FACULTY OF ARTS AND PHILOSOPHY

INVESTIGATING CROSS-DOCUMENT EVENT COREFERENCE RESOLUTION FOR DUTCH

Loic De Langhe, Orphée De Clercq, Véronique Hoste

CRAC 2022 -16/10/2022

GHENT UNIVERSITY

language and translation technology team

OVERVIEW

1. **Introduction**
    - Event Coreference Resolution (ECR)
    - Motivation
2. **The ENCORE Corpus**
3. **Experiments**
    - Experimental Setup
    - Results and Error Analysis
4. **Conclusion and future work**

GHENT UNIVERSITY

EVENT COREFERENCE RESOLUTION (ECR)

1. ***A disciplined France team has beaten the Red Devils in the world cup semis earlier this week.***

2. ***Belgium lost the semi-final of the 2018 World cup against France with a score of 1-0 on Tuesday.***

GHENT UNIVERSITY

CONSIDERATIONS AND MOTIVATION

- Difficulties in ECR
  - Lack of data, sparsity and lack of uniformity
  - Coreference resolution across documents

- Why ECR research is valuable
  1. Move away from the paradigm of lexical semantics
  2. Focus on discourse-level relations to break down topic- and document barriers
  3. Practical multi-document applications such as summarization, content-based news recommendation and reading comprehension

CROSS-DOCUMENT ECR FOR DUTCH

GHENT UNIVERSITY

ENCORE CORPUS

- ENCORE *(De Langhe, De Clercq, Hoste, 2022)*
  - Large-scale Dutch event coreference corpus
  - Data sourced from a variety of Dutch (Flemish) newspapers
  - Focus on unrestricted events
  - Coreference annotated within event clusters

Cluster id	Topic	# of documents
47	Tim Burton exposition in Genk	11
75	Royal Wedding Prince Harry	24
87	Election of Cuban president	12

GHENT UNIVERSITY

ENCORE CORPUS: EVENTS

- Events are annotated based on ECB+ guidelines *(Cybulska & Vossen, 2014)*

In Guatemala, the volcan de fuego has erupted again today

Event Arguments

- Event Action (trigger)
- Event Time
- Event Location
- Event Participants
  - Human Participants
  - Non-Human Participants

Event Properties

- Event Prominence (Main)
- Event Realis (Certain)
- Event Sentiment (Negative)

GHENT UNIVERSITY

ENCORE CORPUS: COREFERENCE

1. **Identity Coreference between events**
   - Event time (1) , location (2) and participants (3) must match

"The 2012 London Olympics were a succes. The games lasted from 27/07/12 to 13/08/12."

2. **Part-whole Coreference between events**
   - One event is fully encompassed by the other

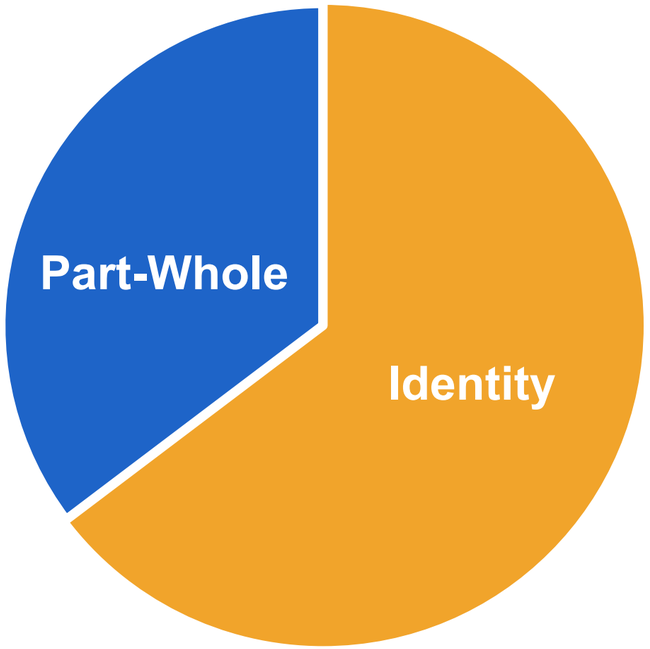
"The oscar ceremony was about to begin, with the presenters preparing for the opening speech."

GHENT UNIVERSITY

ENCORE CORPUS: OVERVIEW

Statistics

Description	Count
Documents	1115
Topics	91
Events	15407
Event Arguments	35315
Intra-document Event Coreference Chains	1018
Cross-Document Event Coreference Chains	1587



***Graph 2**: Distribution of Identity (28561) and Part-whole (15587) links between events*

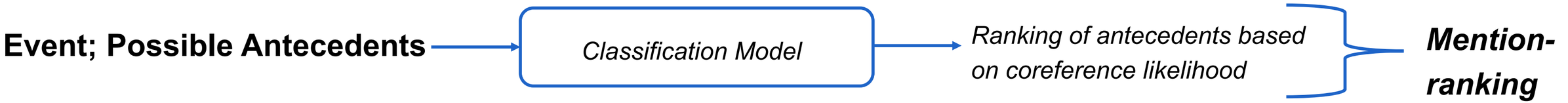
GHENT UNIVERSITY

EXPERIMENTS

GHENT
UNIVERSITY

EXPERIMENTAL SETUP

- **Task Description**
  - Classification of coreference based on *gold-standard event* mentions
  - For now, only identity links are considered
  - *Within-document* and *cross-document setting*
  - Experiments based on two existing paradigms in coreference studies:



GHENT UNIVERSITY

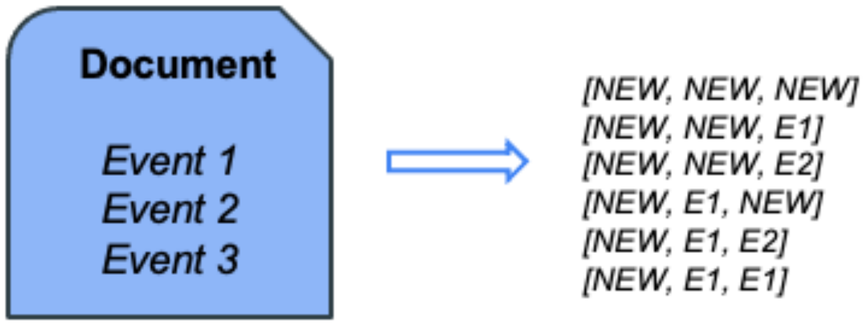
EXPERIMENTAL SETUP: MENTION-PAIR MODELS

- Transformer-based language models
  - BERTje *(de Vries et al., 2019)*
  - RobBERT *(Delobelle et al., 2020)*
  - RobBERTje *(Delobelle et al., 2022)*
  - XLM-RobBERTa *(Lample and Conneau, 2019)*
  - mBERT (*Devlin et al., 2019*)

- Traditional feature-based learning
  - Gradient-boosted Tree algorithm *(Chen et al., 2015)*
  - Features based on those used in English-language ECR studies
    - Lexical Similarity
    - Discourse
    - Logical Constraining

EXPERIMENTAL SETUP: MENTION-RANKING MODEL

- **Feature-based Mention-Ranking** *(Lu and Ng, 2017)*
  - Log-linear coreference model
  - Ranking of Document partitions
  - Task-specific loss based on type of error



$$L(\Theta) = \sum_{d=1}^{t} log \sum_{a \in A(C_d^*)} p(a|d;w)' + \lambda ||\Theta||_1$$

$$p(a|d;w)' \propto p(a|d;w) l(a, C_d^*)$$

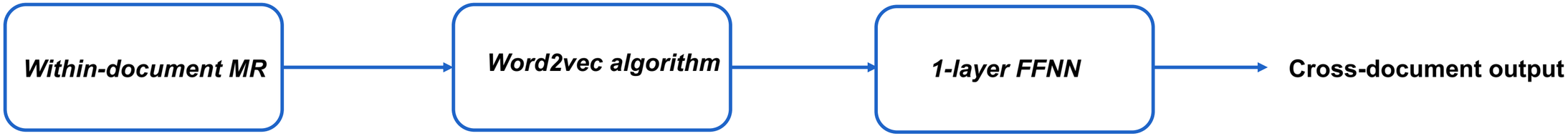
$$p(a|d;w) \propto exp(\sum_{i=1}^{n} w \cdot f(i, a_i, d))$$

GHENT UNIVERSITY

EXPERIMENTAL SETUP: MENTION-RANKING MODELS

- **Cross-Document Mention-Ranking**
  1. Poses scalability issues in memory
  2. Creates artifical sparsity problem in the data

➔ Pairwise classification of within-document coreference chains



GHENT UNIVERSITY

RESULTS AND DISCUSSION

GHENT UNIVERSITY

IDENTIFYING IDENTITY COREFERENCE RELATIONS

- **Results**

	CONLL	LEA
MP XGBoost	0.36	0.21
MR_{base}	0.39	0.25
$MR_{task-specific}$	0.42	0.26
MR $Embedding_{base}$	/	/
MR $Embedding_{task-specific}$	/	/
MP BERTje	**0.52**	**0.33**
MP RobBERT	0.49	0.29
MP RobBERTje	0.48	0.29
MP XLM-RoBERTa	0.17	0.11
MP mBERT	0.14	0.08

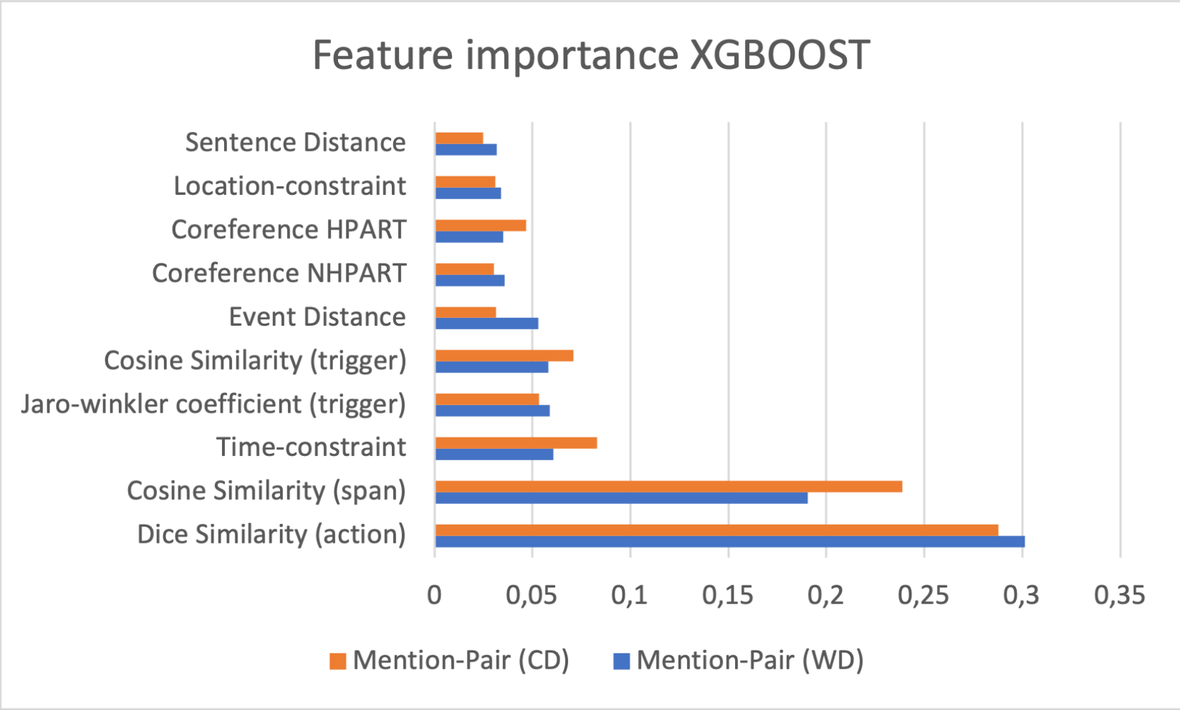
(a) Results for within-document ECR

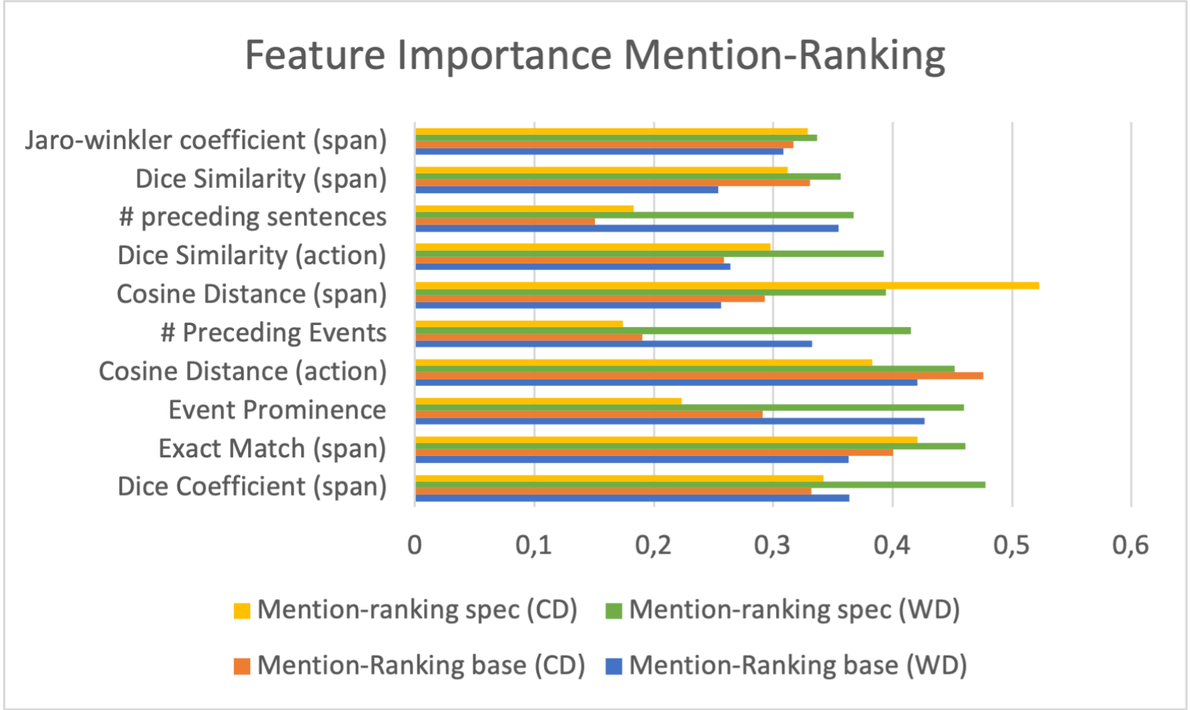
	CONLL	LEA
MP XGBOOST	0.37	0.23
MR_{base}	0.35	0.22
$MR_{task-specific}$	0.38	0.25
MR $Embedding_{base}$	0.36	0.24
MR $Embedding_{task-specific}$	0.40	0.28
MP BERTje	**0.59**	**0.39**
MP RobBERT	0.56	0.38
MP RobBERTje	0.54	0.35
MP XLM-RoBERTa	0.23	0.14
MP mBERT	0.19	0.10

(b) Results for cross-document ECR

GHENT UNIVERSITY

DISCUSSION: FEATURE-BASED MODELS

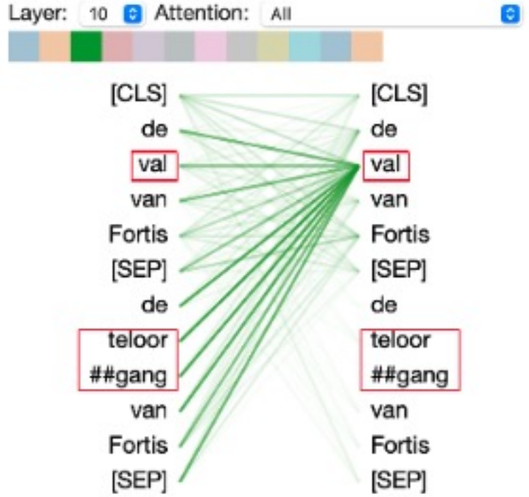




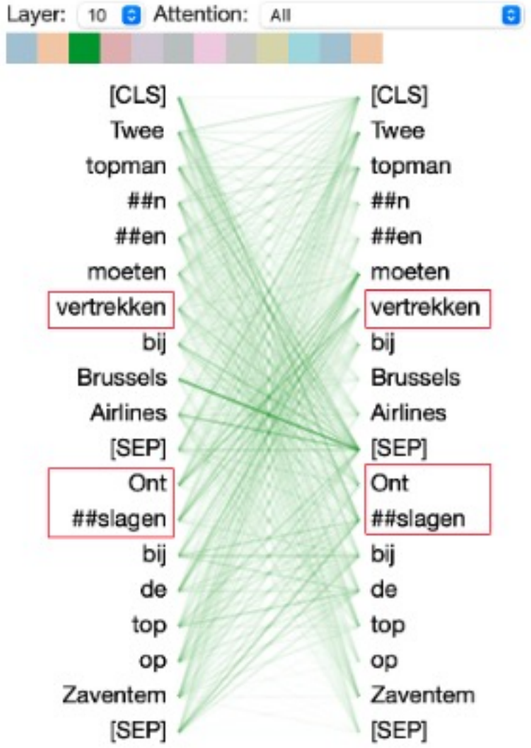
GHENT UNIVERSITY

DISCUSSION: TRANSFORMER-BASED MODELS

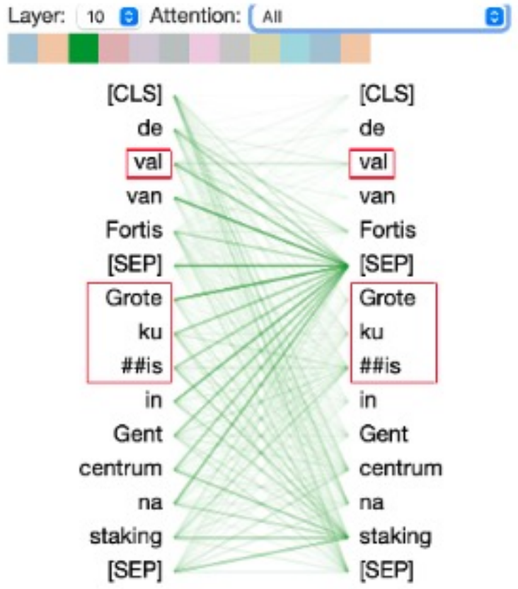
- Visualisation of transformer attention heads (Vig, 2019)



(a) *The [downfall] of Fortis vs. The [decline] of Fortis*



(b) *Two executives have to [leave] at Brussels Airlines vs. [Dismissals] at the top of Zaventem*



(c) *The [downfall] of Fortis vs. [Large cleanup] in Gent city center after strikes*

GHENT UNIVERSITY

CONCLUSION AND FUTURE RESEARCH

GHENT UNIVERSITY

CONCLUSION

- **Conclusion**
  - ENCORE corpus allows exploration of Dutch event coreference resolution
  - Large number of baseline experiments performed using both transformer-based and feature-based methods
  - Analysis of the baseline (feature-based) experiments show trends similar to ECR studies in English

- **Future work**
  - Integration of the baseline models into existing event mention detection sytems (pipeline architecture)
  - Exploratory studies regarding event-subevent relationships
  - Development of joint extraction-coreference systems for Dutch ECR using SpanBERT architectures, Graph Neural Networks (GNN), …

GHENT UNIVERSITY