Automated Testing and Improvement of Named Entity Recognition Systems

Boxi Yu, Yiyan Hu, Qiuyang Mang, Wenhan Hu, and Pinjia He
School of Data Science
The Chinese University of Hong Kong, Shenzhen

香 港 中 文 大 學（深 圳）
The Chinese University of Hong Kong, Shenzhen
数 据 科 学 学 院
School of Data Science

What is Named Entity Recognization

Named Entity Recognition (NER) is the process of identifying and categorizing named entities in a given text.

When Sebastian Thrun PERSON started at Google ORG in 2007 DATE few people outside of the company took him seriously. “I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun PERSON , now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE .

A little less than a decade later DATE , dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

NER: The foundation of various NLP tasks

(1) Information extraction

(2) Question answering

(3) Sentiment analysis

NER Systems are NOT robust

Error Type	Sentence	Predicted Entities	Target Entities
Omission	Sir Paul's command of the stage is so casual that he makes it look easy (**Flair-Ontonotes**).	NULL	["Paul", PER]
Over-labeling	Elon Musk is having a similar effect on the platform (**Azure**).	["Elon Musk", PERSON] ["Platform", LOCATION]	["Elon Musk", PERSON]
Incorrect Category	Norrie believes securing Unesco status could offer new opportunities in sustainable tourism and branding of local produce, while at the same time highlighting the environmental value of the peatland (**Flair-Conll**).	["Norrie", PER] ["Unesco", MISC]	["Norrie", PER] ["Unesco", ORG]
Range Error	Det Supt Rance said the investigation remained active (**AWS**).	["Det", PERSON] ["Supt Rance", PERSON]	["Det Supt Rance", PERSON]

How to detect Potential NER Errors? *Differential Testing?*

Main Challenge: Different NER systems have various standards.

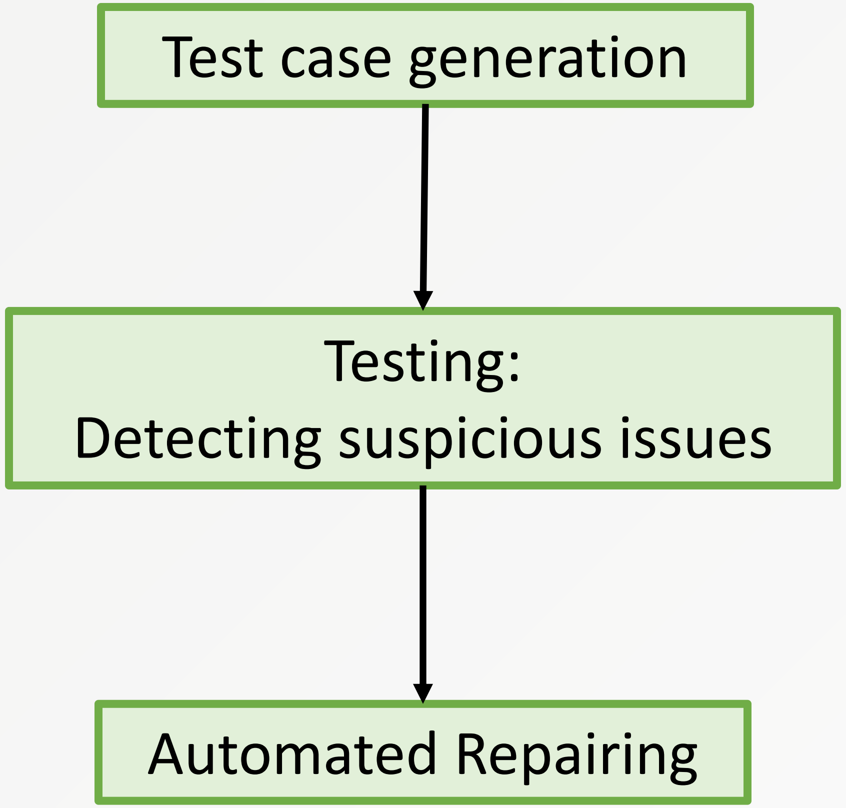
Flair-CONLL: <u>PERSON</u>, <u>ORGANIZATION</u>, <u>LOCATION</u>, MISCELLANEOUS NAMES

Flair-Ontonotes: <u>PERSON</u>, <u>ORGANIZATION</u>, <u>LOCATION</u>, CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE, LAW, MONEY, NORP, ORDINAL, PERCENT, PRODUCT, QUANTITY, TIME, WORK-OF-ART

Azure NER: <u>PERSON</u>, <u>ORGANIZATION</u>, <u>LOCATION</u>, PERSONTYPE, EVENT, PRODUCT, SKILL, ADDRESS, PHONENUMBER, EMAIL, URL, IP, DATETIME, QUANTITY

AWS NER: <u>PERSON</u>, <u>ORGANIZATION</u>, <u>LOCATION</u>, COMMERCIAL ITEM, DATE, EVENT, OTHER, QUANTITY, TITLE

Architecture of TIN



TIN: Detecting Potential NER Errors via *Metamorphic Testing*

Idea: NER predictions of the **same named entities** under **similar contexts** should be **identical**.

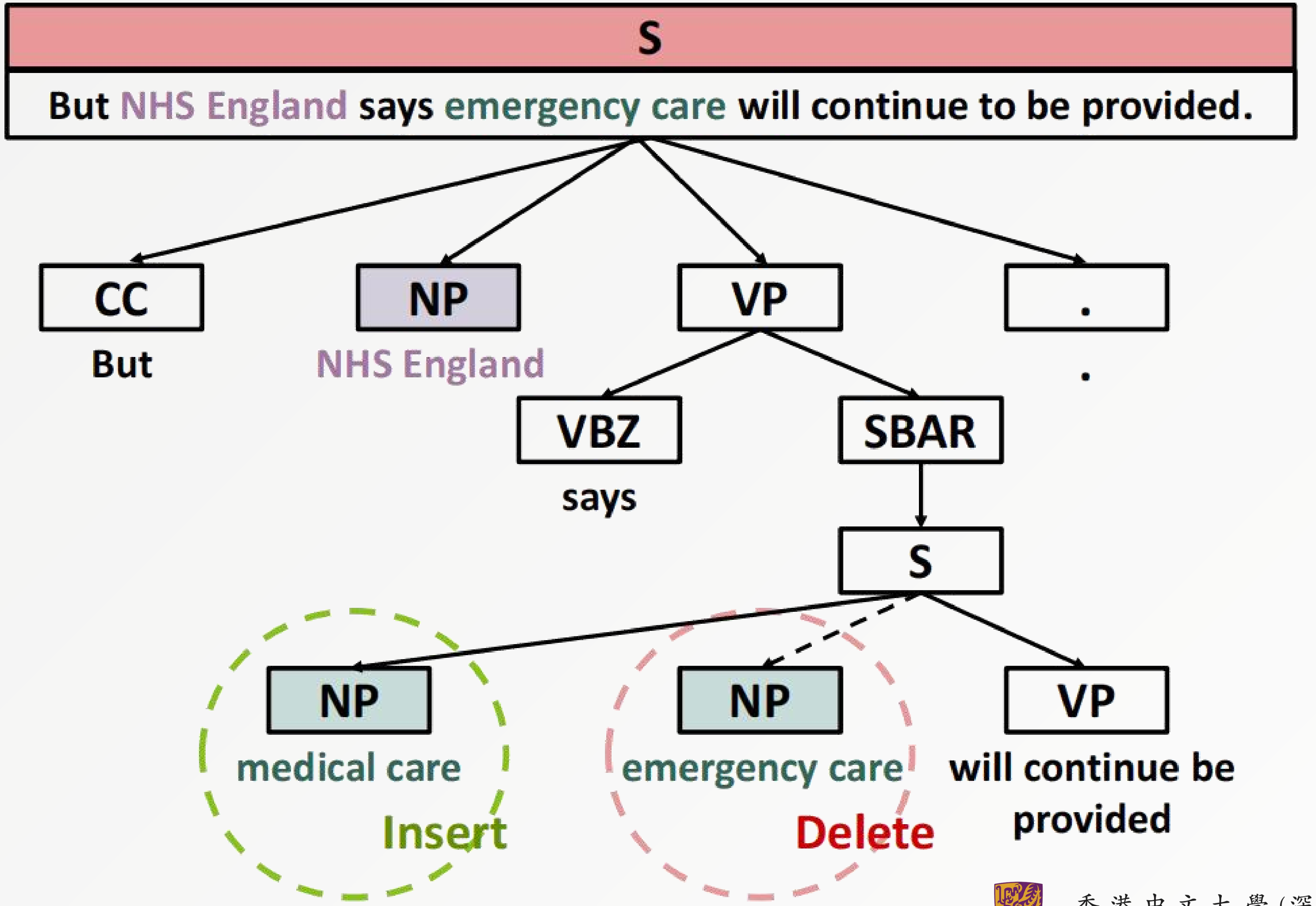
Generating *mutated sentences* with similar contexts

(1) Similar Sentence Generation:
Substituting the words or phrases in the sentences with the ones that have similar semantics.

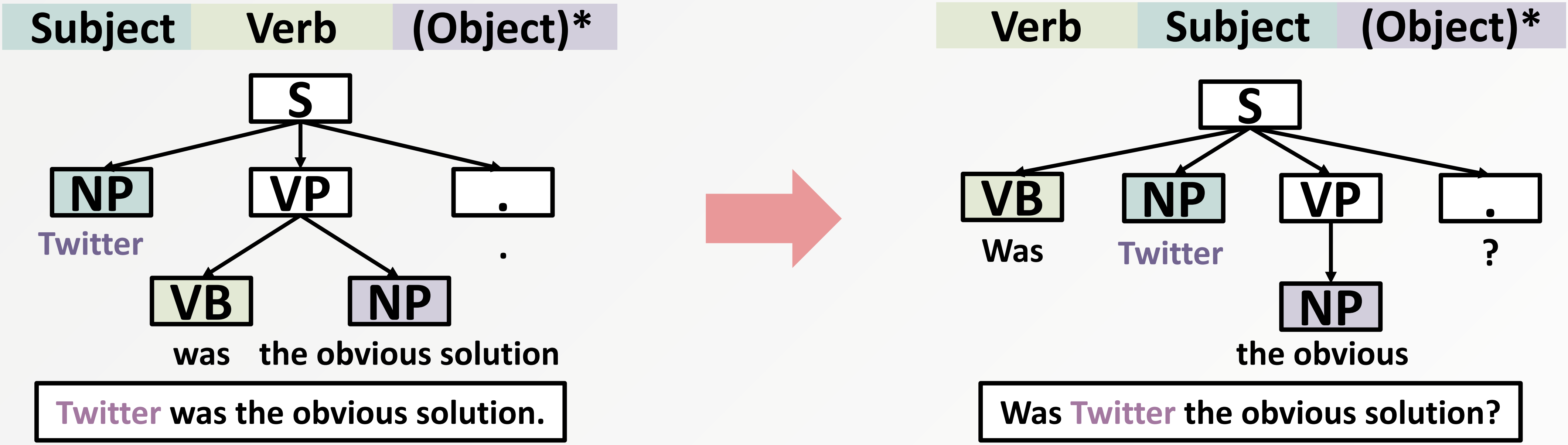
(2) Structure Transformation:
Transforming the declarative sentence into the interrogative sentence.

(3) Random NER Shuffle:
Randomly Shuffling the named entities with the same categories in the sentences.

Similar Sentence Generation via Constituency Parser



Structure Transformation via Constituency Parser



Random NER Shuffle

Spotify, Apple Music, and Deezer all said the track was their top performer of the year, beating competition from Ed Sheeran, Drake, and Taylor Swift.

NER

Spotify, Apple Music, Deezer : ORGNIZATION

Ed Sheeran, Drake, Taylor Swift : PERSON

Apple Music, Spotify, and Deezer all said the track was their top performer of the year, beating competition from Taylor Swift, Drake, and Ed Sheeran.

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen
数据科学学院
School of Data Science

Beyond Testing: Automated repairing NER systems

Idea: similar named entities should have the **same NER prediction** under the **same context**

Steps of Automated NER Repairing

(1) Suspicious Entity Location

(2) Equivalent Sentence Generation

(3) Relabeling NER Predictions

Suspicious Entity Location

BBC News is an operational business division of the BBC.

Named Entity: "BBC News" - LOC

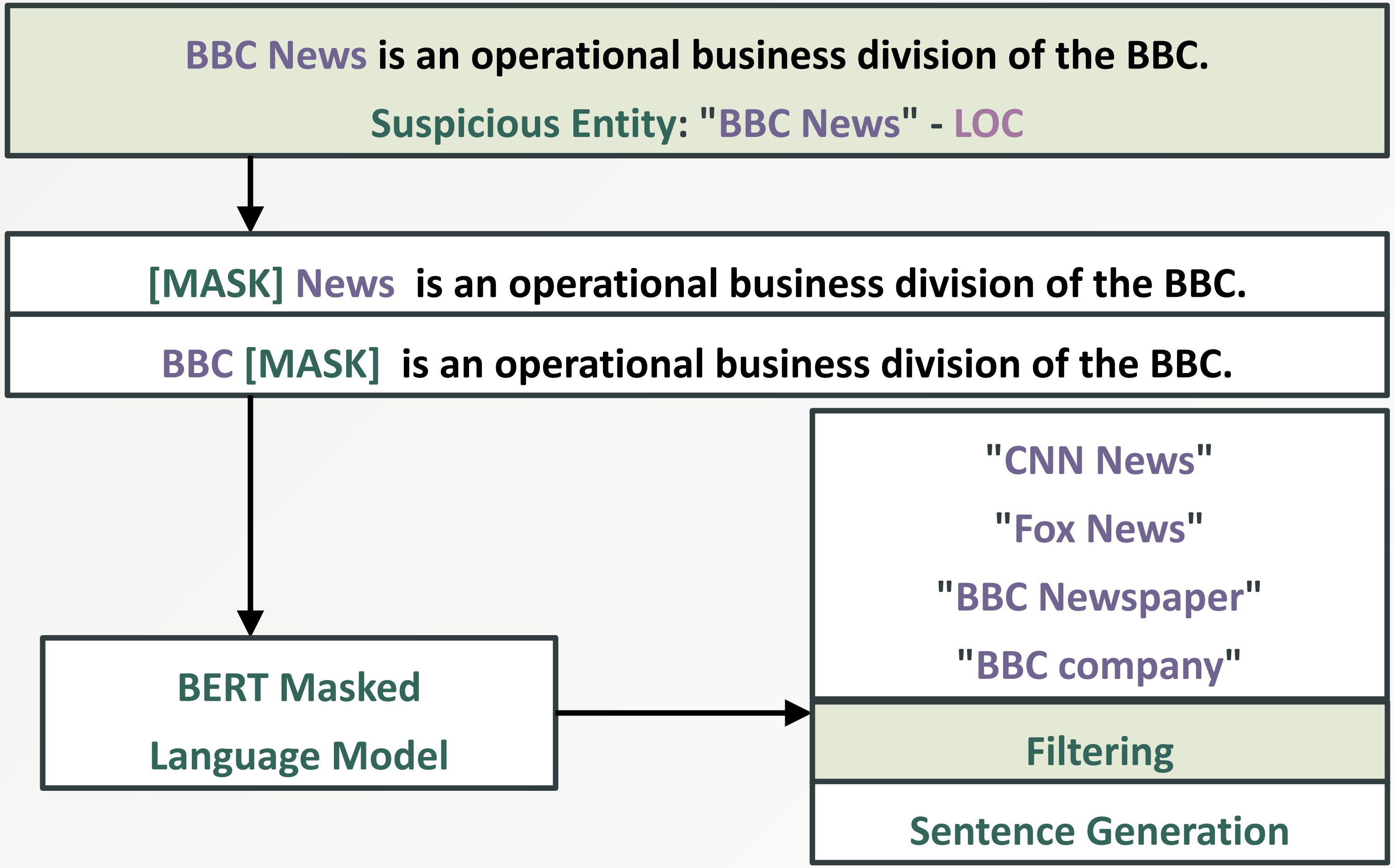
Is BBC News an operational business division of the BBC?

Named Entity: "BBC News" - PER

BBC News is an operational business division of the BBC.

Suspicious Entity: "BBC News" - LOC

Equivalent Sentence Generation



香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen
数据科学学院
School of Data Science

Relabeling NER Predictions

"CNN News"

"Fox News"

"BBC Newspaper"

"BBC company"

Filtering

Sentence Generation

CNN News is an operational business division of the BBC.

Fox News is an operational business division of the BBC.

BBC Newspaper is an operational business division of the BBC.

Evaluation Function

NER System

"CNN News" - ORG - 0.3

"Fox News" - ORG - 0.4

"BBC Newspaper" - MISC - 0.3

Relabeling NER Predictions

"CNN News" - ORG - 0.3

"Fox News" - ORG - 0.4

"BBC Newspaper" - MISC - 0.3

↓

ORG - 0.7,

MISC - 0.3

↓

Select the NER type with the maximum score

↓

Relabeled Named Entity

BBC News - ORG - 0.7

Evaluation Metric - Testing

$$\text{Precision} = \frac{\sum_{p_t \in P_T} \mathbb{1}\{error(p_t)\}}{|P_T|},$$

Evaluation - Testing

NER systems	TIN Overall
Flair-CoNLL	86.6% (161/186)
Flair-Ontonotes	85.0% (170/200)
Azure NER	93.0% (186/200)
AWS NER	93.4% (185/198)

Comparison with baseline - Testing

NER systems	TIN Overall
Flair-CoNLL	86.6% (161/186)
Flair-Ontonotes	85.0% (170/200)
Azure NER	93.0% (186/200)
AWS NER	93.4% (185/198)

Evaluation: four situations of NER relabeling

We use T and F to represent whether the NER prediction is correct or incorrect

- TF: Change true to false
- FT: Change false to true
- FF: Change false to false
- TT: Change true to true

Evaluation metrics of NER repairing

- *Err2Cor*: measures the probability of changing incorrect NER predictions to correct and is calculated as $Err2Cor = \frac{FT}{NumError}$
- Cor2Err: measures the probability of changing correct NER predictions to incorrect and is calculated as $Cor2Err = \frac{TF}{NumCorrect}$
- ErrorReduce: measures the ability to reduce NER errors and is calculated as $ErrorReduce = \frac{FT - TF}{NumError}$

Evaluation - Repairing

NER Systems	Err2Cor	Cor2Err	ErrorReduce
Flair-ConNLL	53.9%	14.4%	**40.4%**
Flair-Ontonotes	48.1%	19.5%	**26.8%**
AWS NER	55.2%	12.1%	**42.6%**
Azure NER	68.6%	17.1%	**50.6%**

Examples of NER Repairing

Error type	Omission
NER System	Flair-CoNLL
Sentence	Ben Johnson, from the Environmental Services Association (**ESA**), told BBC News…
Suspicious entities	“ESA”
Original Prediction	["Ben Johnson", PER] ["Environmental Services Association", ORG] ["BBC News", ORG]
Fixed Prediction	["Ben Johnson", PER] ["Environmental Services Association", ORG] ["BBC News", ORG] [**"ESA", ORG**]

Examples of NER Repairing

Error type	Over-labeling
NER System	Flair-Ontonotes
Sentence	The **halfway** point affords us an opportunity to step back and…
Suspicious entities	“halfway”
Original Prediction	**["halfway", CARDINAL]**
Fixed Prediction	**["halfway", NULL]**

Examples of NER Repairing

Error type	Incorrect Category
NER System	Azure NER
Sentence	They say the only positive thing the federal authorities have done is to return electricity to **Mekelle**.
Suspicious entities	“Mekelle”
Original Prediction	["authorities", PERSONTYPE] [**"Mekelle", PERSON**]
Fixed Prediction	["authorities", PERSONTYPE] [**"Mekelle", LOCATION**]

Examples of NER Repairing

Error type	Range Error
NER System	AWS NER
Sentence	Fibrus is delivering a similar scheme in Northern Ireland known as **Project Stratum**.
Suspicious entities	"Project"] ["Stratum"] ["Project Stratum"]
Original Prediction	["Fibrus", ORGANIZATION] ["Northern Ireland", LOCATION] [**"Project", OTHER**] [**"Stratum", TITLE**]
Fixed Prediction	["Fibrus", ORGANIZATION] ["Northern Ireland", LOCATION] [**"Project Stratum", OTHER**]

Conclusion

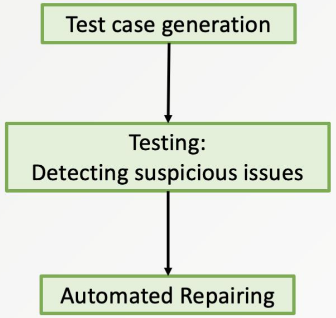
① NER Systems are NOT robust

Error Type	Sentence	Predicted Entities	Target Entities
Omission	Sir Paul's command of the stage is so casual that he makes it look easy (**Flair-Ontonotes**).	NULL	["Paul", PER]
Over-labeling	Elon Musk is having a similar effect on the platform (**Azure**).	["Elon Musk", PERSON] ["Platform", LOCATION]	["Elon Musk", PERSON]
Incorrect Category	Norrie believes securing Unesco status could offer new opportunities in sustainable tourism and branding of local produce, while at the same time highlighting the environmental value of the peatland (**Flair-Conll**).	["Norrie", PER] ["Unesco", MISC]	["Norrie", PER] ["Unesco", ORG]
Range Error	Det Supt Rance said the investigation remained active (**AWS**).	["Det", PERSON] ["Supt Rance", PERSON]	["Det Supt Rance", PERSON]

② Evaluation - Testing

NER systems	TIN Overall
Flair-CoNLL	86.6% (161/186)
Flair-Ontonotes	85.0% (170/200)
Azure NER	93.0% (186/200)
AWS NER	93.4% (185/198)

③ Architecture of TIN



④ Evaluation - Repairing

NER Systems	Err2Cor	Cor2Err	ErrorReduce
Flair-ConNLL	53.9%	14.4%	**40.4%**
Flair-Ontonotes	48.1%	19.5%	**26.8%**
AWS NER	55.2%	12.1%	**42.6%**
Azure NER	68.6%	17.1%	**50.6%**

Open Source on Github

FLOWCODE

PRIVACY.FLOWCODE.COM