Information extraction & automated knowledge graph construction

Lecture 3: Coreference & linking

Lauriane Aufrant (Inria Paris)

Université Paris Cité – UFRL – M2 LI – 23/11/2023

Progress

- ~~09/11 – Overview of information extraction~~
- ~~16/11 – Entities, relations...~~
- 23/11 – Coreference, linking...
- 30/11 – From IE to automated knowledge graph construction
- 07/12 – IE annotations
- 14/12 – IE in a specialty domain

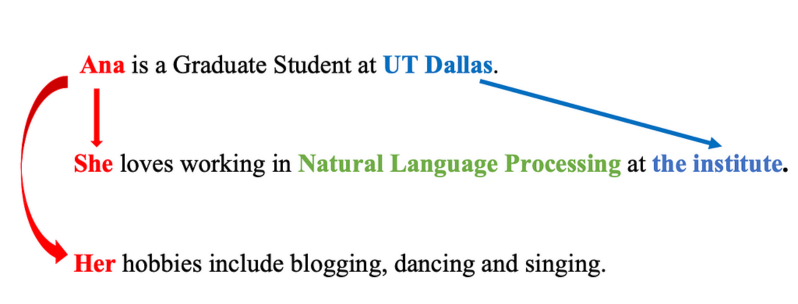
Today: Coreference & linking

- Coreference resolution

  ↪ task & variants, formalism, corpora, methods, evaluation

- Entity linking

Coreference resolution



Ana is a Graduate Student at **UT Dallas**.

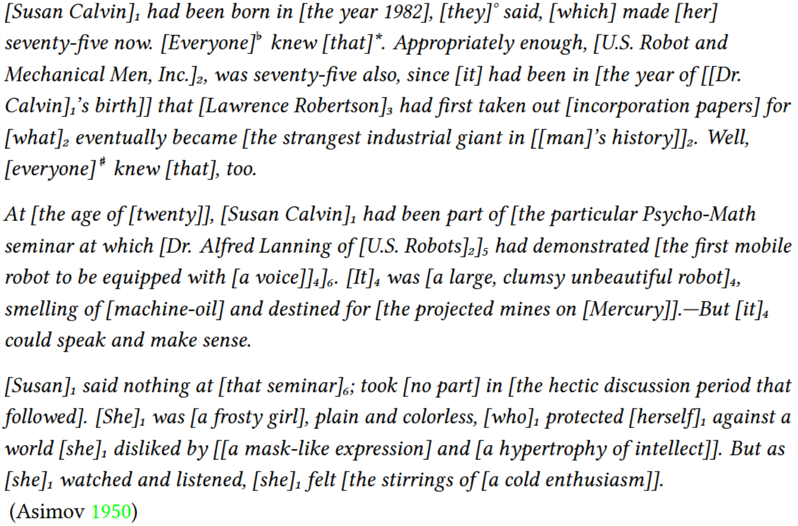
She loves working in **Natural Language Processing** at **the institute**.

Her hobbies include blogging, dancing and singing.

From Towards AI

Coreference resolution (CR)

Detecting mentions referring to the same entity (= same coreference chain)



[Susan Calvin]$_1$ had been born in [the year 1982], [they]$^\circ$ said, [which] made [her] seventy-five now. [Everyone]$^\flat$ knew [that]*. Appropriately enough, [U.S. Robot and Mechanical Men, Inc.]$_2$, was seventy-five also, since [it] had been in [the year of [[Dr. Calvin]$_1$'s birth]] that [Lawrence Robertson]$_3$ had first taken out [incorporation papers] for [what]$_2$ eventually became [the strangest industrial giant in [[man]'s history]]$_2$. Well, [everyone]$^\sharp$ knew [that], too.

At [the age of [twenty]], [Susan Calvin]$_1$ had been part of [the particular Psycho-Math seminar at which [Dr. Alfred Lanning of [U.S. Robots]$_2$]$_5$ had demonstrated [the first mobile robot to be equipped with [a voice]]$_4$]$_6$. [It]$_4$ was [a large, clumsy unbeautiful robot]$_4$, smelling of [machine-oil] and destined for [the projected mines on [Mercury]].—But [it]$_4$ could speak and make sense.

[Susan]$_1$ said nothing at [that seminar]$_6$; took [no part] in [the hectic discussion period that followed]. [She]$_1$ was [a frosty girl], plain and colorless, [who]$_1$ protected [herself]$_1$ against a world [she]$_1$ disliked by [[a mask-like expression] and [a hypertrophy of intellect]]. But as [she]$_1$ watched and listened, [she]$_1$ felt [the stirrings of [a cold enthusiasm]].

(Asimov 1950)

From Grobol, 2020

Mentions

Types of referring expressions:

- Names
- Pronouns
- Noun phrases
- ... any constituent?
- ... any part of a sentence?

More examples

- Bill saw a unicorn. **The unicorn** had a golden mane.
- If Mary had a car, she would take me to work in **it**.
- I must write it down by the way you remind me of **that**.

From Grobol, 2020

Anaphora resolution $\neq$ coreference resolution

Anaphoric expression = dependent on objects previously mentioned

- Every dancer twisted **her** knee.
- No dancer twisted **her** knee.
- We went to see a concert last night. **The tickets** were really expensive.
- She doesn't bike, though she owns **one**.

Non-anaphoric coreferences:

Emmanuel Macron is the French President since 2017. **Macron** has been elected twice.

From Manning. From Grobol, 2020.

Anaphora and cataphora

If **he** had woken up earlier, Pete wouldn't have missed the bus.

Multiple antecedent coreference

Josh saw Teresa yesterday. **They** spent at least two hours together.

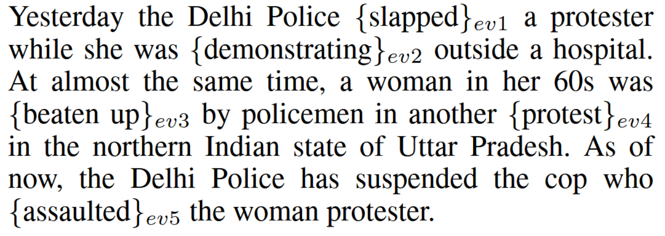
Near-identity

- Laura lives in **Croatia**. **It** has joined the EU in 2013.
- **John** learned to speak at age 2. **He** married at 25.
- Pick **a ripe, plump zucchini**. Prepare **it** for the oven, cut **it** into four pieces and roast **it** with thyme and smoked paprika for 1 hour. Serve **it** with white rice and enjoy **its** sweet and lightly spicy taste.

Event coreference

Ted arrived late. This irritated Mary.

Fred damaged a garment. He stained a shirt.



Yesterday the Delhi Police $\{\text{slapped}\}_{ev1}$ a protester while she was $\{\text{demonstrating}\}_{ev2}$ outside a hospital. At almost the same time, a woman in her 60s was $\{\text{beaten up}\}_{ev3}$ by policemen in another $\{\text{protest}\}_{ev4}$ in the northern Indian state of Uttar Pradesh. As of now, the Delhi Police has suspended the cop who $\{\text{assaulted}\}_{ev5}$ the woman protester.

From Danlos & Gaiffe, 2002. From Lu & Ng, 2021.

Formalizing coreference resolution

(Identify mentions then...)

- **mention pairs:** associate mentions with their antecedent
- **entity-centric:** affect each mention to the corresponding entity
- **mention-centric:** clustering mentions

Formalizing coreference resolution

(Identify mentions then...)

- **mention pairs:** associate mentions with their antecedent
- **entity-centric:** affect each mention to the corresponding entity
- **mention-centric:** clustering mentions

↪ Not fully equivalent: singletons, biases, instability...

Corpora for coreference resolution

- MUC-6 (1995)
- ACE 2004
- CoNLL 2012
- The Winograd Schema Challenge
- OntoNotes 5.0
- GAP (2018): Gender Ambiguous Pronouns

- ANCOR (2014): spoken French
- DEMOCRAT (2016): written French

Winograd Schema (Winograd, 1972)

- She poured water from the pitcher into the cup until **it** was full.
- She poured water from the pitcher into the cup until **it** was empty.

- The city council refused the women a permit because **they** feared violence.
- The city council refused the women a permit because **they** advocated violence.

From Manning

Winograd Schema (Winograd, 1972)

- She poured water from the pitcher into the cup until **it** was full.
- She poured water from the pitcher into the cup until **it** was empty.

- The city council refused the women a permit because **they** feared violence.
- The city council refused the women a permit because **they** advocated violence.

An alternative to the Turing test? (Levesque, 2013)

From Manning

Methods for coreference resolution

Methods for coreference resolution

- Latest mention heuristic... depending on the distance?

Methods for coreference resolution

- Latest mention heuristic... depending on the distance?
- String matching (head match, exact match, partial match...)

Methods for coreference resolution

- Latest mention heuristic... depending on the distance?
- String matching (head match, exact match, partial match...)
- Syntactic patterns

Methods for coreference resolution

- Latest mention heuristic... depending on the distance?
- String matching (head match, exact match, partial match...)
- Syntactic patterns
- Person/number/gender agreement

Methods for coreference resolution

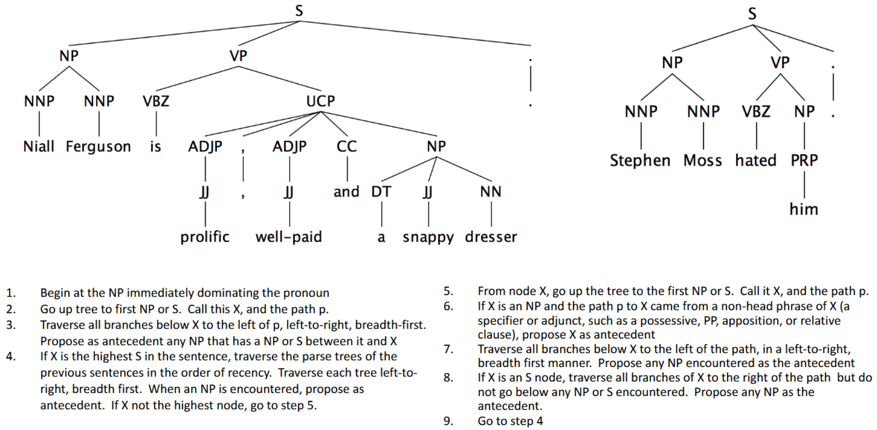
- Latest mention heuristic... depending on the distance?
- String matching (head match, exact match, partial match...)
- Syntactic patterns
- Person/number/gender agreement
- Semantic similarity

Methods for coreference resolution

- Latest mention heuristic... depending on the distance?
- String matching (head match, exact match, partial match...)
- Syntactic patterns
- Person/number/gender agreement
- Semantic similarity
- ...

↪ Relevant both for rule-based methods and feature-based ML

Hobbs' naive(?) algorithm, 1976

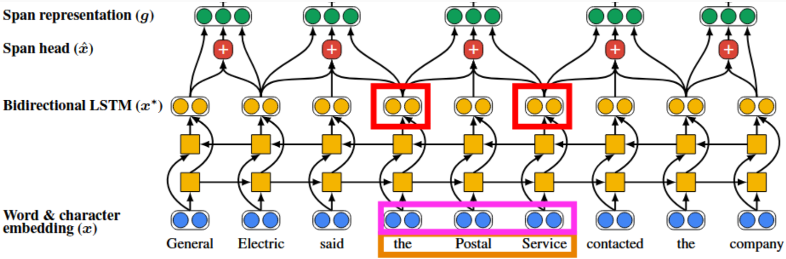


1. Begin at the NP immediately dominating the pronoun
2. Go up tree to first NP or S. Call this X, and the path p.
3. Traverse all branches below X to the left of p, left-to-right, breadth-first. Propose as antecedent any NP that has a NP or S between it and X
4. If X is the highest S in the sentence, traverse the parse trees of the previous sentences in the order of recency. Traverse each tree left-to-right, breadth first. When an NP is encountered, propose as antecedent. If X not the highest node, go to step 5.
5. From node X, go up the tree to the first NP or S. Call it X, and the path p.
6. If X is an NP and the path p to X came from a non-head phrase of X (a specifier or adjunct, such as a possessive, PP, apposition, or relative clause), propose X as antecedent
7. Traverse all branches below X to the left of the path, in a left-to-right, breadth first manner. Propose any NP encountered as the antecedent
8. If X is an S node, traverse all branches of X to the right of the path but do not go below any NP or S encountered. Propose any NP as the antecedent.
9. Go to step 4

The pipeline approach for CR

- Mention detection: based on NER, a PoS tagger, a constituency parser...
- Candidate filtering (heuristics, classifier...) – or keep all and later discard singletons
- Mention pair scoring: heuristics, classifier, similarity over neural representations...
- Clustering: usually agglomerative clustering (possibly with more heuristics)

End-to-end Neural Coreference Resolution (Lee et al., 2017)



Span representation (g)
Span head ($\hat{x}$)
Bidirectional LSTM (x^*)
Word & character embedding (x)
General
Electric
said
the
Postal
Service
contacted
the
company

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

BILSTM hidden states for span's start and end

Attention-based representation (details next slide) of the words in the span

Additional features

From Manning

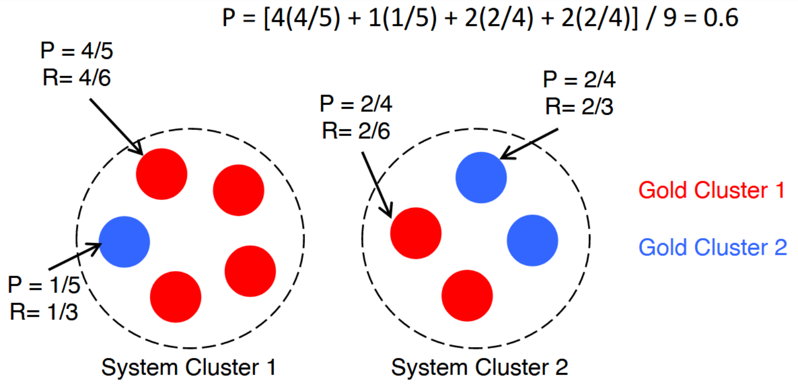
Other methods

- Reformulate antecedent search as question answering
- Word-level (represent mention pairs through their heads – with contextual embeddings): Dobrovolskii, 2021
- Shift-reduce (over Transformer-based representations): Bohnet et al., 2022

Evaluation metrics: F1, F1 and F1

- MUC: precision/recall/F1 over the links (from a mention to its closest antecedent)
- B^3: average P/R over all (true) entities represented in the cluster
- CEAF: best matching (CEAF_m if mention-centric / CEAF_e if entity-centric) of reference/predicted clusters, then P/R/F1
- MELA (=CoNLL): weighted average of F1 from MUC, B^3 and CEAF_e
- BLANC (2011): F1 over all coreferring mention pairs + non-coreferring mention pairs
- LEA (2016): average (weighted by entity size) of P/R over intra-entity mention pairs

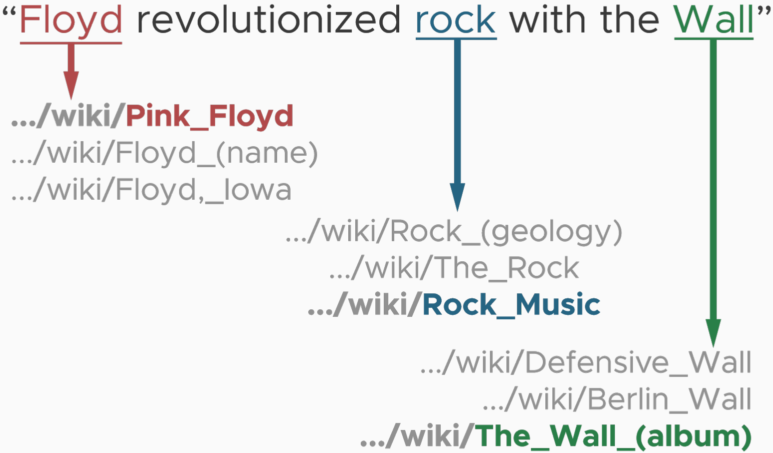
B^3



P = [4(4/5) + 1(1/5) + 2(2/4) + 2(2/4)] / 9 = 0.6

From Manning

Entity linking



"Floyd revolutionized rock with the Wall"

.../wiki/Pink_Floyd

.../wiki/Floyd_(name)

.../wiki/Floyd,_Iowa

.../wiki/Rock_(geology)

.../wiki/The_Rock

.../wiki/Rock_Music

.../wiki/Defensive_Wall

.../wiki/Berlin_Wall

.../wiki/The_Wall_(album)

Alberto Parravicini, 2019

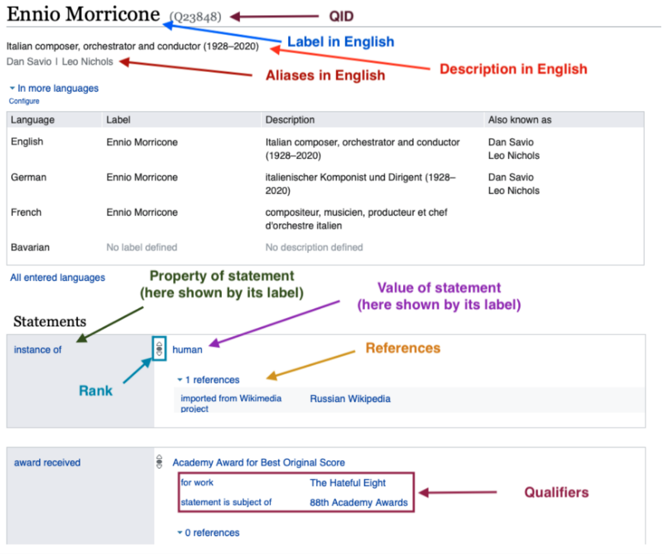
Entity linking (EL)

Mapping entity mentions in text to entities in a knowledge base

Comparison with CR

- Mentions: names (named entity linking, NEL), sometimes with nominal mentions. Not possessives.
- Similar issues: near-identity...
- Singleton $\longleftrightarrow$ unlinkable mentions

Wikidata



Ennio Morricone (Q23848)

Italian composer, orchestrator and conductor (1928–2020)

Dan Savio | Leo Nichols

▾ In more languages

Configure

Language	Label	Description	Also known as
English	Ennio Morricone	Italian composer, orchestrator and conductor (1928–2020)	Dan Savio Leo Nichols
German	Ennio Morricone	italienischer Komponist und Dirigent (1928–2020)	Dan Savio Leo Nichols
French	Ennio Morricone	compositeur, musicien, producteur et chef d'orchestre italien	
Bavarian	No label defined	No description defined	

All entered languages

Statements

instance of — human

▾ 1 references

imported from Wikimedia project — Russian Wikipedia

award received — Academy Award for Best Original Score

for work — The Hateful Eight

statement is subject of — 88th Academy Awards

▾ 0 references

QID

Label in English

Description in English

Aliases in English

Property of statement (here shown by its label)

Value of statement (here shown by its label)

References

Rank

Qualifiers

From Möller et al., 2021

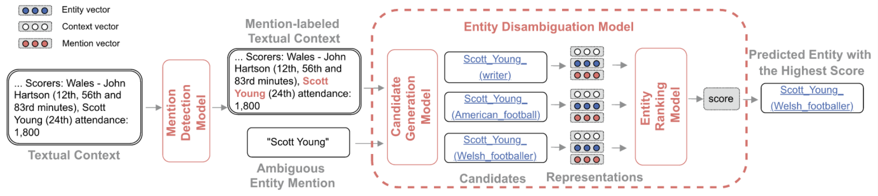
Corpora for entity linking

- AIDA-CoNLL (extension of CoNLL 2003, 27k mentions)
- MSNBC (English, no unlinkable)
- AQUAINT (English, no unlinkable)
- ACE (extension of ACE 2004)
- TAC KBP (Knowledge Base Population) Entity Linking (English / Chinese / Spanish, news)
- TAC KBP Entity Discovery and Linking (English / Chinese / Spanish, news, 50k mentions)
- LORELEI language packs (multilingual, low-resourced languages)
- WikiAnn (282 languages, automatic labelling)

Methods for entity linking

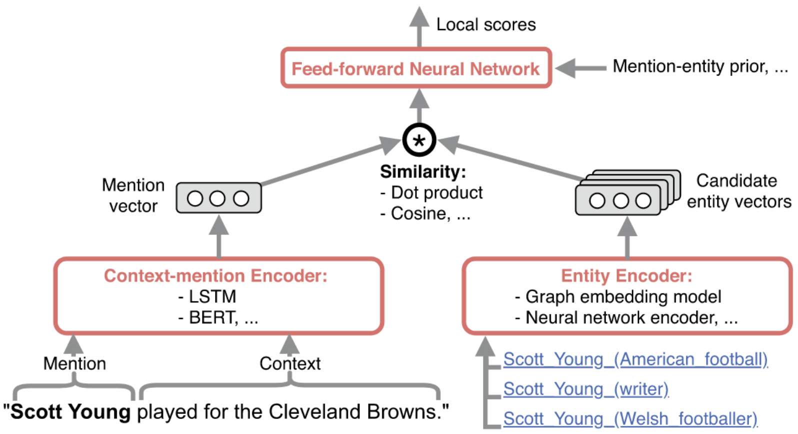
- String matching or similarity
- Popularity-based heuristics (number of inlinks)
- Knowledge matching: facts in sentence ↔ facts in knowledge base
- Feature-based ML to classify unlinkable mentions

The pipeline approach for EL


Entity vector
Context vector
Mention vector
... Scorers: Wales - John Hartson (12th, 56th and 83rd minutes), Scott Young (24th) attendance: 1,800
Textual Context
Mention Detection Model
Mention-labeled Textual Context
... Scorers: Wales - John Hartson (12th, 56th and 83rd minutes), Scott Young (24th) attendance: 1,800
"Scott Young"
Ambiguous Entity Mention
Entity Disambiguation Model
Candidate Generation Model
Scott_Young_(writer)
Scott_Young_(American_football)
Scott_Young_(Welsh_footballer)
Candidates
Representations
Entity Ranking Model
score
Predicted Entity with the Highest Score
Scott_Young_(Welsh_footballer)

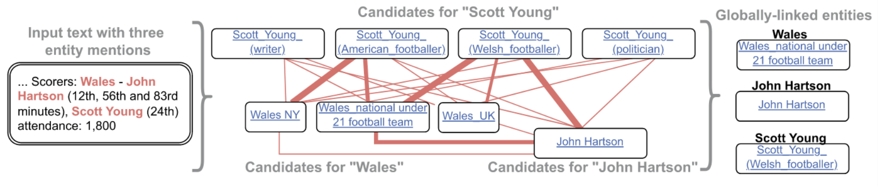
From Sevgili et al., 2022

Entity ranking



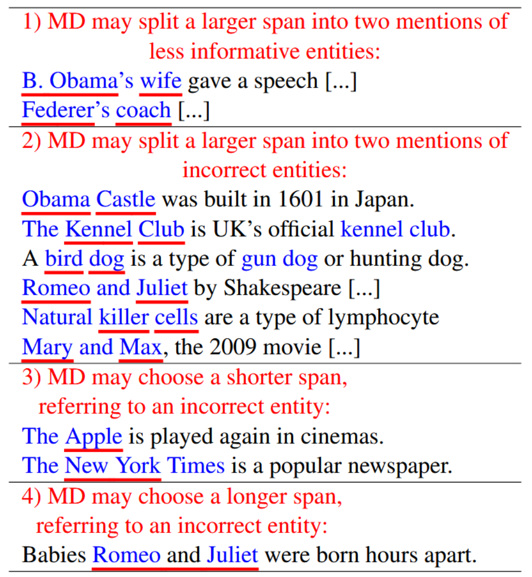
From Sevgili et al., 2022

Global (collective) entity linking


Input text with three entity mentions
... Scorers: Wales - John Hartson (12th, 56th and 83rd minutes), Scott Young (24th) attendance: 1,800
Candidates for "Scott Young"
Scott_Young_(writer)
Scott_Young_(American_football)
Scott_Young_(Welsh_footballer)
Scott_Young_(politician)
Wales NY
Wales_national under 21 football team
Wales_UK
John Hartson
Candidates for "Wales"
Candidates for "John Hartson"
Globally-linked entities
Wales
Wales_national under 21 football team
John Hartson
John Hartson
Scott Young
Scott_Young_(Welsh_footballer)

From Sevgili et al., 2022

End-to-End Neural Entity Linking (Kolitsas et al., 2018)



1) MD may split a larger span into two mentions of less informative entities:

B. Obama's wife gave a speech [...]

Federer's coach [...]

2) MD may split a larger span into two mentions of incorrect entities:

Obama Castle was built in 1601 in Japan.

The Kennel Club is UK's official kennel club.

A bird dog is a type of gun dog or hunting dog.

Romeo and Juliet by Shakespeare [...]

Natural killer cells are a type of lymphocyte

Mary and Max, the 2009 movie [...]

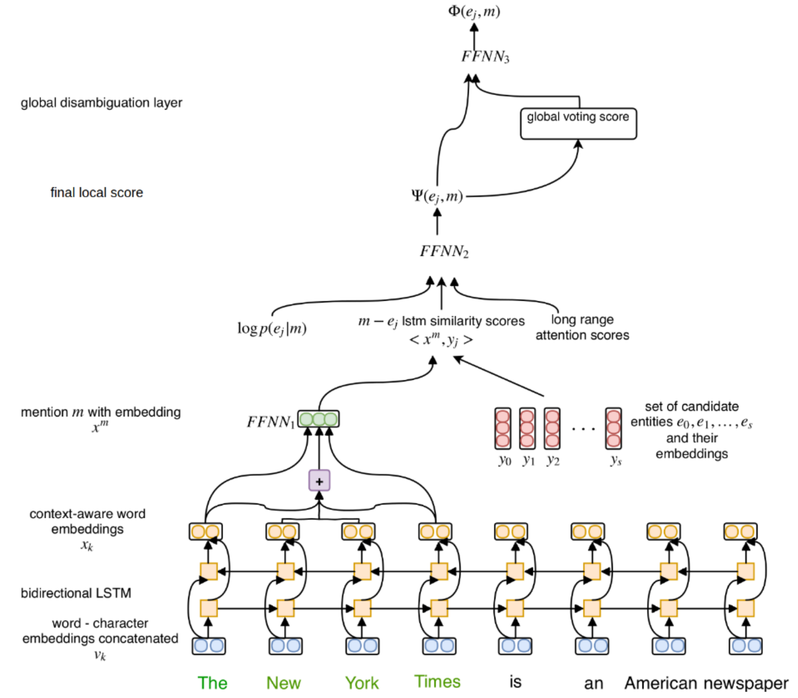
3) MD may choose a shorter span, referring to an incorrect entity:

The Apple is played again in cinemas.

The New York Times is a popular newspaper.

4) MD may choose a longer span, referring to an incorrect entity:

Babies Romeo and Juliet were born hours apart.


$\Phi(e_j, m)$
$FFNN_3$
global disambiguation layer
global voting score
final local score
$\Psi(e_j, m)$
$FFNN_2$
$\log p(e_j|m)$
$m - e_j$ lstm similarity scores
$< x^m, y_j >$
long range attention scores
mention m with embedding x^m
$FFNN_1$
+
set of candidate entities $e_0, e_1, \ldots, e_s$ and their embeddings
y_0
y_1
y_2
. . .
y_s
context-aware word embeddings x_k
bidirectional LSTM
word - character embeddings concatenated v_k
The
New
York
Times
is
an
American newspaper

See you next week!

`first.last@inria.fr`