



XỬ LÝ NGÔN NGỮ TỰ NHIÊN

CHƯƠNG 5: Sequential labeling

ThS. Lưu Thanh Sơn



NỘI DUNG

1. Định nghĩa bài toán.
2. Part of speech tagging
3. Named entity recognition
4. Markov chains
5. HMM Tagger
6. CRF
- 7. Độ đo đánh giá**



7. Các phương pháp đánh giá

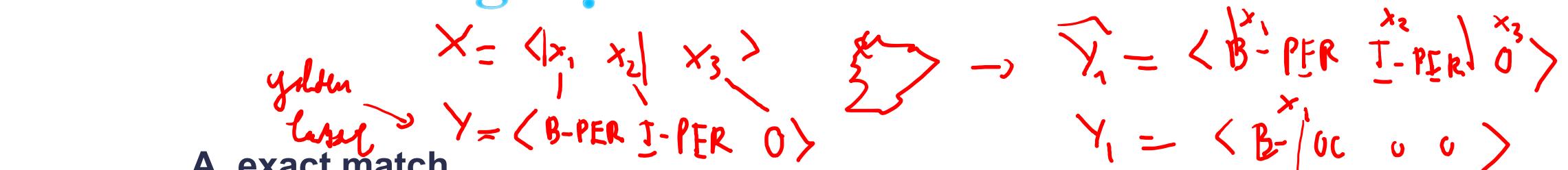


Metric đánh giá

- Đối với bài toán POS Tagging: độ đo chuẩn là *Accuracy*.
- Đối với bài toán NER recognition: Độ đo bao gồm: *recall*, *precision* và *F1-score*.



Các tình huống dự đoán đối với bài toán NER



- + Surface entity and type match (Both entity boundary and type are correct)
- + Type
- 1) Surface entity and type match (Both entity boundary and type are correct)
 - 2) System hypothesized an entity (predict entity that does not exist in ground truth)
 - 3) Systems miss an entity (entity exists in ground truth, but is not predicted by NER system)

B. partial match (overlapping)

- 4) Wrong entity type (correct entity boundary, type disagree)
- 5) Wrong boundaries (boundary overlap)
- 6) Wrong boundaries and wrong entity type



Tính toán precision và recall cho NER

- Trong CoNLL 2003 shared task: *precision is the percentage of named entities found by the learning system that are correct. Recall is the percentage of named entities present in the corpus that are found by the system. A named entity is correct only if it is an exact match of the corresponding entity in the data file.*
- Trong cuộc thi SemEval 2013 Task 9: Sử dụng độ đo MUC.

A

CoNLL 2003 Shared task: <https://www.aclweb.org/anthology/W03-0419>

SemEval 2013 Shared task: <https://www.aclweb.org/anthology/S13-2056.pdf>



Độ đo MUC-5

Các loại dự đoán:

- boundary*
- **Correct (COR)**: match
 - **Incorrect (INC)**: not match
 - **Partial (PAR)**: predicted entity boundary overlap with golden annotation, but they are **not the same**.
- types*
- **Missing (MIS)**: golden annotation boundary is not identified (predictions do not have, but golden label do)
 - **Spurious (SPU)**: predicted entity boundary does not exist in golden annotation (predictions have, but golden label do not)

<https://www.aclweb.org/anthology/M93-1007.pdf>



Độ đo MUC-5

Các tình huống dự đoán đối với nhãn:

- **Strict:** exact boundary surface string **match** and **entity type**.
- **Exact:** exact boundary match over the **surface string**, regardless of the type.
- **Partial:** partial boundary match over the **surface string**, regardless of the type.
- **Type:** some **overlap** between the system tagged entity and the **gold annotation** is required.



Gold-standard and predicted annotations

- Gold-standard (POS):

$$\underline{\text{POSSIBLE}}(\text{POS}) = \underline{\text{COR}} + \underline{\text{INC}} + \underline{\text{PAR}} + \underline{\text{MIS}} = \underline{\text{TP}} + \underline{\text{FN}}$$

- Predicted by NER system

$$\underline{\text{ACTUAL}}(\text{ACT}) = \underline{\text{COR}} + \underline{\text{INC}} + \underline{\text{PAR}} + \underline{\text{SPU}} = \underline{\text{TP}} + \underline{\text{FP}}$$



Tính toán precision/recall/f1 score

- Exact match (strict and exact)

$$\text{Precision} = \frac{COR}{ACT} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{COR}{POS} = \frac{TP}{TP+FN}$$

- Partial match (partial and type)

$$\text{Precision} = \frac{COR + 0.5 \times PAR}{ACT} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{COR + 0.5 \times PAR}{POS} = \frac{COR}{ACT} = \frac{TP}{TP+FP}$$

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$



Ví dụ

Table

Scenario	Golden Standard		System Prediction		Evaluation Schema			
	Entity Type	Surface String	Entity Type	Surface String	Type	Partial	Exact	Strict
III	brand	PIKOSYN			MIS	MIS	MIS	MIS
II			brand	healthy	SPU	SPU	SPU	SPU
V	drug	warfarin	drug	of warfarin	COR	PAR	INC	INC
IV	drug	propranolol	brand	propranolol	INC	COR	COR	INC
I	drug	phenytoin	drug	phenytoin	COR	COR	COR	COR
I	Drug	theophylline	drug	theophylline	COR	COR	COR	COR
VI	group	contraceptives	drug	oral contraceptives	INC	PAR	INC	INC



Tính precision/recall và F1 score

$$\text{Pos} = \text{COR} + \text{INC} + \text{PAR} + \text{MIS} \\ = 3 + 2 + 0 + 1 = 6$$

$$\text{ACT} = \text{COR} + \text{INC} + \text{PAR} + \text{SPU} \\ = 3 + 2 + 0 + 1 = 6$$

$$\text{Precision} = \frac{\text{COR} + 0.5 \times \text{PAR}}{\text{ACT}} = \frac{3+0}{6} = 0.5$$

$$\text{Recall} = \frac{\text{COR} + 0.5 \times \text{PAR}}{\text{Pos}} = \frac{3+0}{6} = 0.5$$

Partial M.I

Exact match (strict)

Measure	Type	Partial	Exact	Strict
Correct	3	3	3	2
Incorrect	2	0	2	3
Partial	0	2	0	0
Missed	1	1	1	1
Spurious	1	1	1	1
Precision	0.5	0.66	0.5	0.33
Recall	0.5	0.66	0.5	0.33
F1	0.5	0.66	0.5	0.33

$$\text{Pos} = 2 + 3 + 0 + 1 \\ = 6$$

$$\text{ACT} = 2 + 3 + 0 + 1 = 6$$

$$\text{Precision} = \frac{\text{COR} + \text{PAR}}{\text{ACT}} = \frac{2}{6} = \frac{1}{3}$$

$$\text{Recall} = \frac{\text{COR} + \text{PAR}}{\text{Pos}} = \frac{2}{6} = \frac{1}{3}$$



Tổng kết

1. Định nghĩa bài toán Sequence labeling: cho một chuỗi (sequence) X, tìm các nhãn (tags) ứng với từng phần tử trong chuỗi X.
2. Hai bài toán cụ thể:
 1. Gán nhãn từ loại (Part of speech tagging – POS tagging).
 2. Nhận diện thực thể có tên (Named entity recognition - NER).
3. Các phương pháp tiếp cận cho bài toán:
 1. HMM.
 2. CRF.
4. Các độ đo đánh giá:
 1. POS Tagging: độ đo Accuracy.
 2. NER: precision, recall và F1 score.

MVC-S