* ***[you]***$_5$ *as well.*

*Exemple de (Wiseman et al., 2016)

Coreference resolution (CR) vs. Anaphora resolution (AR) 1/3

$AR \subset CR$???

- There are people thinking that $AR \subset CR$
- *"Every speaker has to present his paper"*
  - **"his"** needs **"every speaker"** to be understood
  - **"his"** and **"every speaker"** are not coreferent
    Otherwise :
    "Every speaker had to present every speaker's paper"

*Exemples de (Sukthanker et al., 2018)

Coreference resolution (CR) vs. Anaphora resolution (AR) 2/3

$CR \subset AR$???

- There are also people thinking that $CR \subset AR$
- *"If he is unhappy with your work, the CEO will fire you"*
  - **"he"** and **"CEO"** are coreferent
  - **"he"** appears before **"CEO"** (*cataphore*)

*Exemples de (Sukthanker et al., 2018)

Coreference resolution (CR) vs. Anaphora resolution (AR) 3/3

In order to be clear

- Coreference : implies that two mentions refer (clearly) to the same entity
- Anaphore : a mention needs an antecedent in order to be understandable
  –> there is not necessarily coreference

*Exemples de (Sukthanker et al., 2018)

Coreference types 1/2

- **Zero anaphora**
  "You always have **[two fears]** : **[your commitment]** versus **[your fear]**"
- **One anaphora**
  "Since Samantha has set her eyes on **[the beautiful villa by the beach]**, she just wants to buy **[that one]**"
- **Demonstratives**
  "**[This car]** is much more spacious and classy than **[that]**"
- **Presuppositions**
  "If there is **[anyone]** who can break the spell, it is **[you]**"

*Exemples de (Sukthanker et al., 2018)

Coreference types 2/2

- **Split anaphora**
  "**[Kathrine]** and **[Maggie]** love reading. **[They]** really read all the time."
- **Contextual disambiguation**
  "The carpenter built a **[laminate]** and the dentist built **[one]** too."
  –> Useful for WSD
- **Pronominal anaphora** (3 sous-types)
  "She had seen **[the car]** which had met with an accident. **[It]** was an old white ambassador."
- **Cataphore**
  ""“If **[he]** is unhappy with your work, **[the CEO]** will fire you"""

*Exemples de (Sukthanker et al., 2018)

Non-anaphoric pronouns

- **Clefts**
  "**[It]** was Tabby who drank the milk."
- **Pleonastic "Ìt"**
  "**[It]**'s raining man !"

*Exemples de (Sukthanker et al., 2018)

Evaluation metrics 1/4

MUC (1995)

- "Link based"
- T : gold clusters (Truth) ; R : predicted clusters (Response)
- $Precision(T, R) = \sum_{r \in R} \frac{|r| - |partition(r,T)|}{|r| - 1}$
- $Recall(T, R) = \sum_{t \in T} \frac{|t| - |partition(t,R)|}{|t| - 1}$
- $|partition(r, T)|$: number of clusters in T having a non-empty intersection with r

Evaluation metrics 2/4

B^3 (1998)

- "Mention based"
- First computes precision and recall on mentions in every cluster, then computes a weighted sum from these values :
  $FinalPrecision = \sum_{i=1}^{N} w_i \cdot \frac{|R_{m_i} \cap T_{m_i}|}{|R_{m_i}|}$
  $FinalRecall = \sum_{i=1}^{N} w_i \cdot \frac{|R_{m_i} \cap T_{m_i}|}{|T_{m_i}|}$

Evaluation metrics 3/4

CEAF (*Constrained Entity Alignement F-masure*, 2005)

- "Optimal mapping based"
- Perform an optimal mapping m between R and T with a similarity measure ϕ :
  - 4 different ϕ are defined (CEAF$_{\phi_i}$)
  - the most used : $\phi_4(T, R) = 2\frac{R \cap T}{|R| + |T|}$
- $CEAF_{\phi_i} Precision(T, R) = max_m \frac{\sum_{r \in R} \phi_i(r, m(r))}{\sum_{r \in R} \phi_i(r, r)}$
- $CEAF_{\phi_i} Recall(T, R) = max_m \frac{\sum_{r \in R} \phi_i(r, m(r))}{\sum_{t \in T} \phi_i(t, t)}$

Evaluation metrics 4/4

Blanc (2014)

- "Link based"
- Used sets :
  C_T : Gold coreference clusters
  C_R : Predicted coreference clusters
  N_T : Gold non-coreferent mentions
  N_R : Predicted non-coreferent mentions
- Computed metrics :

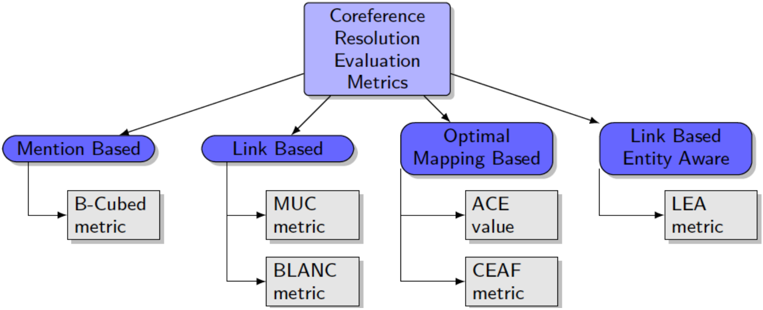
$$R_c = \frac{|C_T \cap C_R|}{|C_T|} \ , \ P_c = \frac{|C_T \cap C_R|}{|C_R|}$$

$$R_n = \frac{|N_T \cap N_R|}{|N_T|} \ , \ P_n = \frac{|N_T \cap N_R|}{|N_R|}$$

- Final metrics :

$$Recall = \frac{R_c + R_n}{2} \ , \ Precision = \frac{P_c + P_n}{2}$$

Evaluation metrics overview



(Sukthanker et al., 2018)

Data for coreference resolution

Today there are "enough" (?) :

- French : ANCOR, Democrat
- English : MUC 6 et 7, Semeval 2011 et 2012
- Several other languages : (Nedoluzhko et al., 2022) *CorefUD 1.0 : Coreference Meets Universal Dependencies*

Semeval 2012 corpus

- Version 5 Ontonotes corpus (Pradhan et al., 2012) –> News data
- In 3 languages (English the most used)
- Annotationtype : coreferences (no non-coreferent anaphora) –> Annotation of singletons
- The most used corpus

General approach to CR

Example

- Sentence 2

1. (Eastern Airlines)$_{a2}$ executives notified (union)$_{e1}$ leaders that the carrier wishes to discuss selective ((wage)$_{c2}$ reductions)$_{d2}$ on (Feb. 3)$_{b2}$.
2. ((Eastern Airlines)$_{5}$ executives)$_{6}$ notified ((union)$_{7}$ leaders)$_{8}$ that (the carrier)$_{9}$ wishes to discuss (selective (wage)$_{10}$ reductions)$_{11}$ on (Feb. 3)$_{12}$.

- Sentence 3

1. ((Union)$_{e2}$ representatives who could be reached)$_{f1}$ said (they)$_{f2}$ hadn't decided whether (they)$_{f3}$ would respond.
2. ((Union)$_{13}$ representatives)$_{14}$ who could be reached said (they)$_{15}$ hadn't decided whether (they)$_{16}$ would respond.

2 steps strategy :

1. Mention detection
2. Clustering of corefering mentions

Scientific articles overview

1. (Soon et al., 2001)
2. (Ng and Cardie, 2002)
3. (Fernandez et al., 2012)
4. (Durrett and Klein, 2013)
5. (Clark and Manning, 2015)
6. (Wiseman et al., 2016)
7. (Lee et al., 2017)

Approche *(Soon et al., 2001)* (1)

Article : *A Machine Learning Approach to Coreference Resolution of Noun Phrases*.
Auteurs : Soon, Ng et Lim

- Premier système (complètement) par apprentissage artificiel
- Représentations des mentions par des vecteurs de caractéristiques discrètes

Approche *(Soon et al., 2001)* (2)

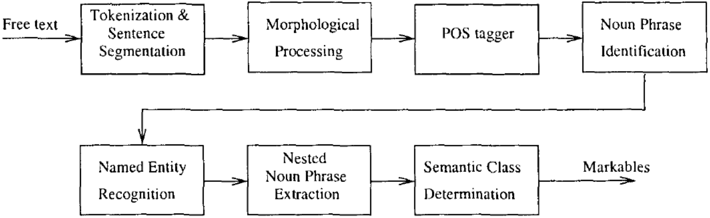


Figure – Processing pipeline (`Soon et al., 2001`)

- Phase 1 : repérage des mentions (*markables*)
- bout-en-bout (! ! !)
- 85% des mentions repérées

Approche *(Soon et al., 2001)* (3)

- Phase 2 : détection des mentions coréférentes
- Vecteurs de caractéristiques discrètes

Feature vector of the markable pair (i = *Frank Newman*, j = *vice chairman*).

Feature	**Value**	**Comments**
DIST	0	i and j are in the same sentence
I_PRONOUN	–	i is not a pronoun
J_PRONOUN	–	j is not a pronoun
STR_MATCH	–	i and j do not match
DEF_NP	–	j is not a definite noun phrase
DEM_NP	–	j is not a demonstrative noun phrase
NUMBER	+	i and j are both singular
SEMCLASS	1	i and j are both persons (This feature has three values: false(0), true(1), unknown(2).)
GENDER	1	i and j are both males (This feature has three values: false(0), true(1), unknown(2).)
PROPER_NAME	–	Only i is a proper name
ALIAS	–	j is not an alias of i
APPOSITIVE	+	j is in apposition to i

Figure – Exemple de caractéristiques (Soon et al., 2001)

Approche *(Soon et al., 2001)* (4)

Training instance generation

Étant donnée :

- une chaîne de coréférences $\boldsymbol{A} = A_1, A_2, A_3, A_4$

- une chaîne hypothétique $\boldsymbol{B}$ dans le même document

- d'autres mentions non-coréférentes a, b, ...

Supposant a, b, B_1 apparaissent par exemple entre A_1 et A_2

- **Exemples positifs** : (A_1, A_2) (A_2, A_3) (A_3, A_4)
- **Exemples négatifs** : (a, A_2) (b, A_2) (B_1, A_2) ...

Approche *(Soon et al., 2001)* (5)

Example

- Sentence 2

1. (Eastern Airlines)$_{a2}$ executives notified (union)$_{e1}$ leaders that the carrier wishes to discuss selective ((wage)$_{c2}$ reductions)$_{d2}$ on (Feb. 3)$_{b2}$.
2. ((Eastern Airlines)$_5$ executives)$_6$ notified ((union)$_7$ leaders)$_8$ that (the carrier)$_9$ wishes to discuss (selective (wage)$_{10}$ reductions)$_{11}$ on (Feb. 3)$_{12}$.

- Sentence 3

1. ((Union)$_{e2}$ representatives who could be reached)$_{f1}$ said (they)$_{f2}$ hadn't decided whether (they)$_{f3}$ would respond.
2. ((Union)$_{13}$ representatives)$_{14}$ who could be reached said (they)$_{15}$ hadn't decided whether (they)$_{16}$ would respond.

Exemples générés pour la chaîne *e* :

- **Positifs :** *(union$_7$, union$_{13}$)*
- **Négatifs :** *(the carrier$_9$, union$_{13}$) (wage$_{10}$, union$_{13}$) (selective wage reductions$_{11}$, union$_{13}$) (Feb. 3$_{12}$, union$_{13}$)*

Approche *(Soon et al., 2001)* (6)

Algorithme d'apprentissage :
arbres de décision (C5 (`Quinlan 1993`))

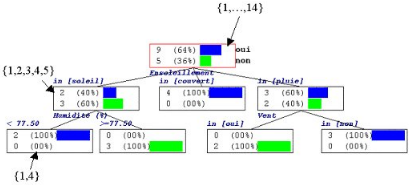


Figure – Exemple d'arbre de décision (*Wikipedia*)

Approche *(Soon et al., 2001)* (7)

Évaluation

- Données : *MUC-6* et *MUC-7* (articles de *news*)
  20910 exemples (6,5% positifs) et 48872 exemples (4,4%), respectivement
- Résultats :
  - MUC-6 : P=67.3, R=58.6, F1=62.6
  - MUC-7 : P=65.5, R=56.1, F1=60.4

Approche *(Ng and Cardie, 2002)* (1)

Article : *Improving Machine Learning Approaches to Coreference Resolution*
Auteurs : Ng et Cardie
Extension de l'approche `(Soon et al., 2001)` :

- Arbres de décision : C4.5 (vs. C5)
- Plus de caractéristiques (53 vs. 12)
- Meilleur algorithme de clusterisation
- Meilleure génération d'exemples positifs

Approche *(Ng and Cardie, 2002)* (2)

Caractéristiques

Feature Type	Feature	Description
Lexical	SOON_STR	C if, after discarding determiners, the string denoting NP_i matches that of NP_j; else I.
Grammatical	PRONOUN_1*	Y if NP_i is a pronoun; else N.
	PRONOUN_2*	Y if NP_j is a pronoun; else N.
	DEFINITE_2	Y if NP_j starts with the word "the;" else N.
	DEMONSTRATIVE_2	Y if NP_j starts with a demonstrative such as "this," "that," "these," or "those;" else N.
	NUMBER*	C if the NP pair agree in number; I if they disagree; NA if number information for one or both NPs cannot be determined.
	GENDER*	C if the NP pair agree in gender; I if they disagree; NA if gender information for one or both NPs cannot be determined.
	BOTH_PROPER_NOUNS*	C if both NPs are proper names; NA if exactly one NP is a proper name; else I.
	APPOSITIVE*	C if the NPs are in an appositive relationship; else I.
Semantic	WNCLASS*	C if the NPs have the same WordNet semantic class; I if they don't; NA if the semantic class information for one or both NPs cannot be determined.
	ALIAS*	C if one NP is an alias of the other; else I.
Positional	SENTNUM*	Distance between the NPs in terms of the number of sentences.

Approche *(Ng and Cardie, 2002)* (3)

- Algorithme de clusterisation *best first*
- Génération d'exemples en distinguant groupes nominaux et pronoms

Approche *(Ng and Cardie, 2002)* (4)

Résultats

	C4.5						RIPPER					
	MUC-6			MUC-7			MUC-6			MUC-7		
System Variation	R	P	F	R	P	F	R	P	F	R	P	F
Original Soon et al.	58.6	67.3	62.6	56.1	65.5	60.4	-	-	-	-	-	-
Duplicated Soon Baseline	62.4	70.7	66.3	55.2	68.5	61.2	60.8	68.4	64.3	54.0	69.5	60.8
Learning Framework	62.4	73.5	67.5	56.3	71.5	63.0	60.8	75.3	67.2	55.3	73.8	**63.2**
String Match	60.4	74.4	66.7	54.3	72.1	62.0	58.5	74.9	65.7	48.9	73.2	58.6
Training Instance Selection	61.9	70.3	65.8	55.2	68.3	61.1	61.3	70.4	65.5	54.2	68.8	60.6
Clustering	62.4	70.8	66.3	56.5	69.6	62.3	60.5	68.4	64.2	55.6	70.7	62.2
All Features	70.3	58.3	63.8	65.5	58.2	61.6	67.0	62.2	64.5	61.9	60.6	61.2
Pronouns only	–	66.3	–	–	62.1	–	–	71.3	–	–	62.0	–
Proper Nouns only	–	84.2	–	–	77.7	–	–	85.5	–	–	75.9	–
Common Nouns only	–	40.1	–	–	45.2	–	–	43.7	–	–	48.0	–
Hand-selected Features	64.1	74.9	**69.1**	57.4	70.8	**63.4**	64.2	78.0	**70.4**	55.7	72.8	63.1
Pronouns only	–	67.4	–	–	54.4	–	–	77.0	–	–	60.8	–
Proper Nouns only	–	93.3	–	–	86.6	–	–	95.2	–	–	88.7	–
Common Nouns only	–	63.0	–	–	64.8	–	–	62.8	–	–	63.5	–

Résultats de (`Soon et al., 2001`) :

- MUC-6 : P=67.3, R=58.6, F1=62.6
- MUC-7 : P=65.5, R=56.1, F1=60.4

Approche *(Fernandes et al., 2012)* (1)

Article : *Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution*

Auteurs : Fernandes, Dos Santos et Milidiù

- Représentation des entités (ou *clusters*) avec des arbres
- Apprentissage de structures latentes avec le *perceptron* structuré
- Optimisation d'une fonction de coût incluant l'information des entités
- Déduction automatique de caractéristiques complexes par entropie
- Évaluation sur les données de la campagne *CoNLL Eval 2012*

Approche *(Fernandes et al., 2012)* (2)

Approche en 2 étapes :

1. Détection des mentions dans le texte
   ⇒ approche par analyse syntaxique (groupes nominaux et pronoms) + entités nommées
   `(dos Santos and Carvalho, 2011)`
2. *Clusterisation* des mentions
   ⇒ perceptron structuré

Approche *(Fernandes et al., 2012)* (3)

Large margin structure perceptron :

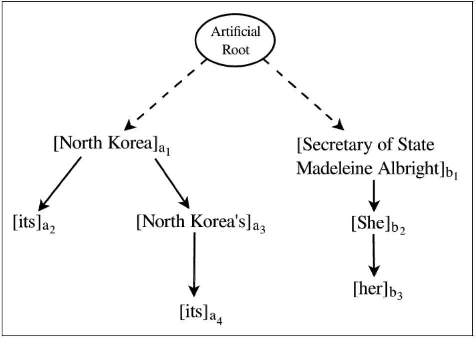
$$F^{\ell}(\boldsymbol{x}) = \arg \max_{y' \in \mathcal{Y}(\boldsymbol{x})} s(\boldsymbol{y}'; \boldsymbol{w}) + \ell(\boldsymbol{y}, \boldsymbol{y}')$$

$$s(\boldsymbol{y}'; \boldsymbol{w}) = \boldsymbol{w}^T \phi(\boldsymbol{x}, \boldsymbol{y}')$$

Approche *(Fernandes et al., 2012)* (4)

Structures latentes : arbres de coréférences

North Korea$_{a_1}$ opened **its**$_{a_2}$ doors to the U.S. today, welcoming **Secretary of State Madeleine Albright**$_{b_1}$. **She**$_{b_2}$ says **her**$_{b_3}$ visit is a good start. The U.S. remains concerned about **North Korea's**$_{a_3}$ missile development program and **its**$_{a_4}$ exports of missiles to Iran.



Approche *(Fernandes et al., 2012)* (5)

Apprentissage des structures latentes

$$F(\boldsymbol{x}) \equiv F_y(F_h(\boldsymbol{x}))$$

$\boldsymbol{w}_0 \leftarrow \boldsymbol{0}$
$t \leftarrow 0$
while no convergence
    for each $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}$
        $\tilde{\boldsymbol{h}} \leftarrow \arg\max_{h \in \mathcal{H}(\boldsymbol{x},\boldsymbol{y})} \langle \boldsymbol{w}_t, \boldsymbol{\Phi}(\boldsymbol{x}, \boldsymbol{h}) \rangle$
        $\hat{\boldsymbol{h}} \leftarrow \arg\max_{h \in \mathcal{H}(\boldsymbol{x})} \langle \boldsymbol{w}_t, \boldsymbol{\Phi}(\boldsymbol{x}, \boldsymbol{h}) \rangle + \ell_r(\boldsymbol{h}, \tilde{\boldsymbol{h}})$
        $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t + \boldsymbol{\Phi}(\boldsymbol{x}, \tilde{\boldsymbol{h}}) - \boldsymbol{\Phi}(\boldsymbol{x}, \hat{\boldsymbol{h}})$
        $t \leftarrow t + 1$
$\boldsymbol{w} \leftarrow \frac{1}{t} \sum_{i=1}^{t} \boldsymbol{w}_i$

$\mathcal{H}(x)$ feasible document trees for x
$\Phi(x, h)$ feature vector representation of x and h

Approche *(Fernandes et al., 2012)* (6)

$\phi(\boldsymbol{x}, \boldsymbol{y})$ utilise 70 caractéristiques de base de 4 types :

- Lexical
- Syntaxique
- Sémantique
- Distance et position

$+$ les caractéristiques complexes déduites automatiquement par entropie

$\Rightarrow$ e.g. 196 caractéristiques au total pour l'anglais

Approche *(Fernandes et al., 2012)* (7)

Résultats

Language	MUC			B^3			$CEAF_e$			Mean
	R	P	F_1	R	P	F_1	R	P	F_1	
Arabic	43.63	49.69	46.46	62.70	72.19	67.11	52.49	46.09	49.08	**54.22**
Chinese	52.69	70.58	60.34	62.99	80.57	70.70	53.75	37.88	44.44	**58.49**
English	65.83	75.91	70.51	65.79	77.69	71.24	55.00	43.17	48.37	**63.37**
								Official Score		**58.69**

Approche *(Durret and Klein, 2013)* (1)

Article : *Easy Victories and Uphill Battles in Coreference Resolution*

Auteurs : Durret and Klein

- Même type de modèle que le précedant (Fernandes et al., 2012) (weighted features)
- Patrons des caractéristiques extraits automatiquement (et non pas par heuristique)
- Caractéristiques assez génériques (et pas beaucoup)
- Système état-de-l'art
- Analyse très intéressante des "succès" (*easy victories*) et des erreurs (*uphill battles*)

Approche *(Durret and Klein, 2013)* (2)

- Détection de mentions : texte annoté avec analyse syntaxique et entités nommées
- 3 types de mentions :
  - pronoms
  - noms propres (entités nommées)
  - groupes nominaux (analyse syntaxique)

Approche *(Durret and Klein, 2013)* (3)

Modèle de coréférence : modèle log-linéaire

$$P(a|x) \propto \exp\left(\sum_{i=1}^{n} \mathbf{w}^\top \mathbf{f}(i, a_i, x)\right)$$

Avec :

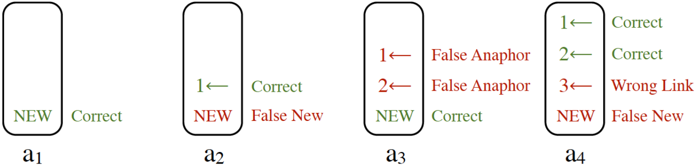
- x ensemble des mentions dans un document
- $a = (a_1, \ldots, a_N)$ un *clustering* particulier où $a_i = j$ implique que l'antécédent de la mention i est la mention j
- f *feature functions*
- w les paramètres du modèle (à apprendre)

Approche *(Durret and Klein, 2013)* (4)

Apprentissage du modèle :

$$\ell(\mathbf{w}) = \sum_{k=1}^{t} \log \left(\sum_{a \in \mathcal{A}(C_k^*)} P'(a|x_k) \right) + \lambda \|\mathbf{w}\|_1$$

$$l(a, C^*) = \alpha_{\mathrm{FA}} \mathrm{FA}(a, C^*) + \alpha_{\mathrm{FN}} \mathrm{FN}(a, C^*) + \alpha_{\mathrm{WL}} \mathrm{WL}(a, C^*)$$



[*Voters*]$_1$ *agree when* [*they*]$_1$ *are given a* [*chance*]$_2$ *to decide if* [*they*]$_1$ ...

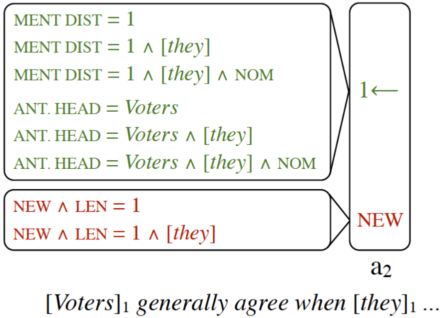
Approche *(Durret and Klein, 2013)* (5)

Caractéristiques :

Feature name	Count
Features on the current mention	
[ANAPHORIC] + [HEAD WORD]	41371
[ANAPHORIC] + [FIRST WORD]	18991
[ANAPHORIC] + [LAST WORD]	19184
[ANAPHORIC] + [PRECEDING WORD]	54605
[ANAPHORIC] + [FOLLOWING WORD]	57239
[ANAPHORIC] + [LENGTH]	4304
Features on the antecedent	
[ANTECEDENT HEAD WORD]	57383
[ANTECEDENT FIRST WORD]	24239
[ANTECEDENT LAST WORD]	23819
[ANTECEDENT PRECEDING WORD]	53421
[ANTECEDENT FOLLOWING WORD]	55718
[ANTECEDENT LENGTH]	4620
Features on the pair	
[EXACT STRING MATCH (T/F)]	47
[HEAD MATCH (T/F)]	46
[SENTENCE DISTANCE, CAPPED AT 10]	2037
[MENTION DISTANCE, CAPPED AT 10]	1680

Approche *(Durret and Klein, 2013)* (6)

Caractéristiques "conjointes" :



Approche *(Durret and Klein, 2013)* (7)

Easy victories :

	MUC	B^3	CEAF_e	Avg.
STANFORD	60.46	65.48	47.07	57.67
IMS	62.15	65.57	46.66	58.13
SURFACE	**64.39**	**66.78**	**49.00**	**60.06**

Système état-de-l'art malgré l'ensemble (restreint) de caractéristiques !

Approche *(Durret and Klein, 2013)* (8)

Analyse : même résultat avec caractéristiques automatiques et heuristiques !

	MUC	B^3	$CEAF_e$	Avg.
SURFACE	64.39	66.78	49.00	60.06
−1STWORD	63.32	66.22	47.89	59.14
+DEF−1STWORD	63.79	66.46	48.35	59.53
−PRONCONJ	59.97	63.46	47.94	57.12
+AGR−PRONCONJ	63.54	66.10	48.72	59.45
−CONTEXT	60.88	64.66	47.60	57.71
+POSN−CONTEXT	62.45	65.44	48.08	58.65
+DEF+AGR+POSN	64.55	66.93	48.94	60.14

Erreurs :

	Nominal/Proper				Pronominal	
	1st w/head		2nd+ w/head			
Singleton	99.7%	18.1K	85.5%	7.3K	66.5%	1.7K
Starts Entity	98.7%	2.1K	78.9%	0.7K	48.5%	0.3K
Anaphoric	7.9%	0.9K	75.5%	3.9K	72.0%	4.4K

Approche *(Durret and Klein, 2013)* (9)

Uphill battles : caractéristiques

- Hyperonymie et synonymie depuis *WordNet*
- Nombre et genre des mentions
- Entités nommées
- *Clusters* latents (e.g. *president*, *leader* ...)

Approche *(Durret and Klein, 2013)* (10)

Uphill battles : résultats

	MUC	B^3	CEAF$_e$	Avg.
SURFACE	64.39	66.78	49.00	60.06
SURFACE+SEM	64.70	**67.27**	**49.28**	**60.42**
SURFACE (G)	82.80	74.10	68.33	75.08
SURFACE+SEM (G)	**84.49**	**75.65**	**69.89**	**76.68**

Pour comparaison (easy victories) :

	MUC	B^3	CEAF$_e$	Avg.
STANFORD	60.46	65.48	47.07	57.67
IMS	62.15	65.57	46.66	58.13
SURFACE	**64.39**	**66.78**	**49.00**	**60.06**

Approche *(Clark and Manning, 2015)* (1)

Article : *Entity-Centric Coreference Resolution with Model Stacking*
Auteurs : Clark and Manning

- 2 modèles locaux sur paires de mentions
- + un modèle de *clustering* (reconstruction de chaînes de coréférences) incrémental
- première approche incrémentale
- Système état-de-l'art

Approche *(Clark and Manning, 2015)* (2)

2 modèles locaux sur paires de mentions :

- modèle de classification
- modèle de *ranking*

Les 2 sous forme de modèle "logistique" :

$$p_{\boldsymbol{\theta}}(a,m) = (1 + e^{\boldsymbol{\theta}^T \boldsymbol{f}(a,m)})^{-1}$$

Mêmes caractéristiques, paramètres $(\boldsymbol{\theta}_c, \boldsymbol{\theta}_r)$ et fonction de coût différents

Approche *(Clark and Manning, 2015)* (3)

Modèles locaux sur paires de mentions, fonctions de coût :

- classificateur

$$\mathcal{L}_c(\boldsymbol{\theta}_c) = -\sum_{m \in \mathcal{M}} \Big(\sum_{t \in \mathcal{T}(m)} \log p_{\boldsymbol{\theta}_c}(t, m) + \sum_{f \in \mathcal{F}(m)} \log(1 - p_{\boldsymbol{\theta}_c}(f, m)) \Big) + \lambda ||\boldsymbol{\theta}_c||_1$$

- modèle de *ranking*

$$\mathcal{L}_r(\boldsymbol{\theta}_r) = -\sum_{m \in \mathcal{M}} \Big(\max_{t \in \mathcal{T}(m)} \log p_{\boldsymbol{\theta}_r}(t, m) + \min_{f \in \mathcal{F}(m)} \log(1 - p_{\boldsymbol{\theta}_r}(f, m)) \Big) + \lambda ||\boldsymbol{\theta}_r||_1$$

$\mathcal{M}$ ensemble de toutes les mentions
$\mathcal{T}(m)$ mentions coréférentes avec m
$\mathcal{F}(m)$ mentions non coréférentes avec m

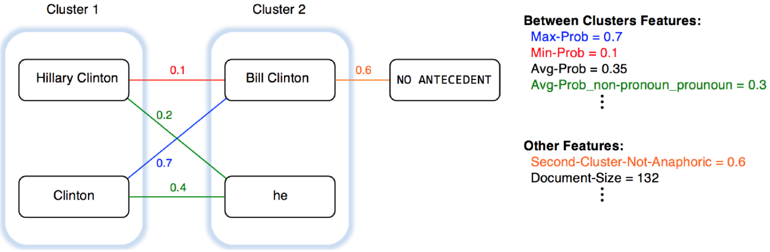
Approche *(Clark and Manning, 2015)* (4)

Caractéristiques :

- Distance
- Syntaxiques
- Sémantiques
- À base de règles
- Lexicales
- Caractéristiques conjointes (Durrett and Klein, 2013)

Approche *(Clark and Manning, 2015)* (5)

Modèle de *clustering* (*Entity-Centric*), exemple :



Approche *(Clark and Manning, 2015)* (6)

Résultats : comparaison avec le modèle *best first*

	MUC	B^3	$CEAF_{\phi_4}$	Avg.
Classification, B.F.	72.00	60.01	55.63	62.55
Ranking, B.F.	71.91	60.63	56.38	62.97
Classification, E.C.	72.34	61.46	57.16	63.65
Ranking, E.C.	72.37	61.34	57.13	63.61
Both, E.C.	72.52	62.02	57.69	64.08

Approche *(Clark and Manning, 2015)* (7)

Résultats : comparaison avec l'état-de-l'art

	MUC			B^3			$CEAF_{\phi_4}$			CoNLL
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Avg. F_1
Fernandes et al.	75.91	65.83	70.51	65.19	51.55	57.58	57.28	50.82	53.86	60.65
Chang et al.	-	-	69.48	-	-	57.44	-	-	53.07	60.00
Björkelund & Kuhn	74.3	67.46	70.72	62.71	54.96	58.58	59.4	52.27	55.61	61.63
Ma et al.	81.03	66.16	**72.84**	66.90	51.10	57.94	68.75	44.34	53.91	61.56
Durrett & Klein (INDEP.)	72.27	69.30	70.75	60.92	55.73	58.21	55.33	54.14	54.73	61.23
Durrett & Klein (JOINT)	72.61	69.91	71.24	61.18	56.43	58.71	56.17	54.23	55.18	61.71
This work	76.12	69.38	72.59	65.64	56.01	**60.44**	59.44	52.98	**56.02**	**63.02**

Approche *(Wiseman et al., 2016)* (1)

Article : *Learning Global Features for Coreference Resolution*
Auteurs : Wiseman, Rush et Shieber

- Modèle local : *mention ranking* ...
- ... mais guidé par une information globale :
  une représentation (vectorielle) des clusters !
- première approche dans son genre
- approche état-de-l'art (évidemment !)

Approche *(Wiseman et al., 2016)* (2)

Motivations de la méthode :

DA: um and [I]$_1$ think that is what's - Go ahead [Linda]$_2$.

LW: Well and uh thanks goes to [you]$_1$ and to [the media]$_3$ to help [us]$_4$...So [our]$_4$ hat is off to all of [you]$_5$ as well.

Approche *(Wiseman et al., 2016)* (3)

Modèle :

$$\underset{y_1,\ldots,y_N}{\arg\max} \sum_{n=1}^{N} f(x_n, y_n) + g(x_n, y_n, \boldsymbol{z}_{1:n-1})$$

Avec :
- $f(x_n, y_n)$ modèle local *mention ranking*
- $g(x_n, y_n, z_{1:n-1})$ modèle global avec *clustering* partiel $z_{1:n-1}$

On note :

- $\mathcal{Y}(x_n)$ les antécédents de x_n, $\mathcal{Y}(x_n) = \{1, \ldots, n-1, \epsilon\}$
- $(X^{(m)})_1^M$ ensemble de M clusters
- $\boldsymbol{z} \in \{1, \ldots, M\}^N$, $z_n = m \Rightarrow x_n \in X^{(m)}$
- $X_j^{(m)}$ est la j-eme mention dans le cluster $X^{(m)}$

Approche *(Wiseman et al., 2016)* (4)

Calcul des représentation des mentions (pour les clusters) :

$$\boldsymbol{h}_c(x_n) \triangleq \tanh(\boldsymbol{W}_c \boldsymbol{\phi}_a(x_n) + \boldsymbol{b}_c)$$

Avec :
- $\phi_a(x_n)$ vecteur creux ($\{0, 1\}^F$) représentant des caractéristiques discrètes
- $\boldsymbol{W}_c, \boldsymbol{b}_c$ paramètres (à apprendre)

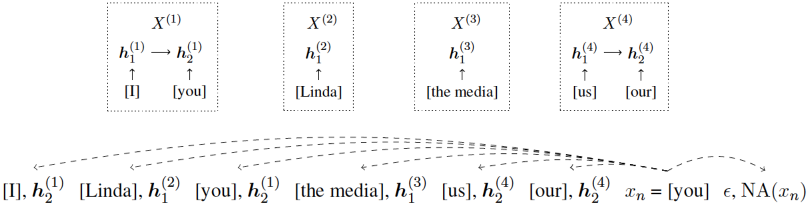
Approche *(Wiseman et al., 2016)* (5)

Calcul des représentation des clusters :

$$\boldsymbol{h}_j^{(m)} \leftarrow \mathbf{RNN}(\boldsymbol{h}_c(X_j^{(m)}), \boldsymbol{h}_{j-1}^{(m)}; \boldsymbol{\theta})$$

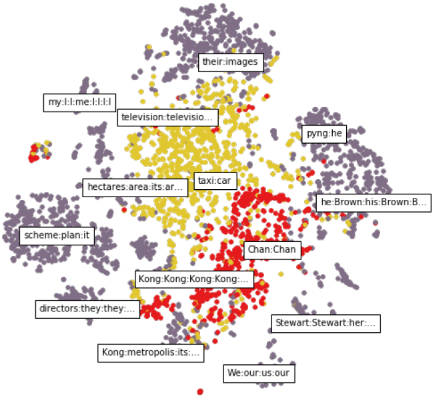
DA: um and [I]$_1$ think that is what's - Go ahead [Linda]$_2$.

LW: Well and thanks goes to [you]$_1$ and to [the media]$_3$ to help [us]$_4$...So [our]$_4$ hat is off to all of [you]$_5$...



Approche *(Wiseman et al., 2016)* (6)

Calcul des représentation des clusters :



Approche *(Wiseman et al., 2016)* (7)

Modèle local (*mention ranking*) $f(x_n, y)$

$$f(x_n, y) \triangleq \begin{cases} \boldsymbol{u}^\mathsf{T} \begin{bmatrix} \boldsymbol{h}_\mathrm{a}(x_n) \\ \boldsymbol{h}_\mathrm{p}(x_n, y) \end{bmatrix} + u_0 & \text{if } y \neq \epsilon \\ \boldsymbol{v}^\mathsf{T} \boldsymbol{h}_\mathrm{a}(x_n) + v_0 & \text{if } y = \epsilon \end{cases}$$

$$\boldsymbol{h}_\mathrm{a}(x_n) \triangleq \tanh(\boldsymbol{W}_\mathrm{a}\, \boldsymbol{\phi}_\mathrm{a}(x_n) + \boldsymbol{b}_\mathrm{a})$$
$$\boldsymbol{h}_\mathrm{p}(x_n, y) \triangleq \tanh(\boldsymbol{W}_\mathrm{p}\, \boldsymbol{\phi}_\mathrm{p}(x_n, y) + \boldsymbol{b}_\mathrm{p})$$

Approche *(Wiseman et al., 2016)* (8)

Modèle global $g(x_n, y, \boldsymbol{z}_{1:n-1})$:

$$g(x_n, y, \boldsymbol{z}_{1:n-1}) \triangleq \begin{cases} \boldsymbol{h}_c(x_n)^\top \boldsymbol{h}_{<n}^{(z_y)} & \text{if } y \neq \epsilon \\ \mathrm{NA}(x_n) & \text{if } y = \epsilon \end{cases}$$

$$\mathrm{NA}(x_n) = \boldsymbol{q}^\top \tanh\left(\boldsymbol{W}_s \begin{bmatrix} \boldsymbol{\phi}_a(x_n) \\ \sum_{m=1}^{M} \boldsymbol{h}_{<n}^{(m)} \end{bmatrix} + \boldsymbol{b}_s\right)$$

Approche *(Wiseman et al., 2016)* (9)

Apprentissage :

$$\sum_{n=1}^{N} \max_{\hat{y} \in \mathcal{Y}(x_n)} \Delta(x_n, \hat{y})(1 + f(x_n, \hat{y}) + g(x_n, \hat{y}, \boldsymbol{z}^{(o)}) - f(x_n, y_n^{\ell}) - g(x_n, y_n^{\ell}, \boldsymbol{z}^{(o)})),$$

$$y_n^{\ell} \triangleq \operatorname*{arg\,max}_{y \in \mathcal{Y}(x_n): z_y^{(o)} = z_n^{(o)}} f(x_n, y) + g(x_n, y, \boldsymbol{z}^{(o)})$$

Approche *(Wiseman et al., 2016)* (10)

Apprentissage :

Algorithm 1 Greedy search with global RNNs

procedure GREEDYCLUSTER($x_1, \ldots, x_N$)
  Initialize clusters $X^{(1)} \ldots$ as empty lists, hidden states $\boldsymbol{h}^{(0)}, \ldots$ as $\mathbf{0}$ vectors in $\mathbb{R}^D$, $\boldsymbol{z}$ as map from mention to cluster, and cluster counter $M \leftarrow 0$
  for $n = 2 \ldots N$ **do**
    $y^* \leftarrow \arg\max_{y \in \mathcal{Y}(x_n)} f(x_n, y) + g(x_n, y, \boldsymbol{z}_{1:n-1})$
    $m \leftarrow z_{y^*}$
    if $y^* = \epsilon$ **then**
      $M \leftarrow M + 1$
      $m \leftarrow M$
    append x_n to $X^{(m)}$
    $z_n \leftarrow m$
    $\boldsymbol{h}^{(m)} \leftarrow \mathbf{RNN}(\boldsymbol{h}_c(x_n), \boldsymbol{h}^{(m)})$
  return $X^{(1)}, \ldots, X^{(M)}$

Approche *(Wiseman et al., 2016)* (11)

Résultats :

System	MUC			B^3			$CEAF_e$			
	P	R	F_1	P	R	F_1	P	R	F_1	CoNLL
B&K (2014)	74.3	67.46	70.72	62.71	54.96	58.58	59.4	52.27	55.61	61.63
M&S (2015)	76.72	68.13	72.17	66.12	54.22	59.58	59.47	52.33	55.67	62.47
C&M (2015)	76.12	69.38	72.59	65.64	56.01	60.44	59.44	52.98	56.02	63.02
Peng et al. (2015)	-	-	72.22	-	-	60.50	-	-	56.37	63.03
Wiseman et al. (2015)	76.23	69.31	72.60	66.07	55.83	60.52	59.41	54.88	57.05	63.39
This work	77.49	69.75	**73.42**	66.83	56.95	**61.50**	62.14	53.85	**57.70**	**64.21**

Approche "Kenton Lee"

Article : *End-to-end Neural Coreference Resolution*
Auteurs : Lee, He, Lewis et Zettlemoyer

Caractéristiques

- Premier système neuronal bout-à-bout (*end-to-end*)
- Il se passe de l'annotation *gold* des mentions
- Il résolue implicitement le problème des enchâssements

Approche “Kenton Lee” (suite...)

Le modèle

- $P(y_1, \ldots, y_N | D) = \prod_{i=1}^{N} P(y_i | D) = \prod_{i=1}^{N} \frac{exp(s(i,y_i))}{\sum_{y' \in Y(i)} exp(s(i,y'))}$
- $$s(i,j) = \begin{cases} 0 & j = \epsilon \\ s_m(i) + s_m(j) + s_a(i,j) & j \neq \epsilon \end{cases}$$
- $s_m(i) = w_m \cdot \text{FFNN}(g_i)$
- $s_a(i,j) = w_a \cdot \text{FFNN}(g_i, g_j, g_i \odot g_j, \phi(i,j))$
- g_i sont les représentations des mentions de mots
- $\phi(i,j)$ encode orateur, genre, distance entre les mentions

Approche “Kenton Lee” (suite...)

Représentations des mentions :

- x_t^* représentation cachée calculée par un LSTM bidirectionnel
- *soft (syntactic) head* :

$$\alpha_t = \boldsymbol{w}_\alpha \cdot \mathrm{FFNN}_\alpha(\boldsymbol{x}_t^*)$$

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum\limits_{k=\mathrm{START}(i)}^{\mathrm{END}(i)} \exp(\alpha_k)}$$

$$\hat{\boldsymbol{x}}_i = \sum_{t=\mathrm{START}(i)}^{\mathrm{END}(i)} a_{i,t} \cdot \boldsymbol{x}_t$$

- représentation finale :

$$\boldsymbol{g}_i = [\boldsymbol{x}^*_{\mathrm{START}(i)}, \boldsymbol{x}^*_{\mathrm{END}(i)}, \hat{\boldsymbol{x}}_i, \phi(i)]$$

Approche "Kenton Lee" (suite...)

Architecture neuronale

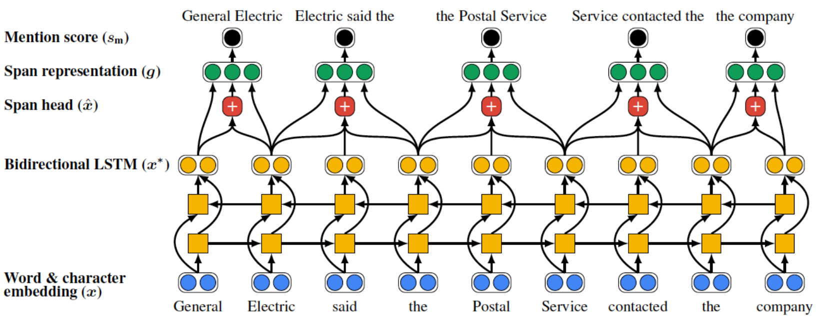


*Image de (Lee et al. 2017)

Approche "Kenton Lee" (suite...)

Problème

- Génère toutes les segmentations d'un texte de longueur T ($O(T^4)$)
- Pour résoudre ça :
  - L : longueur maximale d'un segment ("span")
  - λT : fraction des meilleurs segments gardés (scorés avec $s_m(i)$)
  - K : nombre d'antécédents pour chaque segment
  - Les segments ne peuvent pas se croiser

Apprentissage

Utilise la log-vraisemblance sur les clusters *gold*

Approche "Kenton Lee" : évaluation 1/3

Résultats globaux

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg. F1
Our model (ensemble)	**81.2**	**73.6**	**77.2**	**72.3**	**61.7**	**66.6**	**65.2**	**60.2**	**62.6**	**68.8**
Our model (single)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Durrett and Klein (2014)	72.6	69.9	71.2	61.2	56.4	58.7	56.2	54.2	55.2	61.7
Björkelund and Kuhn (2014)	74.3	67.5	70.7	62.7	55.0	58.6	59.4	52.3	55.6	61.6
Durrett and Klein (2013)	72.9	65.9	69.2	63.6	52.5	57.5	54.3	54.4	54.3	60.3

Approche “Kenton Lee” : évaluation 2/3

Ablation test

	Avg. F1	Δ
Our model (ensemble)	69.0	+1.3
Our model (single)	67.7	
– distance and width features	63.9	-3.8
– GloVe embeddings	65.3	-2.4
– speaker and genre metadata	66.3	-1.4
– head-finding attention	66.4	-1.3
– character CNN	66.8	-0.9
– Turian embeddings	66.9	-0.8

Approche “Kenton Lee” : évaluation 3/3

Attention !

1	(**A fire in a Bangladeshi garment factory**) has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee (**the blaze**) in the four-story building.
	A fire in (**a Bangladeshi garment factory**) has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in (**the four-story building**).
2	We are looking for (**a region of central Italy bordering the Adriatic Sea**). (**The area**) is mostly mountainous and includes Mt. Corno, the highest peak of the Apennines. (**It**) also includes a lot of sheep, good clean-living, healthy sheep, and an Italian entrepreneur has an idea about how to make a little money of them.
3	(**The flight attendants**) have until 6:00 today to ratify labor concessions. (**The pilots'**) union and ground crew did so yesterday.
4	(**Prince Charles and his new wife Camilla**) have jumped across the pond and are touring the United States making (**their**) first stop today in New York. It's Charles' first opportunity to showcase his new wife, but few Americans seem to care. Here's Jeanie Mowth. What a difference two decades make. (**Charles and Diana**) visited a JC Penney's on the prince's last official US tour. Twenty years later here's the prince with his new wife.
5	Also such location devices, (**some ships**) have smoke floats (**they**) can toss out so the man overboard will be able to use smoke signals as a way of trying to, let the rescuer locate (**them**).