Coreference Resolution

Marco Dinarelli

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

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.*

*Example from (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"

*Examples from (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*)

*Examples from `(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

*Examples from (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]**"

*Examples from (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**
  "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""

*Examples from (Sukthanker et al., 2018)

Non-anaphoric pronouns

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

*Examples from (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_{ϕ_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

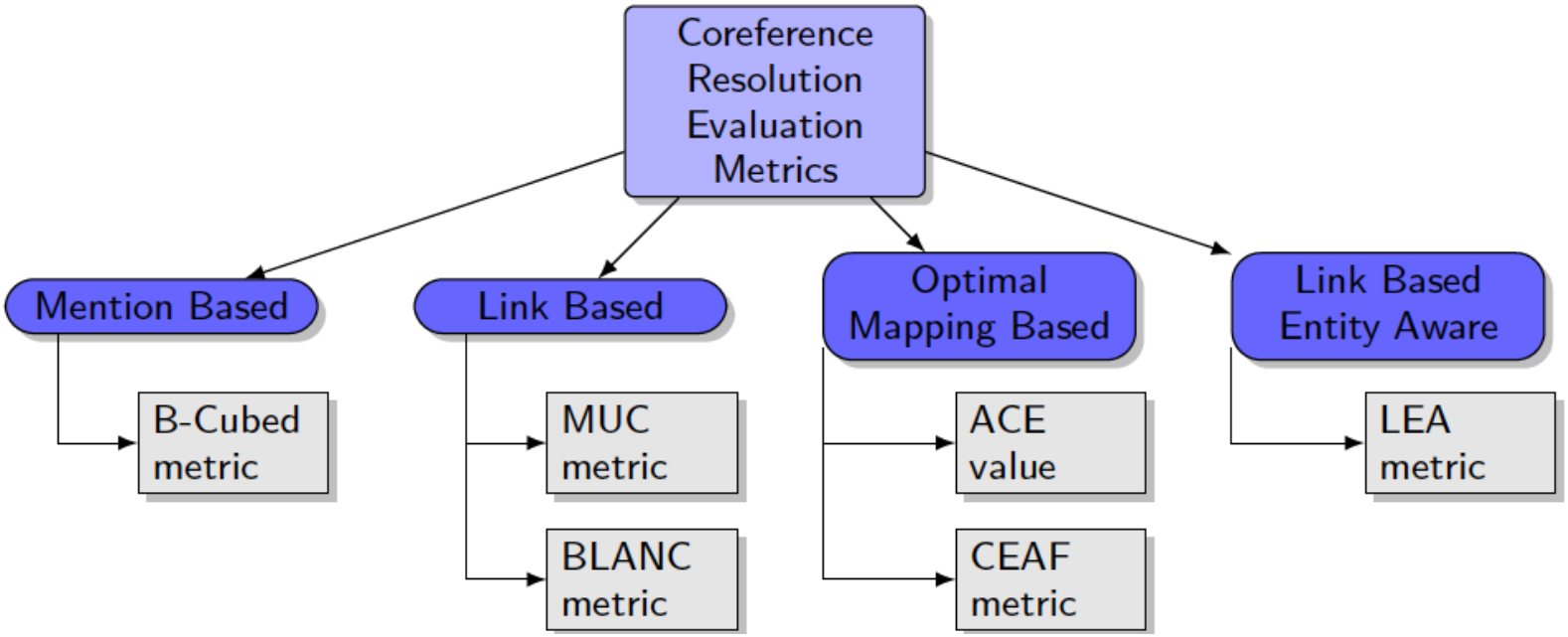
  N_R : Predicted non-coreferent mentions
- Computed metrics :

$$R_c = \frac{|C_T \cap C_R|}{|C_T|} \ , \ P_c = \frac{|C_T \cap C_R|}{|C_R|}$$
$$R_n = \frac{|N_T \cap N_R|}{|N_T|} \ , \ P_n = \frac{|N_T \cap N_R|}{|N_R|}$$

- 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

Still a relatively rare resource, much less since few years :

- 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)
- Annotation type : coreferences (no non-coreferent anaphora)
  –> No 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 (end-to-end or not) :
  - 1. Mention detection
  - 2. Clustering of corefering mentions (entity detection)
- *Neural* end-to-end : all-in-one step, or all steps at the same time
- Seq-to-seq (with LLMs)

Best (imho) 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)
   *** Transition to neural models ***
6. (Wiseman et al., 2016)
   *** Full neural models ***
7. (Lee et al., 2017)
   *** Seq-to-seq neural models (since 2021) ***
8. (Zhang et al., 2023)

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

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

- First full machine learning based system
- Mention pairs representation with discrete feature vectors

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

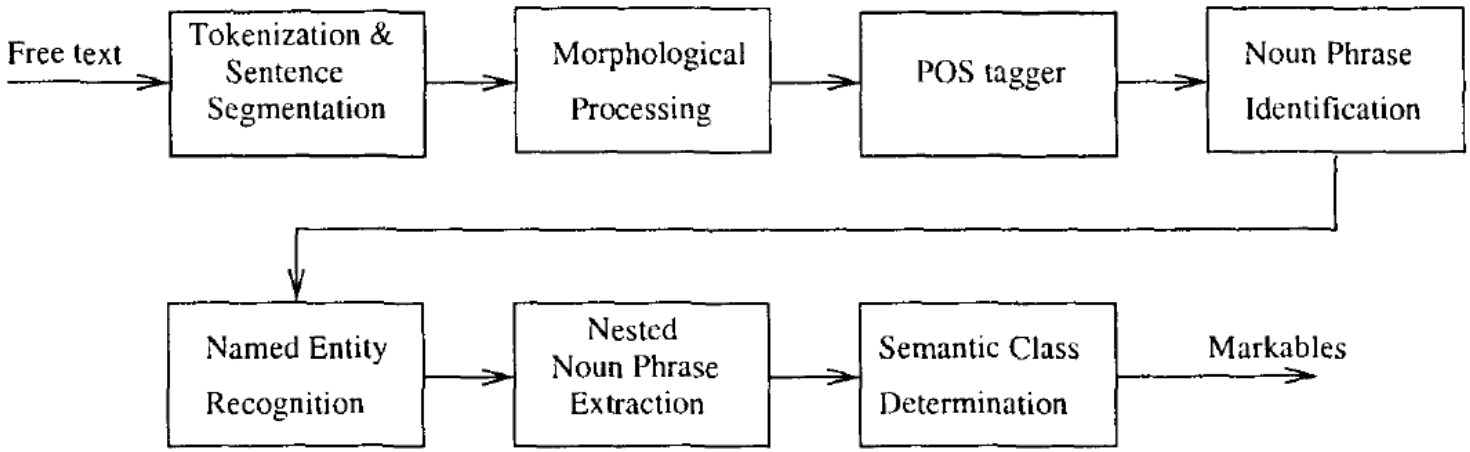


FIGURE – Processing pipeline (Soon et al., 2001)

- Step 1 : mention detection (*markables*)
- end-to-end (!!!)
- Detects 85% of mentions

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

- Step 2 : Detection of coreferent mentions
- Discrete features vectors

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 – Example of feature instantiation (Soon et al., 2001)

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

Training instance generation

Given :

- A coreference chain $\boldsymbol{A} = A_1, A_2, A_3, A_4$
- Another coreference $\boldsymbol{B}$ in the same document
- Other possibly non-coreferent mentions a, b, ...

If a, b, B_1 appears for example between A_1 and A_2

- **Positive examples** : (A_1, A_2) (A_2, A_3) (A_3, A_4)
- **Negative examples** : (a, A_2) (b, A_2) (B_1, A_2) ...

Paper *(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.

Training instances generated for the coreference chain *e* :

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

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

Training algorithm :
Decision trees (C5 (`Quinlan 1993`))

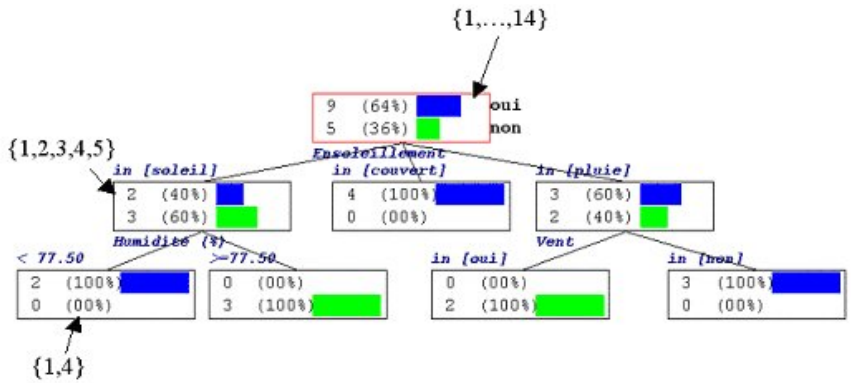


FIGURE – Example of decision tree (*Wikipedia*)

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

Evaluation

- Data : *MUC-6* et *MUC-7* (news articles)
  respectively 20910 instances (6,5% positive) and 48872 instances (4,4%)
- Results :
  - MUC-6 : P=67.3, R=58.6, F1=62.6
  - MUC-7 : P=65.5, R=56.1, F1=60.4

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

Title : *Improving Machine Learning Approaches to Coreference Resolution*
Authors : Ng et Cardie
Extension of previous approach `(Soon et al., 2001)` :

- Decision trees : C4.5 (vs. C5)
- More features (53 vs. 12)
- Better clustering strategy
- Better positive examples generation

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

Features

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.

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

- *Best first* clustering algorithm
- Positive instances generation distinguishing noun-phrase and pronominal mention

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

Results

	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	–

Results of `(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

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

Title : *Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution*
Authors : Fernandes, Dos Santos et Milidiù

- Representation of Entities (*Clusters*) with coreference trees
- Learning of latent structures (trees) with the structured *perceptron*
- Optimization of an entity-level loss function
- Entropy-based deduction of complex features
- Evaluation on the CoNLL Semeval 2012 data
  $\rightarrow$ not comparable with previous papers

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

Two-steps approach :

1. Detection of mentions in raw text
   $\Rightarrow$ based on syntactic analysis (noun phrases and pronouns) + named entities
   `(dos Santos and Carvalho, 2011)`
2. Mention clustering $\Rightarrow$ structured perceptron

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

Large margin structure perceptron :

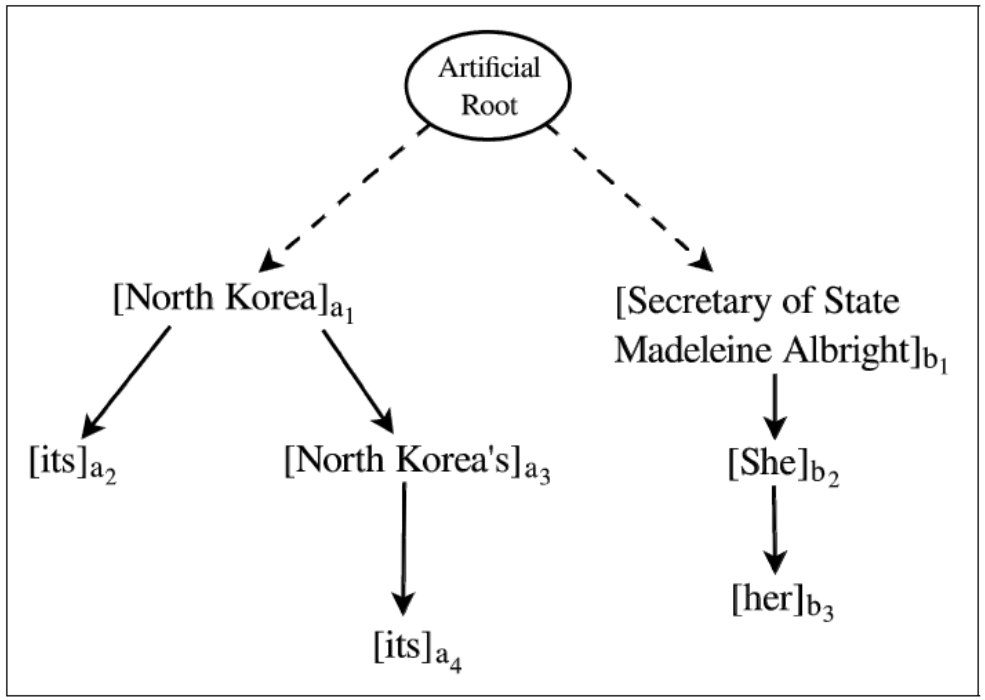
$$F^{\ell}(\boldsymbol{x}) = \arg \max_{y' \in \mathcal{Y}(\boldsymbol{x})} s(\boldsymbol{y}'; \boldsymbol{w}) + \ell(\boldsymbol{y}, \boldsymbol{y}')$$

$s(\boldsymbol{y}'; \boldsymbol{w})$ = a predictor with parameters $\boldsymbol{w}$
$L(\boldsymbol{y}, \boldsymbol{y}')$ = loss function (margin)

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

Latent structures : coreference trees

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.



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

Latent structure learning

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

$\boldsymbol{w}_0 \leftarrow \mathbf{0}$
$t \leftarrow 0$
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

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

$\phi(\boldsymbol{x}, \boldsymbol{y})$ uses 70 features from 4 categories :

- Lexical
- Syntactic
- Semantic
- Distance and position

$+$ complex features automatically induced with entropy information
$\Rightarrow$ e.g. 196 features in total for English

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

Results

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**

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

Title : *Easy Victories and Uphill Battles in Coreference Resolution*
Authors : Durret and Klein

- Same model type as previous paper (`Fernandes et al.`, 2012) (weighted features)
- Automatically extracted features (not based on domain knowledge)
- *General purpose* features (and not as many as previous paper)
- SOTA results
- Interesting analysis of "good outcomes" (*easy victories*) and errors (*uphill battles*)

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

- Mention detection : texts annotated with syntactic analysis and named entities
- 3 types of mentions :
  - pronouns (POS tags in syntactic analysis)
  - proper names (from named entities)
  - noun phrases (from syntactic analysis)

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

Coreference model : log-linear model

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

Avec :

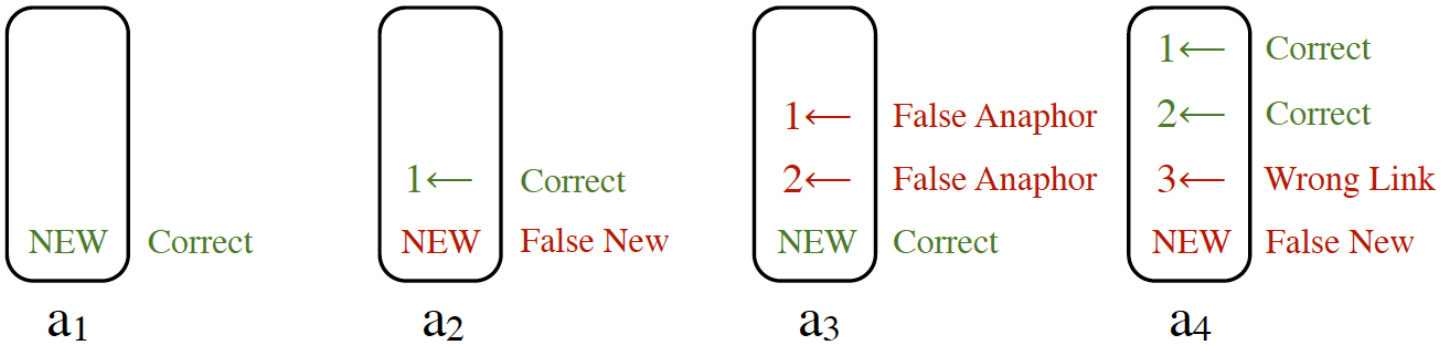
- x : surface-level document context (plus any information...)
- $a = (a_1, \ldots, a_N)$ a particular *clustering* where $a_i = j$ means the antecedent of mention i is mention j
  $a_i \in \{1, \ldots, i-1, NEW\}$
- $\boldsymbol{f}$ *feature functions*
- $\boldsymbol{w}$ model parameters

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

Model training :

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

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



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

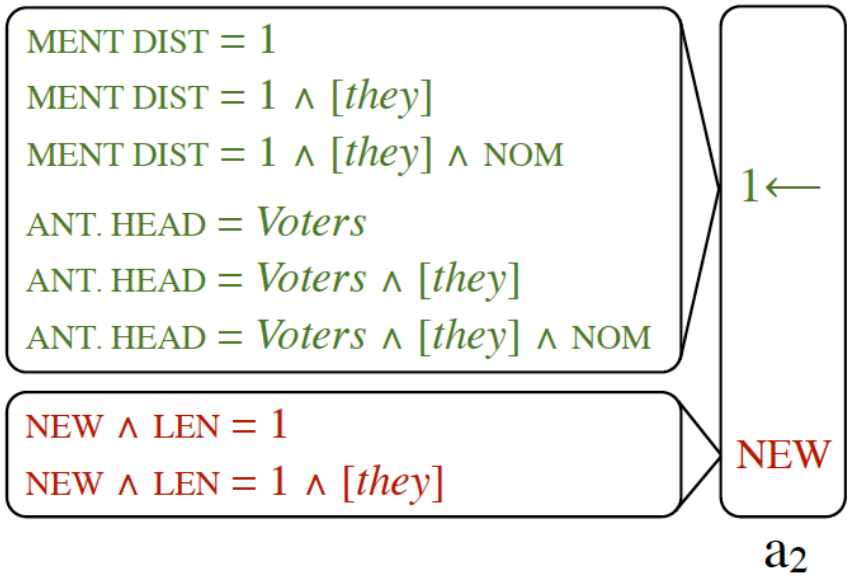
Paper *(Durret and Klein, 2013)* (5)

Features :

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

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

Joint features :



[*Voters*]$_1$ *generally agree when* [*they*]$_1$ ...

Paper *(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**

State-of-the-art results despite a relatively small set of features

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

Analysis : same results with automatic features and heuristics !

	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

Errors :

	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

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

Uphill battles : features

- Hyperonyms et synonims from *WordNet*
- Number and gender of mentions
- Named entities
- Latent *clusters* (e.g. *president*, *leader* ...)

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

Uphill battles : results

	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**

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

Title : *Entity-Centric Coreference Resolution with Model Stacking*
Authors : Clark and Manning

- 2 local models on mention pairs
- + an incremental *clustering* model (generating coreference chains)
- First incremental approach
- State-of-the-art results

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

2 local models on mention pairs :

- classification model
- ranking model

Both formalized as *logistic* models :

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

Same features, different parameters $(\boldsymbol{\theta}_c, \boldsymbol{\theta}_r)$ and loss function

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

Local models on mention pairs, loss functions :
- classifier

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

- ranking model

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

$\mathcal{M}$ (all) mention set
$\mathcal{T}(m)$ mentions coreferent with m (True)
$\mathcal{F}(m)$ mentions not coreferent with m (False)

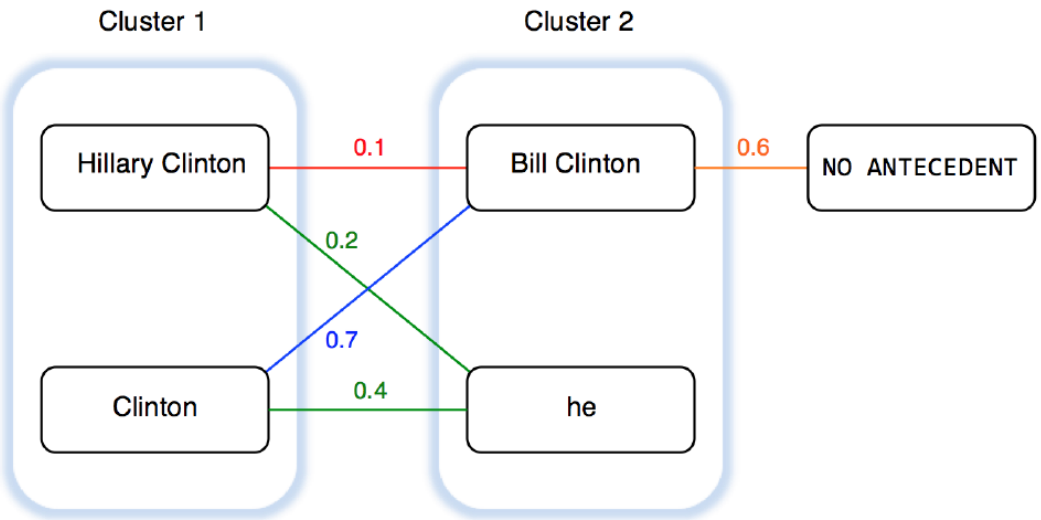
Paper *(Clark and Manning, 2015)* (4)

Features :

- Distance
- Syntactic
- Semantic
- Rule based
- Lexical
- Joint features (`Durrett and Klein, 2013`)

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

Clustering model (*Entity-Centric*), example :



Between Clusters Features:
Max-Prob = 0.7
Min-Prob = 0.1
Avg-Prob = 0.35
Avg-Prob_non-pronoun_prounoun = 0.3
⋮

Other Features:
Second-Cluster-Not-Anaphoric = 0.6
Document-Size = 132
⋮

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

Results : comparison to a *best first* strategy

	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

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

Results : comparison to the litterature

	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**

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

Title : *Learning Global Features for Coreference Resolution*
Authors : Wiseman, Rush et Shieber

- Local model : *mention ranking* ...
- ... but informed with global information : a (vector) representation of clusters !
- First approach of this type (using cluster vector representations)
- SOTA results (of course !)

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

Motivations :

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.

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

Model :

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

Where :

- $f(x_n, y_n)$ local *mention ranking* model
- $g(x_n, y_n, z_{1:n-1})$ global model with partial *clustering* $z_{1:n-1}$

We define :

- $\mathcal{Y}(x_n)$ the possible antecedents of x_n, $\mathcal{Y}(x_n) = \{1, \ldots, n-1, \epsilon\}$
- $(X^{(m)})_1^M$ set of M clusters
- $\boldsymbol{z} \in \{1, \ldots, M\}^N$, $z_n = m \Rightarrow x_n \in X^{(m)}$
- $X_j^{(m)}$ is the j-th mention in cluster $X^{(m)}$

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

Computation of mention representations (for cluster) :

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

Avec :
- $\phi_a(x_n)$ sparse vector ($\{0,1\}^F$) representing some discrete features
- $\boldsymbol{W}_c, \boldsymbol{b}_c$ parameters (to be learned)

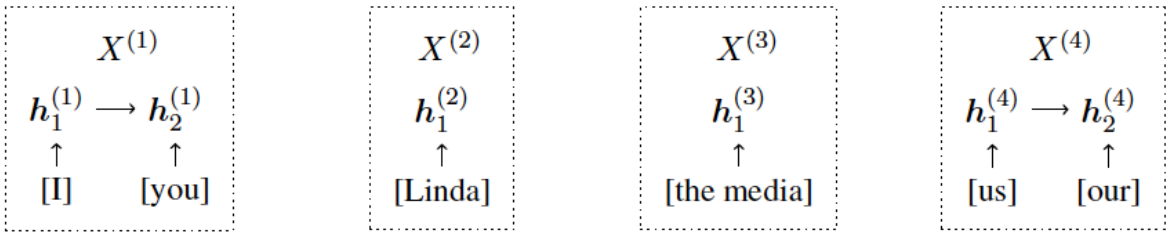
Paper *(Wiseman et al., 2016)* (5)

Computation of cluster representations :

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

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

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



[I], $\boldsymbol{h}_2^{(1)}$ [Linda], $\boldsymbol{h}_1^{(2)}$ [you], $\boldsymbol{h}_2^{(1)}$ [the media], $\boldsymbol{h}_1^{(3)}$ [us], $\boldsymbol{h}_2^{(4)}$ [our], $\boldsymbol{h}_2^{(4)}$ x_n = [you] ϵ, $\text{NA}(x_n)$

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

Visualization of cluster representations :

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

Local (*mention ranking*) model $f(x_n, y)$

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

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

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

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

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

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

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

Learning function (loss) :

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

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

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

Inference :

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 $X^{(1)}, \ldots, X^{(M)}$

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

Results :

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**

Paper "Kenton Lee" 2017

Title : *End-to-end Neural Coreference Resolution*
Authors : Lee, He, Lewis et Zettlemoyer

Caractéristiques

- First end-to-end neural system
- It does not use gold mentions
- Implicitly solve the nested mention problem

Paper "Kenton Lee" 2017 (continued...)

The model

- $P(y_1, \ldots, y_N|D) = \prod_{i=1}^{N} P(y_i|D) = \prod_{i=1}^{N} \frac{exp(s(i,y_i))}{\sum_{y' \in Y(i)} exp(s(i,y'))}$
- $$s(i,j) = \begin{cases} 0 & j = \epsilon \\ s_m(i) + s_m(j) + s_a(i,j) & j \neq \epsilon \end{cases}$$
- $s_m(i) = w_m \cdot \text{FFNN}(g_i)$
- $s_a(i,j) = w_a \cdot \text{FFNN}(g_i, g_j, g_i \odot g_j, \phi(i,j))$
- g_i representation of word mentions
- $\phi(i,j)$ encode speaker, gender, mention distance

Paper "Kenton Lee" 2017 (continued...)

Mention representation :

- x_t^* hidden state from a bidirectional LSTM
- *soft (syntactic) head* :

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

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum_{k=\text{START}(i)}^{\text{END}(i)} \exp(\alpha_k)}$$

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

- final representation :

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

Paper "Kenton Lee" 2017 (continued...)

Neural architecture

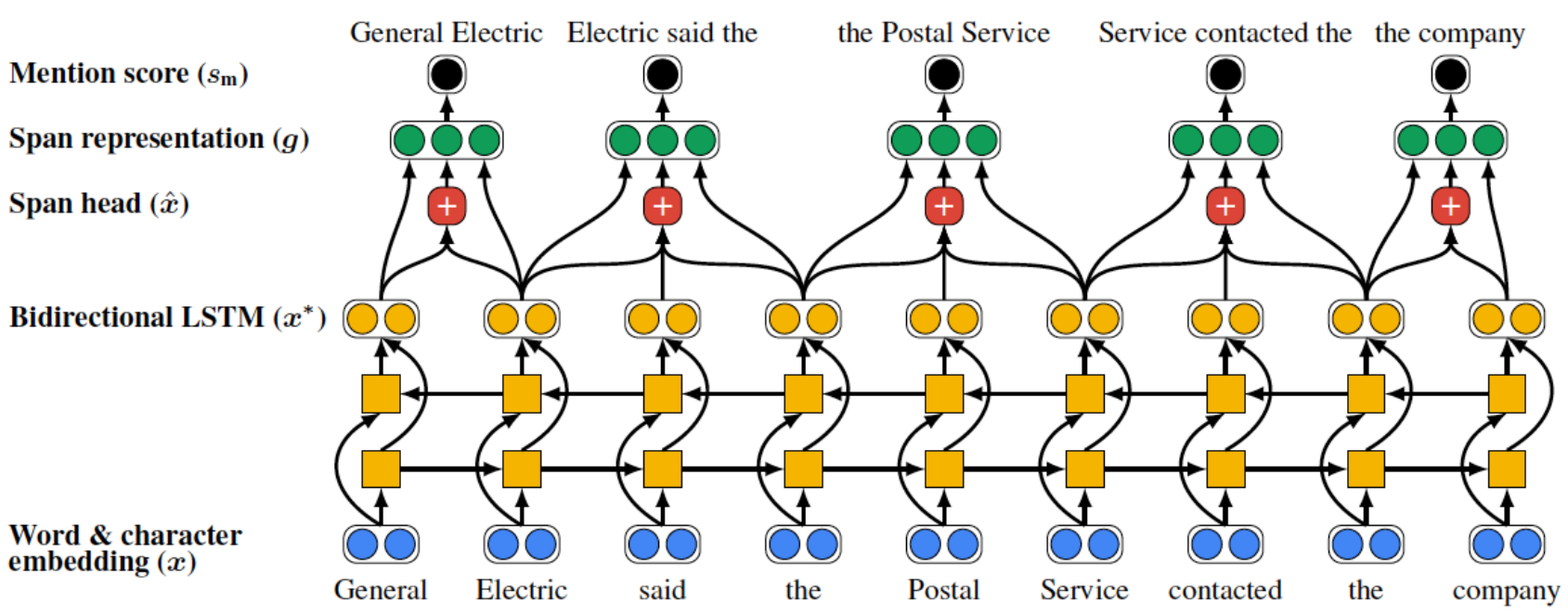


*Image de (Lee et al. 2017)

Paper "Kenton Lee" 2017 (continued...)

Problem

- Generate all possible segmentations of a text of length T ($O(T^4)$)
- In order to overcome such complexity :
  - L : maximum span length
  - λT : fraction of best mentions kept (scored with $s_m(i)$)
  - K : maximum number of antecedent for each mention
  - Segments cannot overlap (cross)

Learning function

Use the log-likelihood on the gold clusters

Paper "Kenton Lee" 2017 : evaluation 1/3

Global results

	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

Paper "Kenton Lee" 2017 : evaluation 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

Paper "Kenton Lee" 2017 : evaluation 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**).

Paper *Sequence-to-sequence* 2023

Title : *Seq2seq is All You Need for Coreference Resolution*
Authors : Wenzheng Zhang, Sam Wiseman, Karl Stratos

Key points

- Full *sequence-to-sequence* model
- It relies (heavely) on a *T5* model
  - for encoding text
  - for encoding the context ...
- No specific functionality for coreference resolution (what a pity !)
- SOTA results (or almost)

Paper *Sequence-to-sequence* 2023 (continued)

Linearization of coreference annotation

- Input : a, b, c, d, e
- Clusters : (2, 2, 1), (5, 5, 2), (2, 3, 2)
  - format : (start-token, end-token, cluster-id)
- Output : a <m> <m> b | 1 </m> c | 2 </m> d <m> e | 2 </m>
- Constrained decoding

Paper *Sequence-to-sequence* 2023 (continued)

Results

		MUC			B^3			$CEAF_{\phi_4}$			Avg.
	Model	P	R	F1	P	R	F1	P	R	F1	F1
Non-Seq2seq	Lee et al., 2017	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
	Lee et al. (2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
	Joshi et al. (2019)	84.7	82.4	83.5	76.5	74.0	75.3	74.1	69.8	71.9	76.9
	Yu et al. (2020)	82.7	83.3	83.0	73.8	75.6	74.7	72.2	71.0	71.6	76.4
	Joshi et al. (2020)	85.8	84.8	85.3	78.3	77.9	78.1	76.4	74.2	75.3	79.6
	Xia et al. (2020)	85.7	84.8	85.3	78.1	77.5	77.8	76.3	74.1	75.2	79.4
	Toshniwal et al. (2020)	85.5	85.1	85.3	78.7	77.3	78.0	74.2	76.5	75.3	79.6
	Wu et al. (2020)*	88.6	87.4	88.0	82.4	82.0	82.2	79.9	78.3	79.1	83.1
	Xu and Choi (2020)	85.9	85.5	85.7	79.0	78.9	79.0	76.7	75.2	75.9	80.2
	Kirstain et al. (2021)	86.5	85.1	85.8	80.3	77.9	79.1	76.8	75.4	76.1	80.3
	Dobrovolskii (2021)	84.9	87.9	86.3	77.4	82.6	79.9	76.1	77.1	76.6	81.0
	Toshniwal et al. (2021)	-	-	-	-	-	-	-	-	-	79.6
	Liu et al. (2022) + $T0_{3B}$	85.8	88.3	86.9	79.6	83.3	81.5	78.3	78.5	78.4	82.3
	Liu et al. (2022) + FLAN-$T5_{XXL}$	86.1	88.4	87.2	80.2	83.2	81.7	78.9	78.3	78.6	82.5
Transition Seq2seq	Bohnet et al. (2023) + $mT5_{XXL}$	87.4	88.3	87.8	81.8	83.4	82.6	79.1	79.9	79.5	83.3
Seq2seq	Paolini et al. (2021)+$T5_{base}$	-	-	81.0	-	-	69.0	-	-	68.4	72.8
	Paolini et al. (2021)+$T0^{\dagger}_{3B}$	85.0	86.0	85.2	76.1	78.5	77.3	76.5	75.6	76.0	79.6
	Partial linear + $T0_{3B}$	83.9	87.6	85.7	76.6	82.1	79.3	77.7	76.5	77.1	80.7
	Integer free + $T0_{3B}$	84.9	88.8	86.8	78.9	84.0	81.4	78.1	79.3	78.7	82.3
	Full inear + token action + $T0_{3B}$	85.9	88.6	87.2	79.6	83.5	81.5	78.9	78.0	78.5	82.4
	Full linear + copy action + $T0_{3B}$	85.8	89.0	87.4	80.0	84.3	82.1	79.1	79.4	79.3	82.9
	Full linear + copy action + $T0_{pp}$	86.1	89.2	87.6	80.6	84.3	82.4	78.9	80.1	79.5	83.2