BERT meets d'Artagnan : Data Augmentation for Robust Character Detection in Novels

COMHUM 2022

Arthur Amalvy[1] Vincent Labatut[1] Richard Dufour[2]

[1]Laboratoire Informatique d'Avignon
{firstname.lastname}@univ-avignon.fr

[2]Laboratoire des Sciences du Numérique de Nantes
richard.dufour@univ-nantes.fr

June 9, 2022

LABORATOIRE INFORMATIQUE D'AVIGNON

AVIGNON UNIVERSITÉ

Outline

The Named Entity Recognition (NER) Task
Task Description

- Extract named entities and their type (person, organisation...) from a text

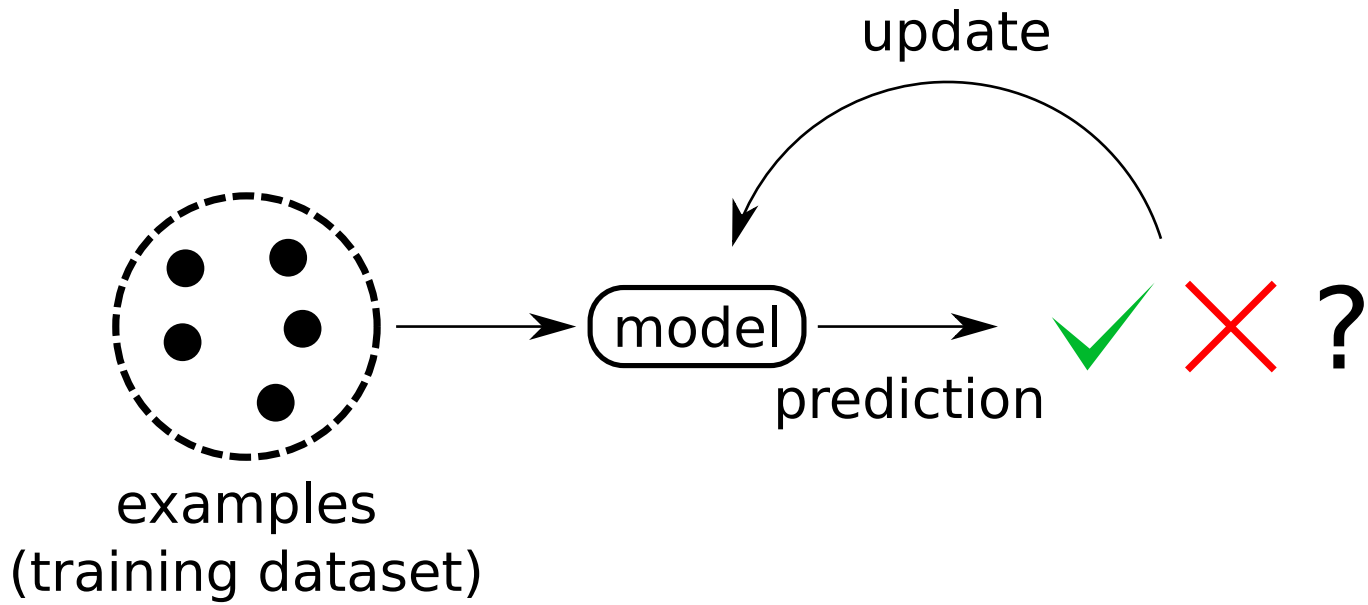
"My lord. Gandalf the Grey *PERSON* is coming."

- Useful for higher level tasks (such as character network extraction)
- Formulated as token classification

My	lord	.	Gandalf	the	Grey	is	coming	.
O	O	O	PER	PER	PER	O	O	O

The Named Entity Recognition (NER) task
Machine-Learning Based NER

- General training process of a NER classifier



- Training dataset role is central
- NER systems often trained on journalistic datasets

Dataset Domain Issue
NER Recall Errors in Literary Texts

- Dekker et al. 2019 : Evaluation of NER systems for character networks extraction in novels
- They annotated the first chapter of 40 novels
- Recurring recall issues :

word names

Dancing	was	an	old	time	general
O	O	O	O	O	O

names with special characters (common in fantasy !)

I	've	not	much	good	to	say	about	Nynaeve al'Meara
O	O	O	O	O	O	O	O	O O

Dataset Domain Issue
How to Fix the Domain Issue ?

- Certain types of names are not in training datasets
- Domain-specific datasets are needed
- Annotating is expensive !
- Automatic methods to extend datasets

Data Augmentation
Definition

- Data augmentation : generation of new examples by modification
- Very sucessful in image processing

Label : **BIRD**

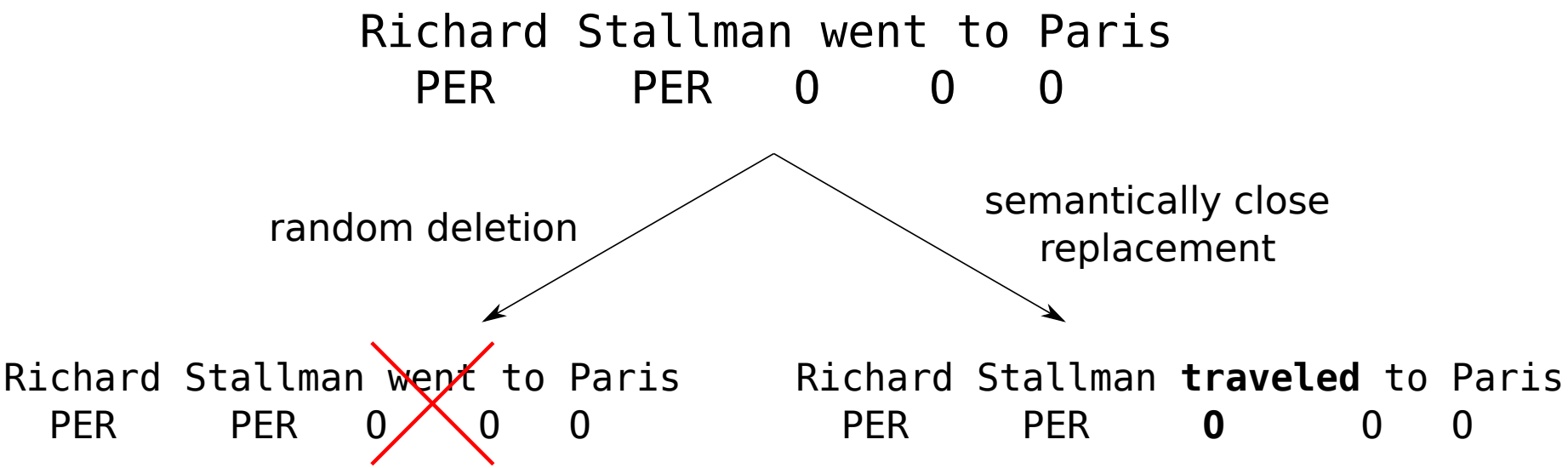
Flip augmentation

Label : **BIRD**

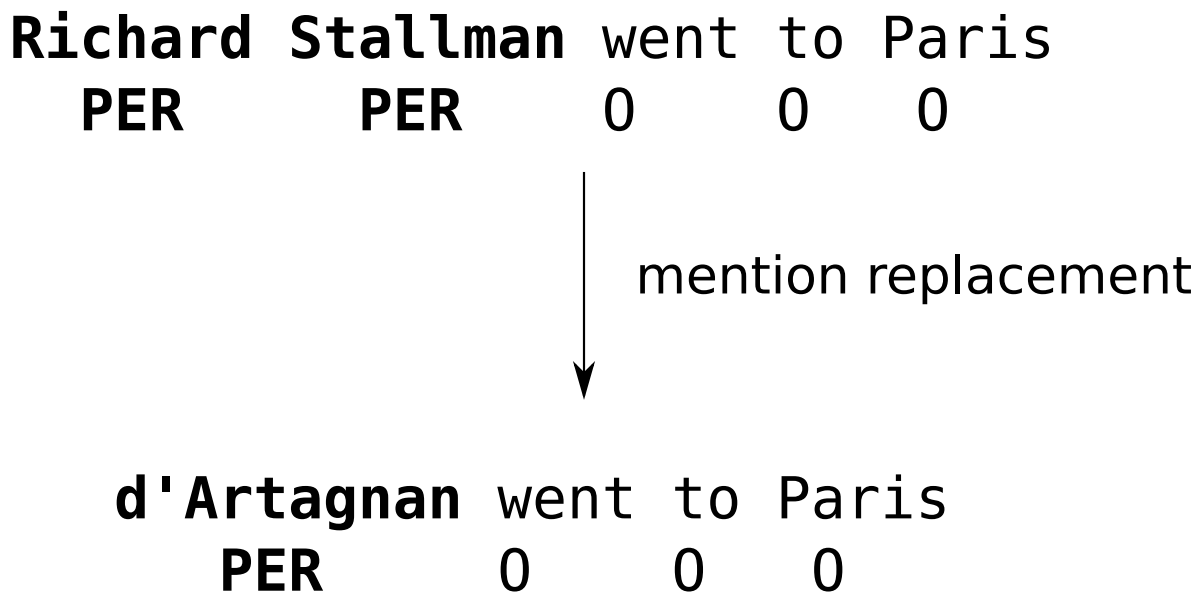
- Harder / less explored for text

Data Augmentation
Transposition to Text

- Some examples of data augmentation applied to NER :



Method
Contribution : Using Mention Replacement



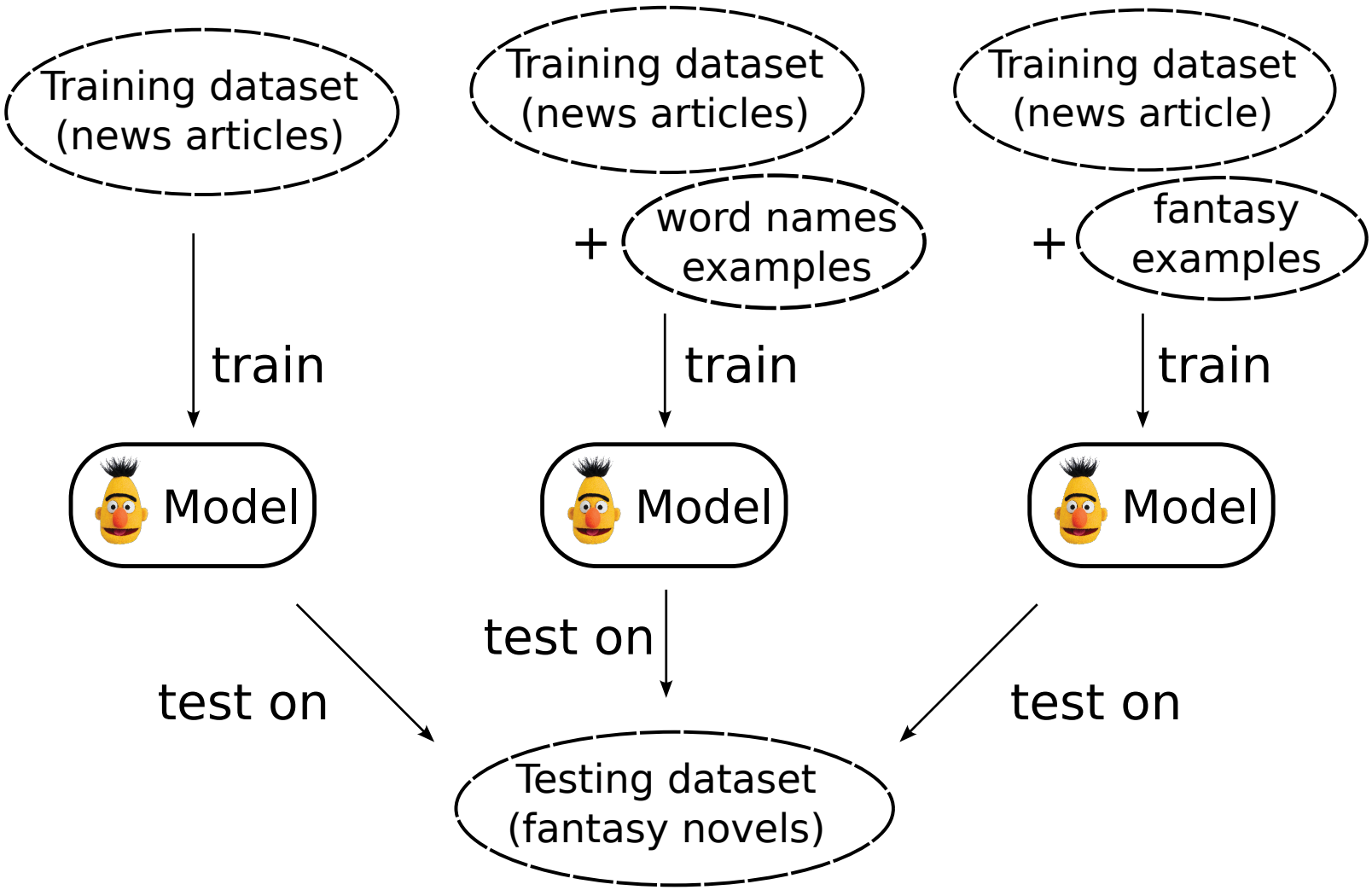
- Might help the model to pick up specific types of names

Method
Contribution : Mention Replacement Lists

- Word Names list
  - English first names ∩ common nouns [1]
  - 493 names
  - Examples : `"Baron"`, `"Red"`, `"Pearl"` ...
- Fantasy list
  - All characters from `The Elder Scrolls III : Morrowind`
  - 1324 names
  - Examples : `"Ma'Jidarr"`, `"Sul-Matuul"`, `"Radd Hard-Heart"` ...

[1] using WordNet (Miller 1995)

Experiments
Experiments Overview



Experiments
Training Dataset : Modified CoNLL-2003 (Tjong Kim Sang and De Meulder 2003)

- A journalistic dataset
- One of the most common NER dataset
- 14041 sentences

Example

```
The mutiny forced Patasse to miss the summit .
 O    O       O      PER  O   O    O    O    O
```

Experiments
Test Dataset

- Fantasy books from the dataset of Dekker et al. 2019
- First chapter of 17 novels, 5518 sentences
- We corrected annotations with a semi-automatic process

Example

```
Mr.   Bilbo Baggins has gone away
PER I-PER  I-PER   O   O    O
```

Results
Effects of Mention Remplacement

Augmentation	Precision	Recall	F1-score
`none`	93.11	87.03	89.89
`word names`	90.52	90.70	90.57
`fantasy`	89.80	91.36	90.51

Insights :

- F1 is slightly improved
- Recall is improved but precision decreases
- False positives can be filtered

Results
Some examples

Word name fixed detection

Bug	passed	the	steering	pole	to	Calo
PER	O	O	O	O	O	PER

Fantasy-style fixed detection

I	've	not	much	good	to	say	about	Nynaeve	al'Meara
O	O	O	O	O	O	O	O	PER	PER

New precision error (ambiguity ?)

"	Drink	,	friend	.	[…]	"
O	PER	O	O	O	O	

Conclusion

- Some character names in novels are hard to detect
- Mention replacement increases recall

- Precision decreases
  - What is the cause ?
  - Can we fix it ?
- Can we fix other issues with mention replacement ?

Section 7

References

AVIGNON
UNIVERSITÉ

References I

[1] N. Dekker, T. Kuhn, and M. van Erp. "Evaluating named entity recognition tools for extracting social networks from novels". In: *PeerJ Computer Science* 5 (2019), e189. DOI: 10.7717/peerj-cs.189.

[2] G. A. Miller. "WordNet: A Lexical Database for English". In: *Communications of the ACM* 38.11 (1995), pp. 39–41. DOI: 10.1145/219717.219748.

[3] E. F. Tjong Kim Sang and F. De Meulder. "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *7th Conference on Natural Language Learning*. 2003, pp. 142–147. URL: https://aclanthology.org/W03-0419.