Deep learning for automatic coreference detection

Loïc Grobol (Lattice / ALMAnaCH)

Inria junior seminar

2018-04-17

Coreferences and how to find them

Anaphora

Sam Vimes sighed when he heard the scream.

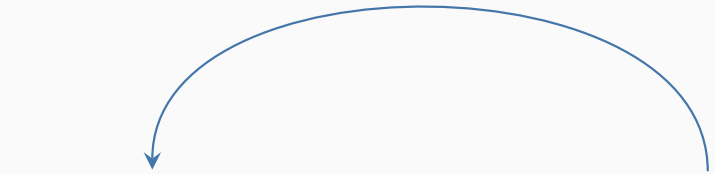
Figure 1: Anaphora

Anaphora

Sam Vimes sighed when he heard the scream.

Figure 1: Anaphora

Anaphora



Sam Vimes sighed when he heard the scream.

Figure 1: Anaphora

Coreference

The Eyjafjallajökull volcano, one of Iceland's largest had been dormant for nearly two centuries before returning gently to life in the late evening of March 20, 2010, noticeable at first by the emergence of a red cloud glowing above the vast glacier that covers it. In the following days, fire fountains jetted from a dozen vents on the volcano, reaching as high as 100 meters

Figure 2: Coreference

Discourse model

Bill saw a unicorn. The unicorn had a gold mane.

Figure 3: Referring to inexistent objects

Discourse model

If I had a hammer I would use it to break your head.

Figure 3: Referring to inexistent objects

Discourse model

A four-sided triangle, now that would be a sight.

Figure 3: Referring to inexistent objects

Applications

- Information extraction

  The French Institute for Research in Computer Science and Automation is a French national research institution focusing on computer science and applied mathematics. It was created under the name Institut de recherche en informatique et en automatique (IRIA) in 1967

- Translation

  Once upon a time I wrote a thesis. You are reading it now.

  Es war einmal als Ich eine Doktorarbeit schrieb. Sie lesen die jetzt

- ...

Problem statement

Beethoven's first music teacher was his father. Although tradition has it that Johann van Beethoven was a harsh instructor, and that the child Beethoven, "made to stand at the keyboard, was often in tears", the Grove Dictionary of Music and Musicians claimed that no solid documentation supported this.

Problem statement

[[*Beethoven***]***'s first music teacher***]** *was* **[[***his***]** *father***]**. *Although* **[***tradition***]** *has it that* **[***Johann van Beethoven***]** *was a harsh instructor, and that* **[***the child Beethoven***]**, *"made to stand at* **[***the keyboard***]**, *was often in tears"*, **[***the Grove Dictionary of Music and Musicians***]** *claimed that no solid documentation supported this.*

- Identify referential expressions

Problem statement

What do we want? Given a raw text

> [[*Beethoven*]*'s first music teacher*] *was* [[*his*] *father*]. *Although* [*tradition*] *has it that* [*Johann van Beethoven*] *was a harsh instructor, and that* [*the child Beethoven*]*, "made to stand at* [*the keyboard*]*, was often in tears",* [*the Grove Dictionary of Music and Musicians*] *claimed that no solid documentation supported this.*

- Identify referential expressions
- Cluster coreferring mentions together in *coreference chains*

Problem statement

> *Beethoven's first music teacher was his father. Although tradition has it that Johann van Beethoven was a harsh instructor, and that the child Beethoven, "made to stand at the keyboard, was often in tears", the Grove Dictionary of Music and Musicians claimed that no solid documentation supported this.*

With no absolute consensus on what are acceptable mentions, e.g. for

Problem statement

> *Beethoven's first music teacher was his father. Although tradition has it that Johann van Beethoven was a harsh instructor, and that the child Beethoven, "made to stand at the keyboard, was often in tears", the Grove Dictionary of Music and Musicians claimed that no solid documentation supported this.*

With no absolute consensus on what are acceptable mentions, e.g. for

- Full sentences

Problem statement

> *Beethoven's first music teacher was his father. Although tradition has it that Johann van Beethoven was a harsh instructor, and that the child Beethoven, "made to stand at the keyboard, was often in tears", the Grove Dictionary of Music and Musicians claimed that no solid documentation supported this.*

With no absolute consensus on what are acceptable mentions, e.g. for

- Full sentences
- Non coreferring mentions

Problem statement

> *Beethoven's first music teacher was his father. Although tradition has it that Johann van Beethoven was a harsh instructor, and that the child Beethoven, "made to stand at the keyboard, was often in tears", the Grove Dictionary of Music and Musicians claimed that no solid documentation supported this.*

With no absolute consensus on what are acceptable mentions, e.g. for

- Full sentences
- Non coreferring mentions
- Copula

Detecting coreference

Given the set of mentions, how to build coreference chains ?

Outside of early specific systems that only dealt with pronoun resolution, two main families of models

- Entity-mention models
- Mention-pair models

With countless variations.

Entity-mention models

Basic principle:

- Process the document in reading order
- Every time we encounter a mention, either
  - Add it to an existing chain
  - Create a new chain for it
- At the end of the documents we have our chains

A lot of possible refinements

- Processing the document globally
- Allow postponing decisions or changing previous ones
- Allow merging existing chains
- ...

Entity-mention models

Basic principle:

- Process the document in reading order
- Every time we encounter a mention, either
  - Add it to an existing chain
  - Create a new chain for it
- At the end of the documents we have our chains

Advantages:

- Cluster consistency is free
- Use of cluster-level information
- Reify the concept of “discourse entities” from cognitive linguistics

Main drawback: demand an efficient representation of clusters

Mention pair models

Basic principle:

- Considering all pairs of mentions in the document
- For each, decide if it is coreferring or not (usually with a confidence score)
- Then build chains that are consistent with these relations

Many variants, mostly in the clustering phase

- Antecedent finding (hence *chains*): for a given mention, chose an antecedent antecedent among coreferring mentions
  - → Two main heuristics : Best-First and Closest-First
- Graph optimization: find the clusters that maximize the global confidence score
  - Integer Linear Programming
  - Graph cutting
  - ...

Encoding mentions and entities

Non-neural models, typically model mentions using extensive lists of features

- Basic morphology: gender, number, case...
- Semantic class (for named entities: person, organization, location...)
- Simple syntax: syntactic head, constituent type...
- Discourse feature: speaker, inclusion in direct quotation...

Additional features for mention pairs

- (Sub)string matching
- Distance (in words and in number of mentions)

Entity features are usually derived from the features of the mentions they contain

The curse of linguistic ressources

Using handcrafted combinations of these features have proved to be very efficient

→ Enough for a rule-based system (H. Lee et al. 2013) to outperform statistical ones

But it is very demanding in linguistic ressources that are not always available

- Data with human annotation for both coreference and morphology, syntax... is very sparse in most languages
- Automatic annotations tool are not always available
- And if they are, they are not always performant enough
  → Particularly on non-standard language: oral, social media...
- Combining raw linguistic information into useful features is highly non-trivial
- Since 2013, only marginal performance improvements

Neural models

Neural networks in 5 seconds

A single neuron

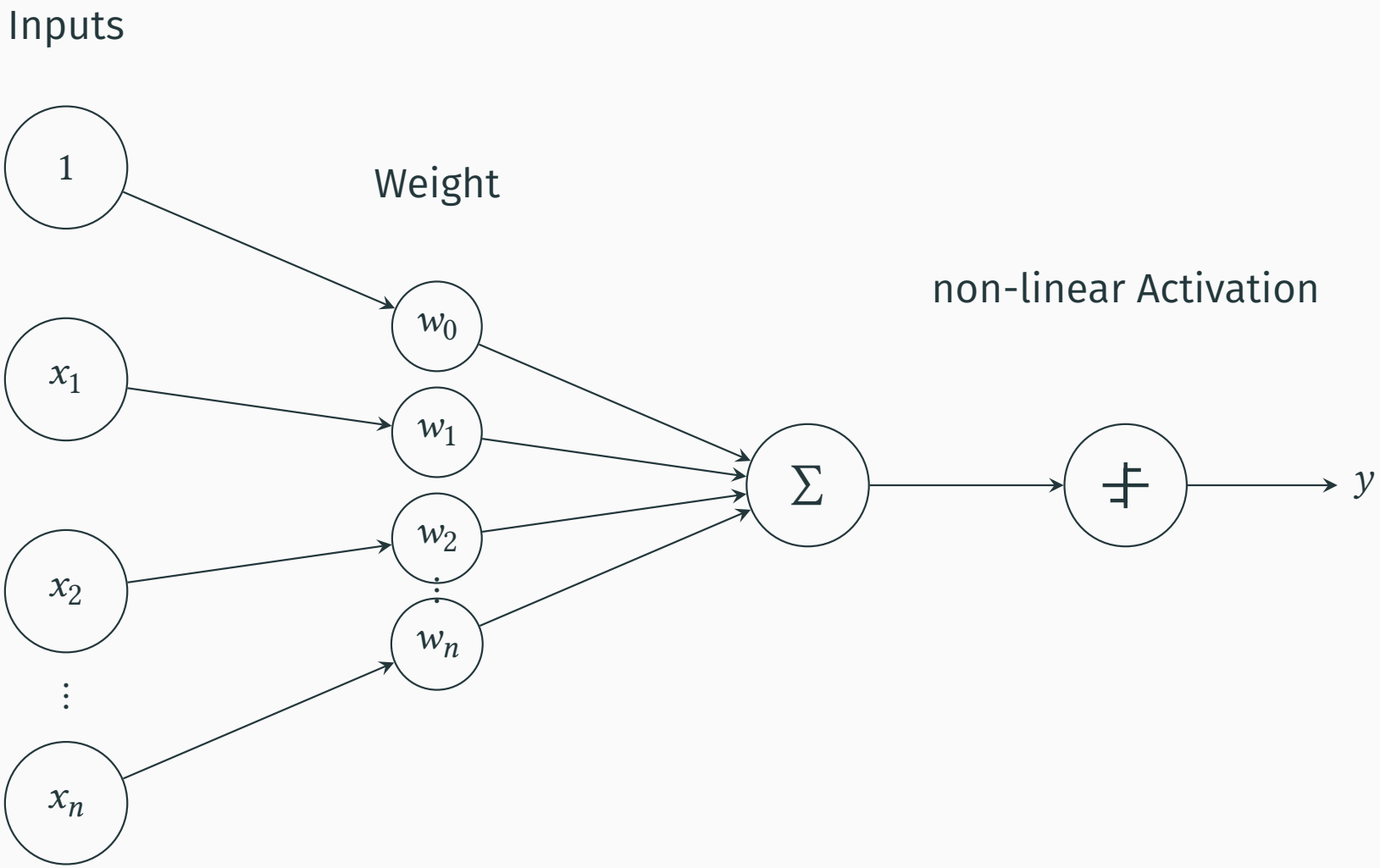


Figure 4: Simple perceptron

Neural networks in 5 seconds

Several layers of connected neurons

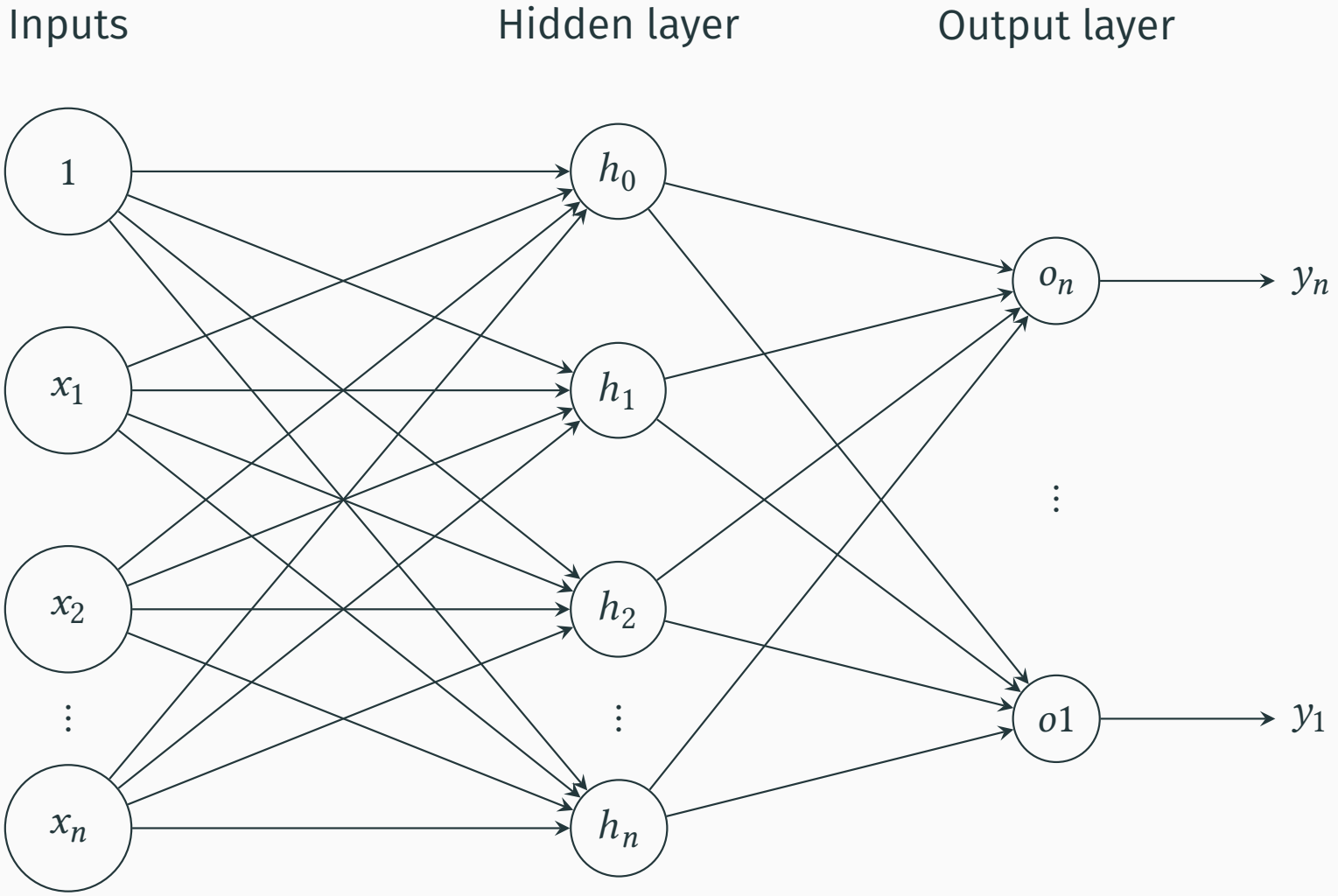


Figure 5: Feedforward neural network

Neural networks in 5 seconds

Neural networks can be used to approximate continuous functions on compact subsets of $\mathbb{R}^n$ with arbitrary precision (given enough neurons).

- Although *how* to find the better weights is still an ongoing question...

Neural models have achieved outstanding performance in machine learning tasks

→ Computer vision
→ Machine translation
→ Syntactic analysis

It is natural to try to apply them to coreference detection.

Neural models for coreference detection

Neural networks allow us to reduce the dependency on linguistic ressources and expert knowledge

- (Wiseman, Rush, Shieber, and Weston 2015), (Wiseman, Rush, and Shieber 2016)
  - → Uses features similar to those of (H. Lee et al. 2013), but lets the model learn the rules instead of handcoding them.
- (Clark and Manning 2016a), (Clark and Manning 2016b)
  - → Uses only lexical features and one syntactic feature

Both models evolved from mention-pair with antecedent ranking to entity-mention models, by learning to represent entities using mention-oriented features.

Neural models for coreference detection

"End-to-end Neural Coreference Resolution" (K. Lee et al. 2017) is even more interesting

- Almost completely resourceless
- Beats the previous two model by a non-negligible margin
- Performs both mention detection and coreference resolution by considering *all possible text spans*
  - → Compute a "mentionness" score for all text spans and keep only the credible ones

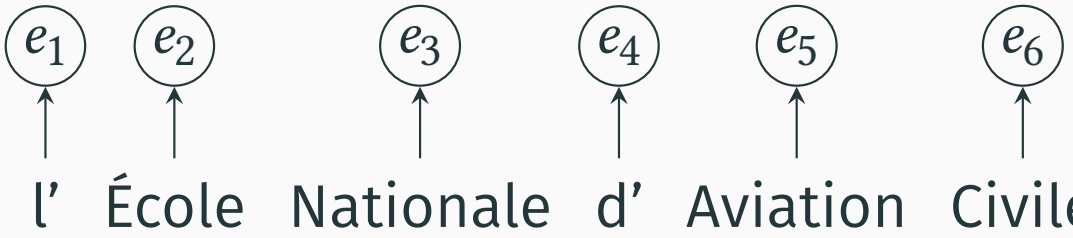
.

Encoding text spans

l’ École Nationale d’ Aviation Civile

Figure 6: Recurrent neural encoder

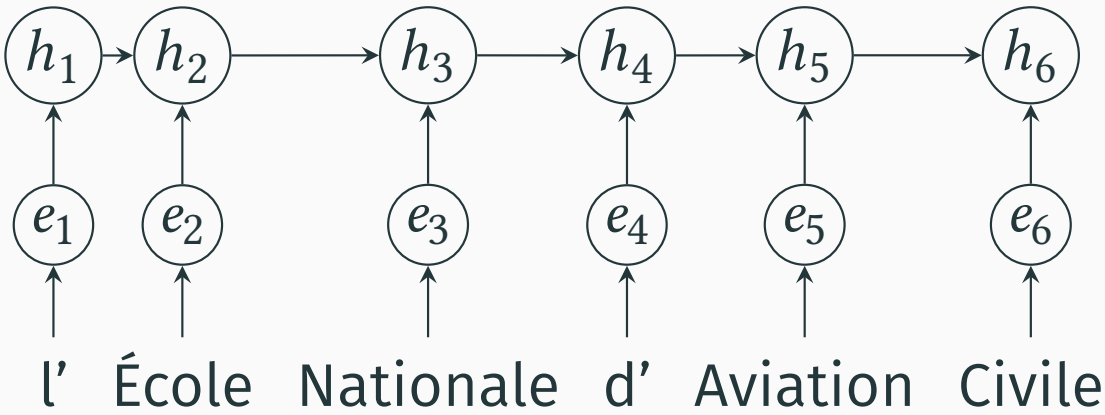
Encoding text spans



Word embeddings $\in \mathbb{R}^d$

Figure 6: Recurrent neural encoder

Encoding text spans

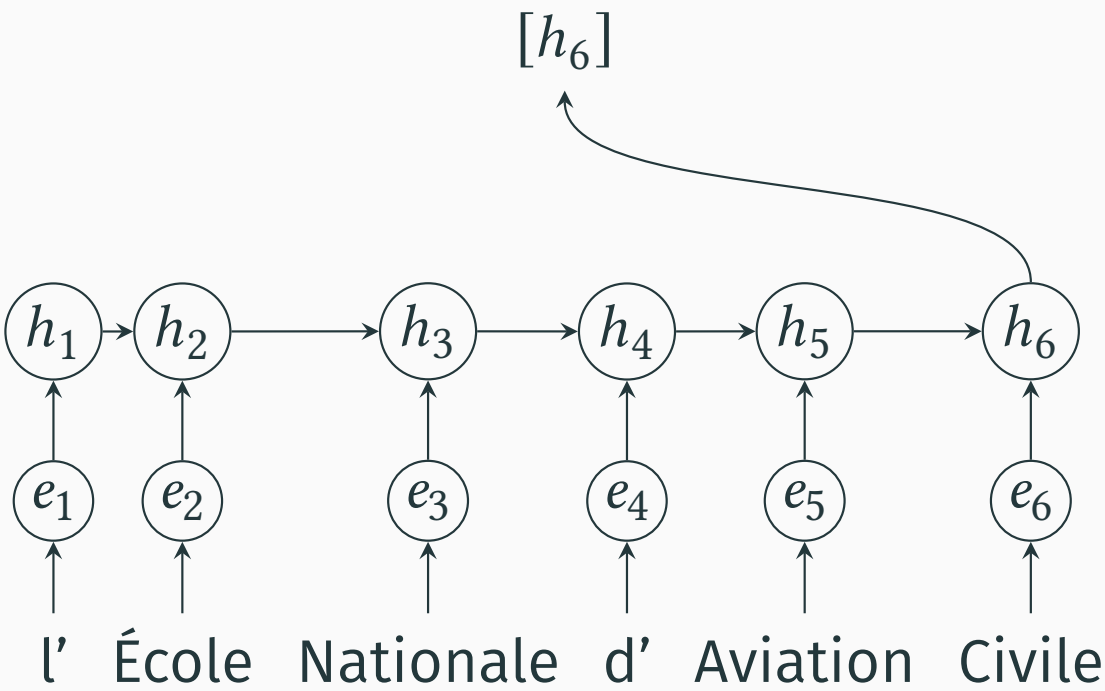


Hidden states $\in \mathbb{R}^k$

Word embeddings $\in \mathbb{R}^d$

Figure 6: Recurrent neural encoder

Encoding text spans



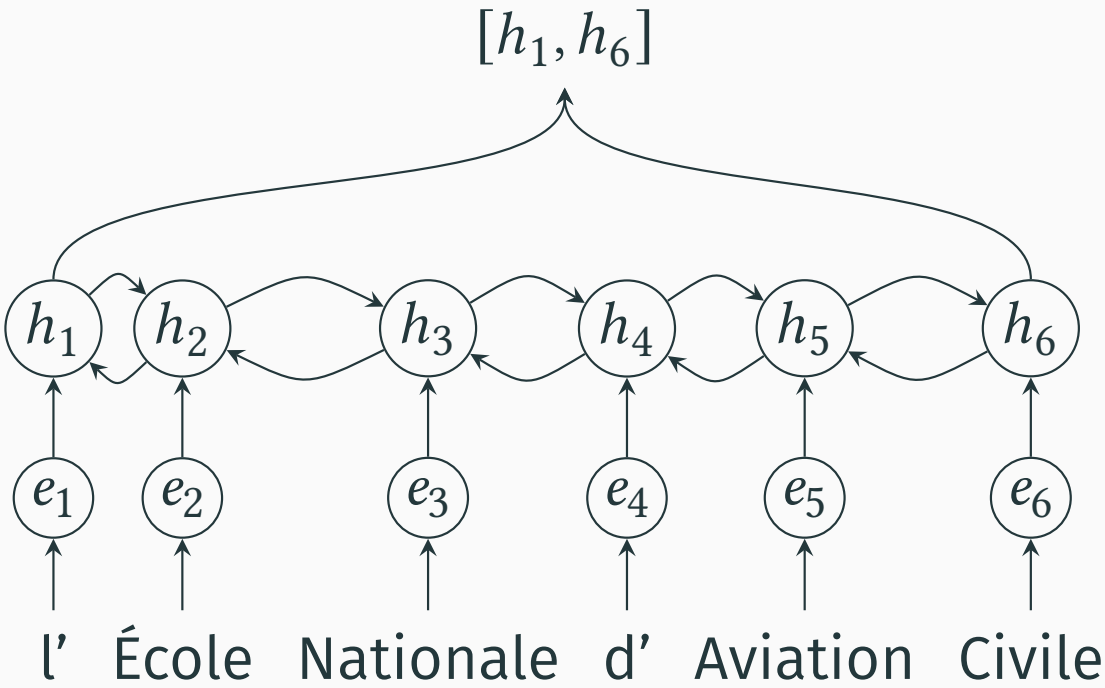
Span representation $\in \mathbb{R}^n$

Hidden states $\in \mathbb{R}^k$

Word embeddings $\in \mathbb{R}^d$

Figure 6: Recurrent neural encoder

Encoding text spans



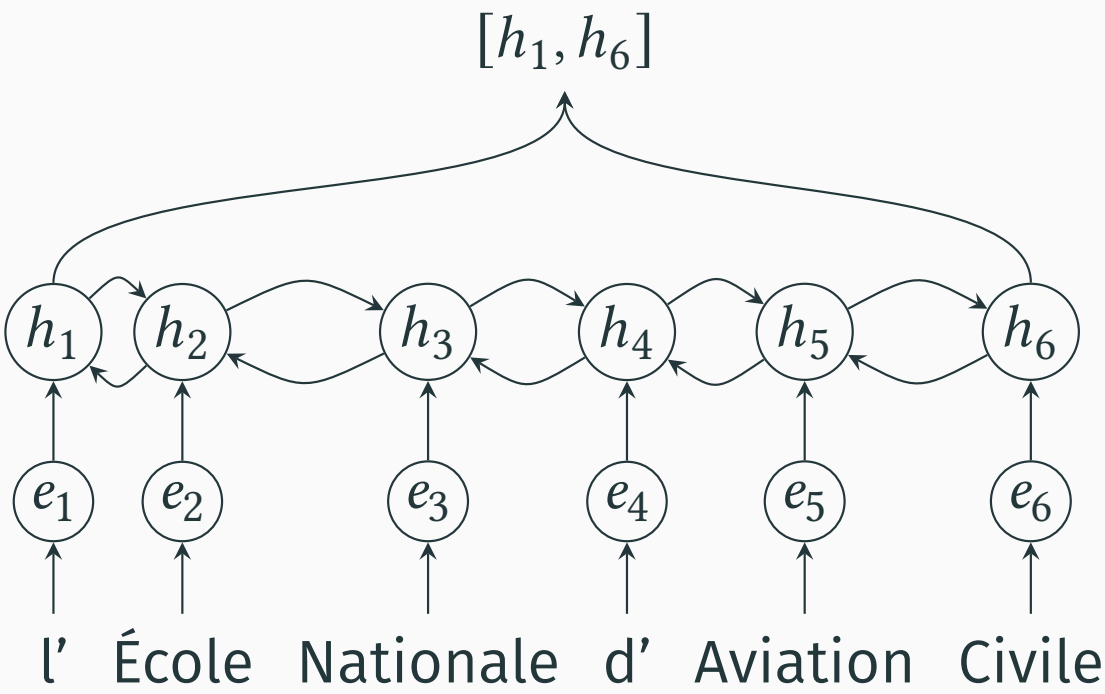
Span representation $\in \mathbb{R}^{2n}$

Hidden states $\in \mathbb{R}^{k}$

Word embeddings $\in \mathbb{R}^{d}$

Figure 6: Recurrent neural encoder

Encoding text spans



Span representation $\in \mathbb{R}^{2n}$

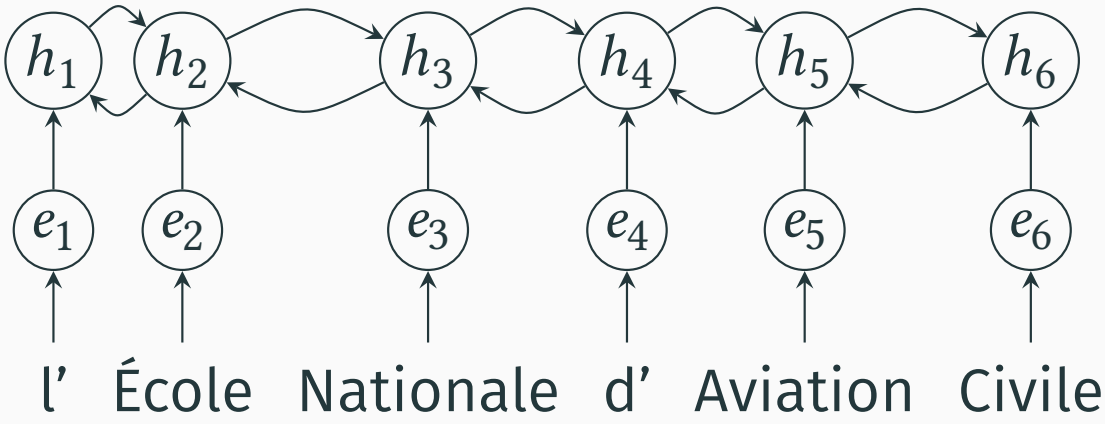
Hidden states $\in \mathbb{R}^{k}$

Word embeddings $\in \mathbb{R}^{d}$

Figure 6: Recurrent neural encoder

- Initially developed for machine translation
- Allow fixed-length representation of variable-length sequences

Encoding text spans

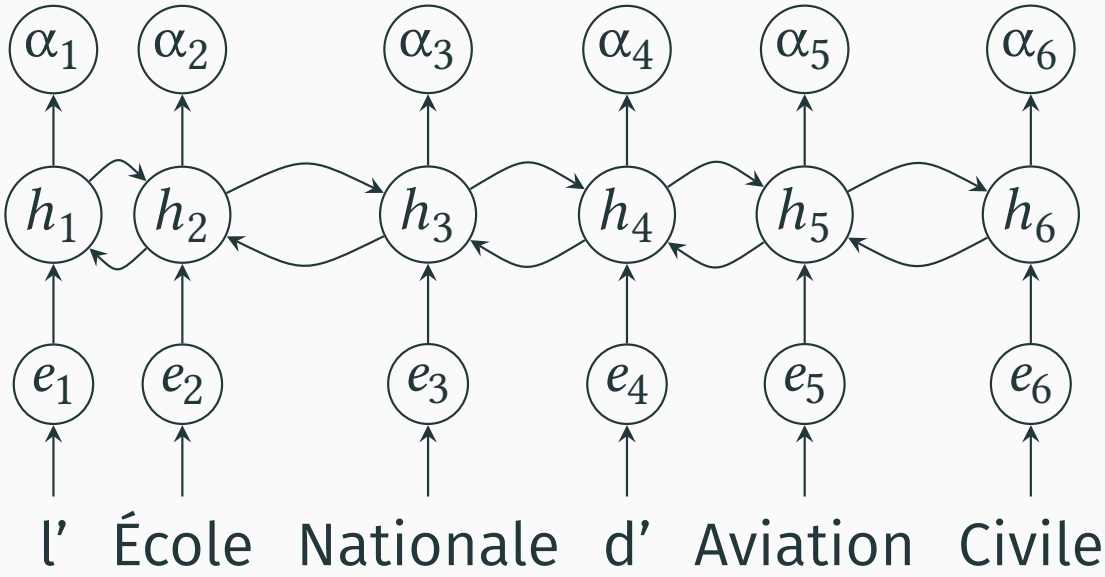


États cachés $\in \mathbb{R}^k$

Embeddings $\in \mathbb{R}^d$

Figure 7: Soft-head attention

Encoding text spans



Poids $\in [0, 1]$

États cachés $\in \mathbb{R}^k$

Embeddings $\in \mathbb{R}^d$

Figure 7: Soft-head attention

Encoding text spans

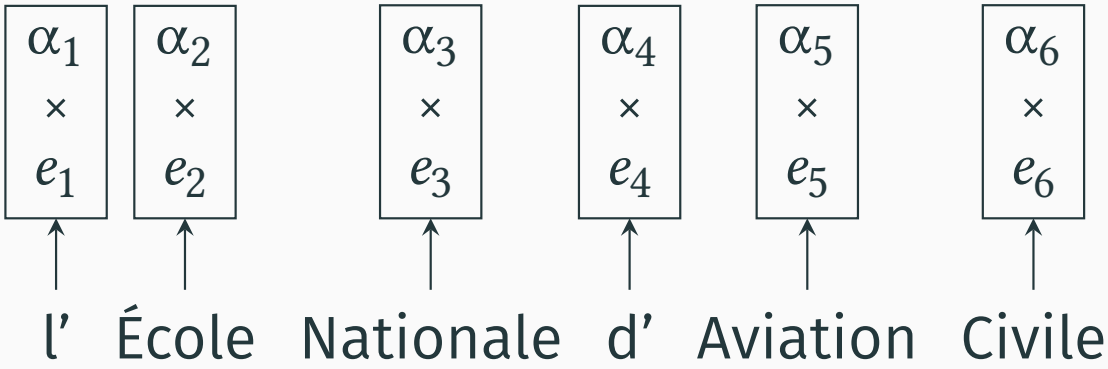
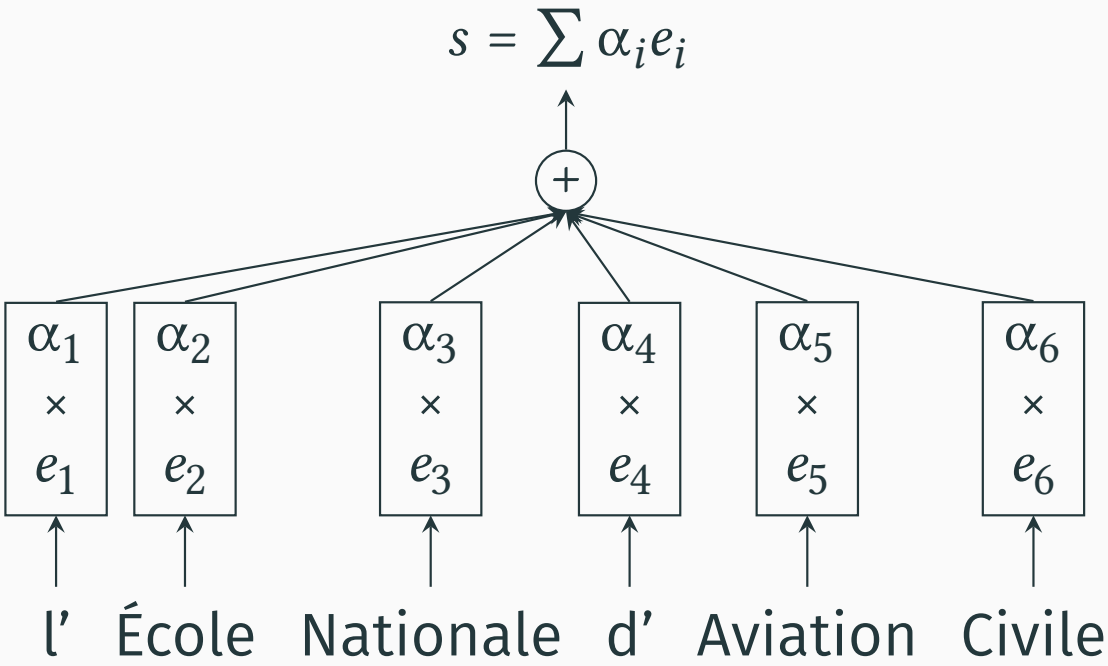


Figure 7: Soft-head attention

Encoding text spans



Soft-head $\in \mathbb{R}^d$

Learned weights $\in [0, 1]$

Embeddings $\in \mathbb{R}^d$

Figure 7: Soft-head attention

Encoding text spans

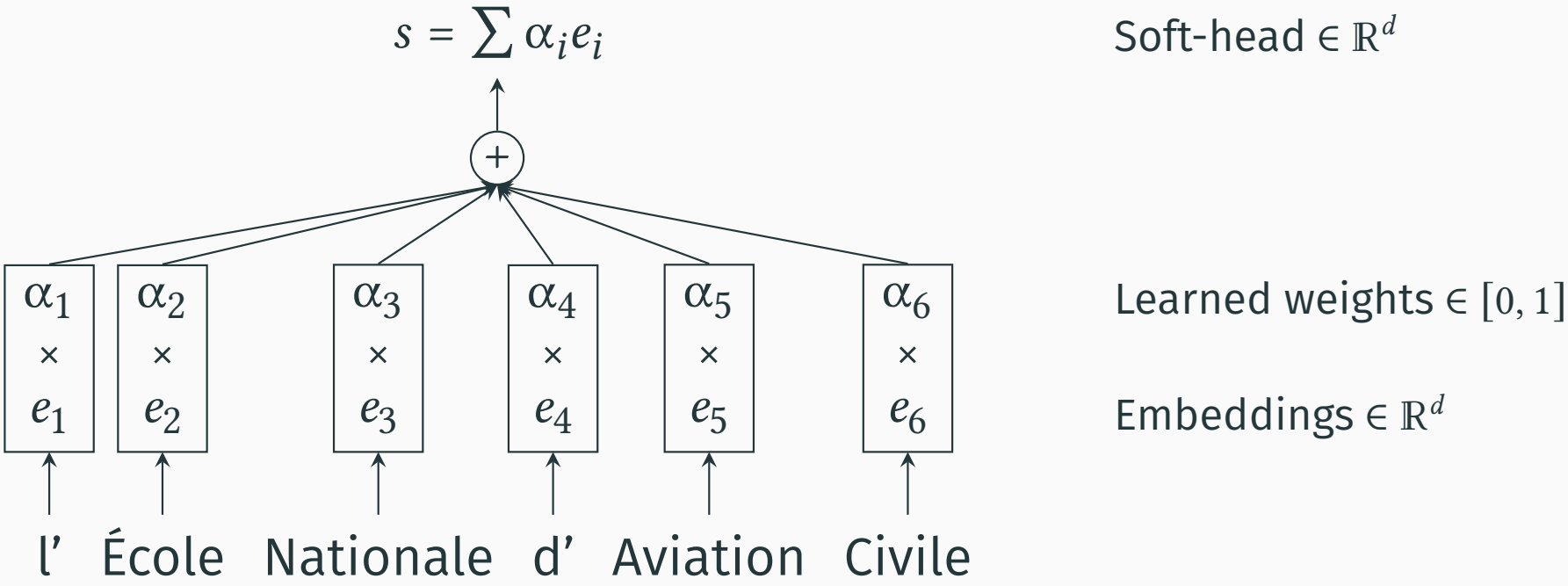


Figure 7: Soft-head attention

- Replace the one syntactic feature used by (Clark and Manning 2016b): the syntactic head of the mention
- Actually, in most cases the maximal α_i is attributed to the syntactic head

Our models

We are improving (K. Lee et al. 2017) by learning richer objectives : instead of simply learning to score, we also learn to classify

- Mention: common noun, proper noun, pronoun
- Relations: direct or indirect coreference, pronominal anaphora, bridging anaphora

Boostrapping the learning phase

1. Train only on mention detection
2. Train on coreference detection on reference (gold) mention
3. Train the full pipeline together

Speeds up and stabilize the convergence.

Final network structure

j’ ai fait mes études au lycée Pothier

Figure 8: Mention detection

Final network structure

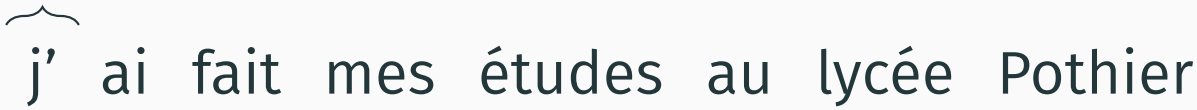


Figure 8: Mention detection

Final network structure

Figure 8: Mention detection

Final network structure

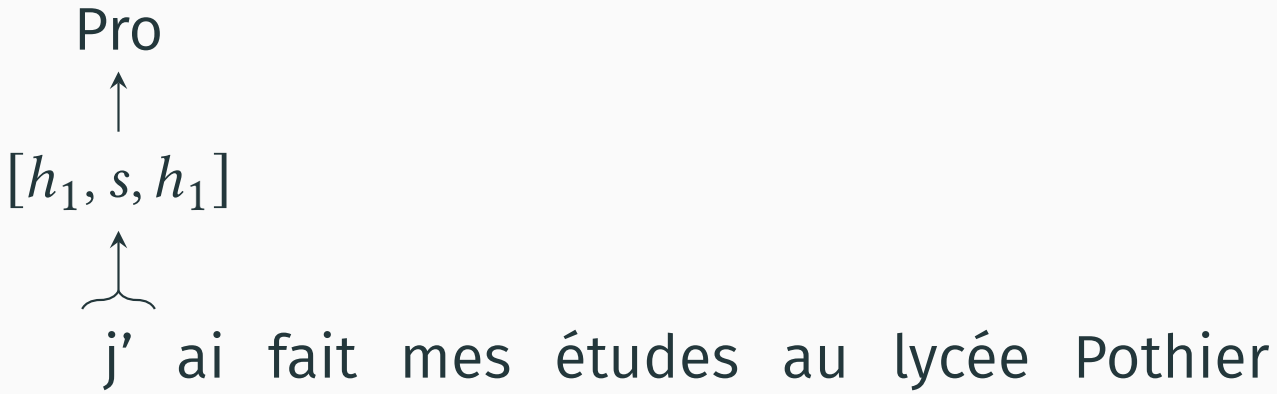


Figure 8: Mention detection

Final network structure

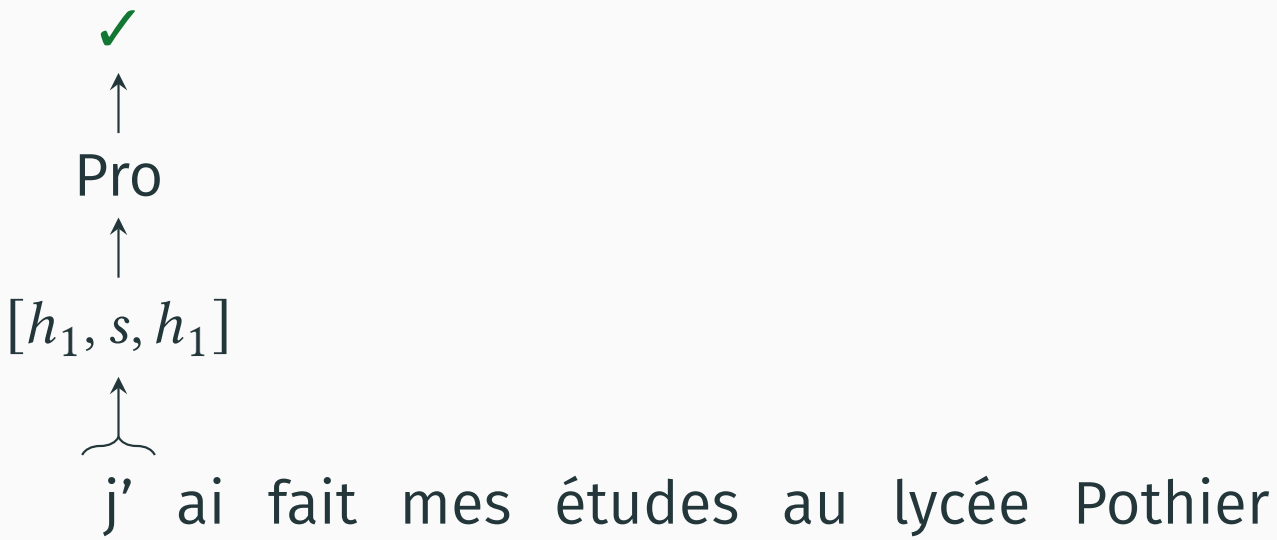


Figure 8: Mention detection

Final network structure

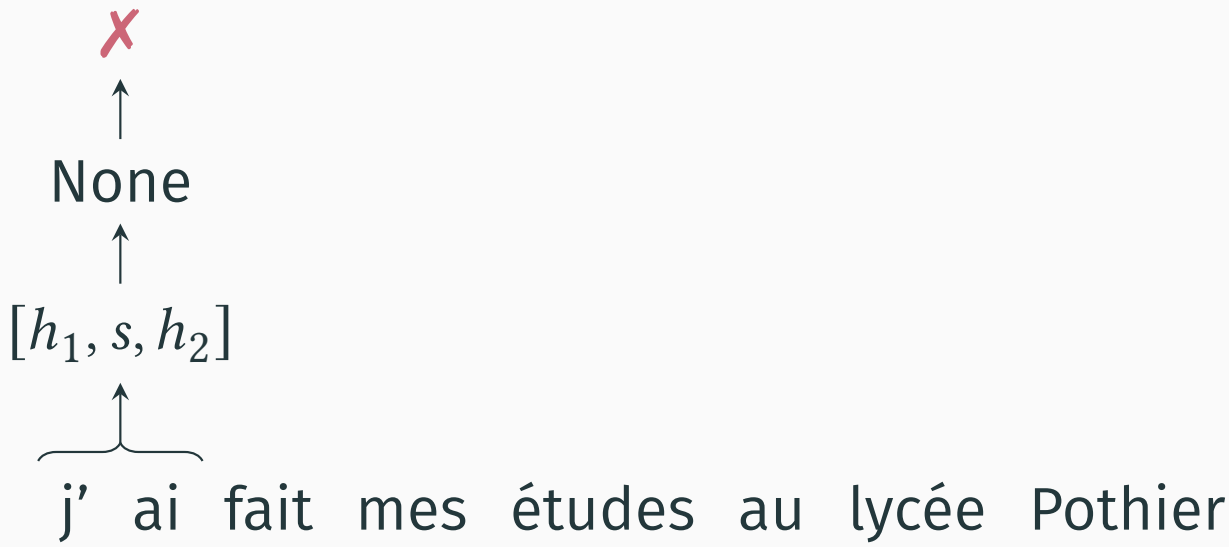


Figure 8: Mention detection

Final network structure

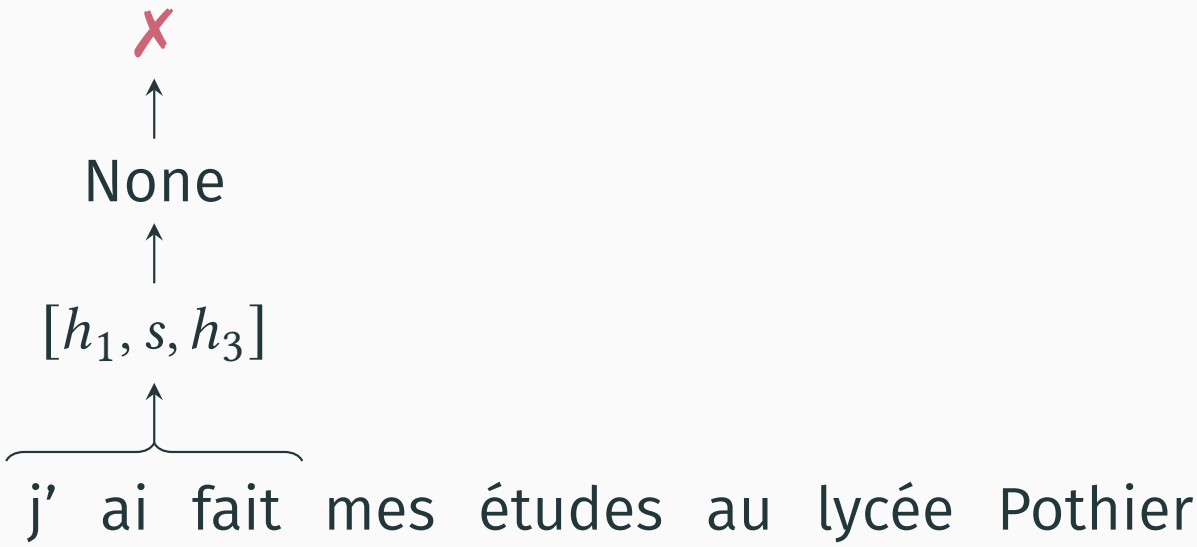


Figure 8: Mention detection

Final network structure

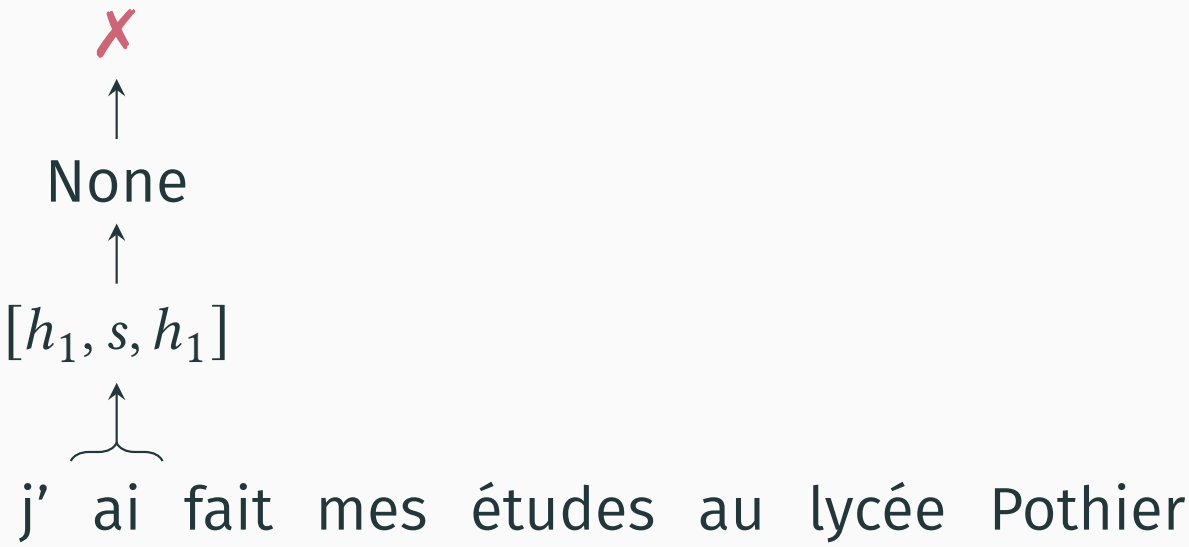


Figure 8: Mention detection

Final network structure

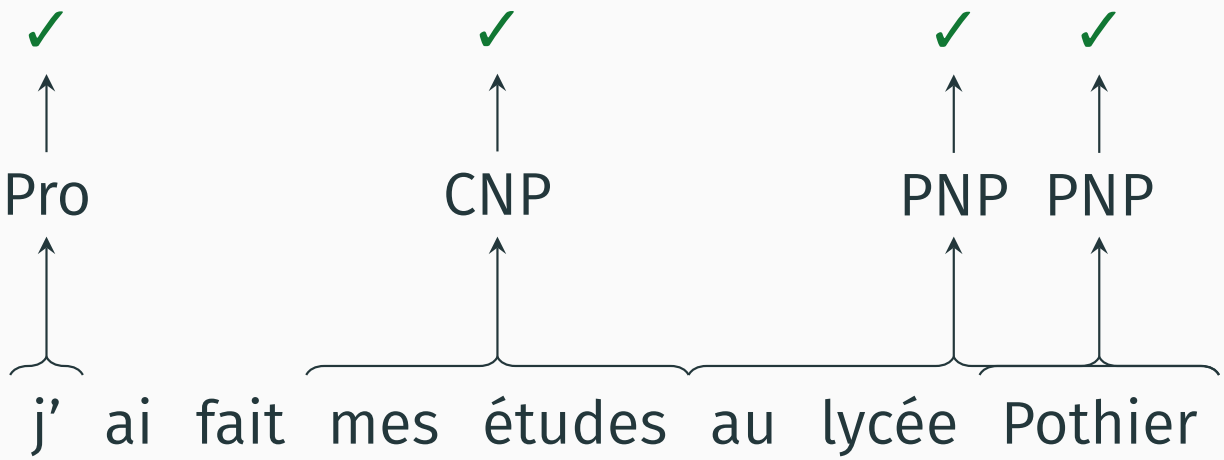


Figure 8: Mention detection

Network structure

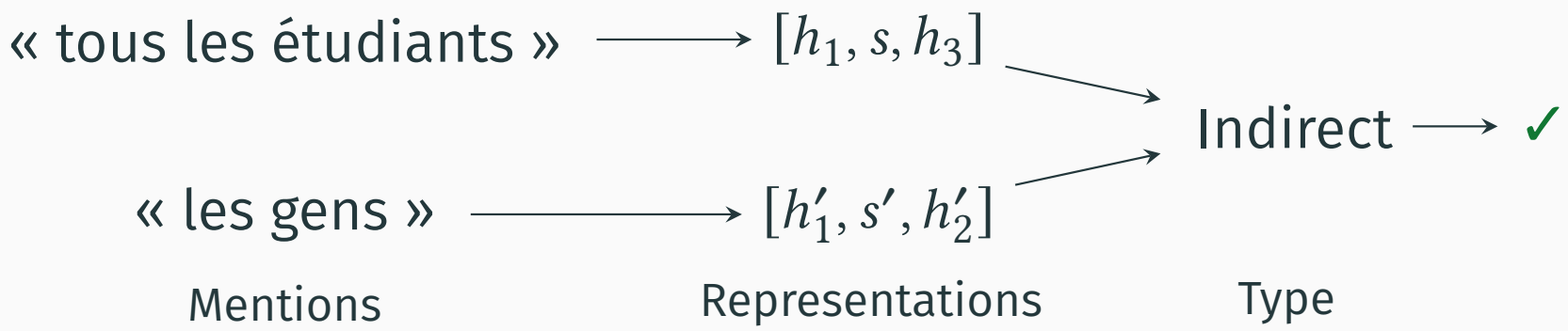


Figure 9: Structure du réseau : détection des paires coréférentes

Network structure

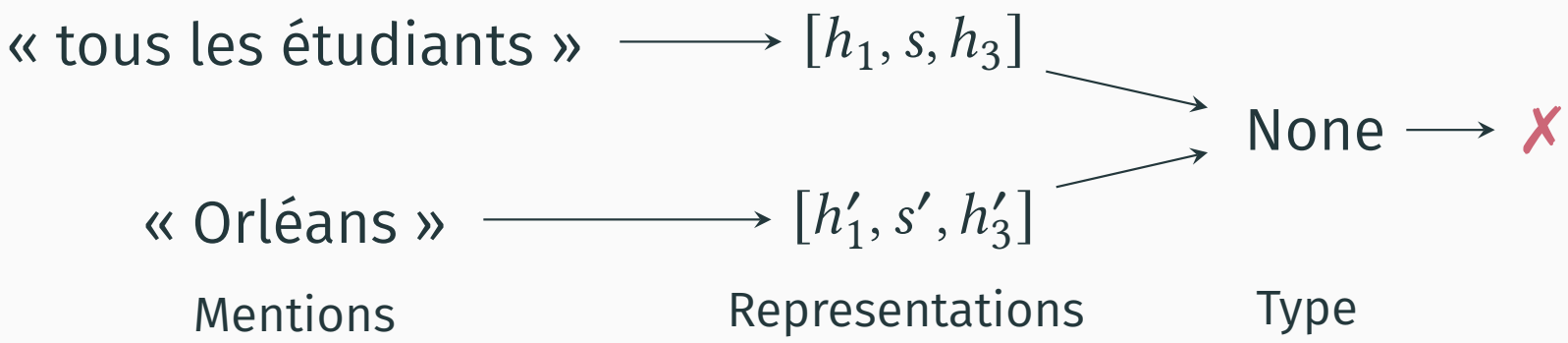


Figure 9: Structure du réseau : détection des paires coréférentes

Findings

Our experiments on French seem to confirm the trend observed on English

- Neural network outperform both rule-based- and traditional statistical models
- These models are not as reliant on linguistic knowledge and feature engineering
- They are proving robust even on non-standard language (in our case, oral French)

But

The pesky case of the Winograd challenges

The city councilmen refused the demonstrators a permit because they **feared** *violence.*

The pesky case of the Winograd challenges

The city councilmen refused the demonstrators a permit because they **feared** *violence.*

The pesky case of the Winograd challenges

The city councilmen refused the demonstrators a permit because they **advocated** *violence.*

The pesky case of the Winograd challenges

The city councilmen refused the demonstrators a permit because they **advocated** *violence.*

The pesky case of the Winograd challenges

The city councilmen refused the demonstrators a permit because they **feared** *violence.*

The pesky case of the Winograd challenges

The city councilmen refused the demonstrators a permit because they **feared** *violence.*

The pesky case of the Winograd challenges

The city councilmen refused the demonstrators a permit because they **advocated** *violence.*

The pesky case of the Winograd challenges

The city councilmen refused the demonstrators a permit because they **advocated** *violence.*

The pesky case of the Winograd challenges

The trophy would not fit in the brown suitcase because it was too big.

The pesky case of the Winograd challenges

The trophy would not fit in the brown suitcase because it was too small.

The puzzling case of encyclopedic knowledge

Good Omens, the new TV series from literary dream team Terry Pratchett and Neil Gaiman, is one of the most hotly anticipated shows of 2019. Since the disparition of the author of *The Colour of Magic*, the project had seemingly ground to a halt [...]

The puzzling case of encyclopedic knowledge

Good Omens, the new TV series from literary dream team Terry Pratchett and Neil Gaiman, is one of the most hotly anticipated shows of 2019. Since the disparition of the author of *The Colour of Magic*, the project had seemingly ground to a halt [...]

The puzzling case of encyclopedic knowledge

Now, the president of the United States is faced with a dilemma.

The puzzling case of encyclopedic knowledge

Obama has been chewing on his legacy for months. Now, the president of the United States is faced with a dilemma.

The road ahead

While neural models have been successful at breaking the stall in coreference detection, they have a major point of failure: world knowledge and common sense.

- Not surprising: efficiently collecting and modelling world knowledge is a hard™ problem
- Historically, applications (of coreference) have mostly been concerned by anaphoric pronoun resolution
  - → Mostly a syntactic or discursive — rather than semantic — phenomenon
  - → … except in contrieved cases

Even worse: well performing ((K. Lee et al. 2017) and our own) almost completely ignore the context of the mentions, relying instead on the consistence of the coreference chains.

This suggest that the real breakthrough is still to come.

Conclusion

- Coreference resolution is not a simple task
- Historical models had a high demand for ressources
- Neural models allow us to be less reliant on those
- Solving the problem in the general case will likely be harder than that
- Meanwhile, there is still a lot of room for improvement, even in this paradigm
  - Better context encoding for neural networks
  - More sophisticated clustering strategies
  - Better training strategies: generative adversarial networks, unsupervised/indirect learning...

Appendix

Acknowledgements

This work is part of the “Investissements d’Avenir” overseen by the French National Research Agency ANR-10-LABX-0083 (Labex EFL).

This work has been supported by the ANR DEMOCRAT (Description et modélisation des chaînes de référence: outils pour l’annotation de corpus et le traitement automatique) project ANR-15-CE38-0008.

References i

Clark, Kevin and Christopher D. Manning (2016a).
"Deep Reinforcement Learning for Mention-Ranking Coreference Models".
In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. URL: http://aclweb.org/anthology/D/D16/D16-1245.pdf.

— (2016b). "Improving Coreference Resolution by Learning Entity-Level Distributed Representations".
In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.
URL: http://aclweb.org/anthology/P/P16/P16-1061.pdf.

Lee, Heeyoung et al. (Dec. 2013).
"Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules".
In: *Computational Linguistics* 39.4. URL: http://dx.doi.org/10.1162/COLI_a_00152.

Lee, Kenton et al. (Sept. 2017). "End-to-end Neural Coreference Resolution".
In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.
URL: https://www.aclweb.org/anthology/D17-1018.

Wiseman, Sam, Alexander M. Rush, and Stuart M. Shieber (2016).
"Learning Global Features for Coreference Resolution". In: *CoRR* abs/1604.03035.
URL: http://arxiv.org/abs/1604.03035.

Wiseman, Sam, Alexander M. Rush, Stuart M. Shieber, and Jason Weston (2015).
"Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution".
In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*.
URL: http://aclweb.org/anthology/P/P15/P15-1137.pdf.

Licence