

# Розпізнавання іменованих сутностей

В роботі розглянуто застосування CRF (Conditional Random Fields) для розпізнавання іменованих сутностей на такі класи: PERSON, ORGANIZATION, LOCATION, MISC. Було проведено тестування на іспанському та нідерландському корпусі та отримані F1 score 0.75 та 0.72 відповідно використовуючи BI тегування. Було проаналізовано вплив різновидів тегування (BI, Pure Name, IL та інші) на результат (F1, Precision, Recall). Проаналізована стабільність features на двох мовах та вплив вибору комбінацій features на результат.

---

## Вступ

---

*Проблематика задачі.* Named Entity Recognition (NER) - це задача ідентифікації та класифікації імен людей, організації, місць та інших об'єктів в межах тексту. Вона займає центральне місце багатьох завдань НЛП.

Завдання розпізнавання іменованих сутностей (named entity recognition, NER), полягає в тому, що б виділити і класифікувати певні фрагменти тексту на заздалегідь відомих типах. Наприклад, на такі 4 типи:

- PERSON (люди)
- ORGANIZATION (організації)
- LOCATION (географічні об'єкти)
- MISC (різне, в широкому сенсі, включаючи, наприклад, події, твір мистецтва і національності)

Іменованими сутностями є фрази, які містять імена осіб, організації та місця. Наприклад:

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad].

Іншим більш не тривіальним є наступний приклад:

[PERSON Paris Hilton] visited the [LOCATION Paris] [ORGANIZATION Hilton]

Різні входження слова Paris відповідають імені та географічній назві.

Вирішення подібних багатозначних робить задачу розпізнавання іменованих сутностей складним завданням семантичної обробки текстів.

---

## Метрики

---

Точність:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Повнота:

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

F1 score - середнє гармонійне точності та повноти.

**Таблиця 1.** Features

Context word feature	Контекст
Word suffix	Суфікс слова
Word prefix	Префікс
LengthIF	Чи більша довжина слова деякої заданої костянти
First word	Чи є першим словом у реченні
Last word	Чи є останнім словом у реченні
Digit features	Чи міститься цифра у слові
Initial capital	Слово з великої букви
All capitals	Всі букви великі
All lowercase	Всі букви малі
Capitals mix	Мікс
Roman	Римські цифри
Single character	Лишє один символ
Part of Speech (POS)	Частина мови, суфікс частини мови

---

### Features

---

Було розглянути та досліджено такі набори feature функцій. [Таблиця 1]

---

### Модель

---

В якості моделі будемо використовувати CRF [1]. Його використовують в завданнях розпізнавання мови і образів, обробки текстової інформації, а також і в інших предметних областях: біоінформатики, комп’ютерної графіки та ін.

На сьогоднішній день саме метод CRF є найбільш популярним і точним способом для NER. Наприклад, він був реалізований в проекті Стенфордського університету Stanford Named Entity Recognizer.

---

### Корпус

---

Було використано корпус Conll2002, мова іспанська та голандська, дані корпуса у форматі: Слово – Pos-tag – BIO-teg.

---

### Реалізація

---

Для реалізації було використано мову Python та модуль pycrfsuite для CRF.

Були використані такі feature-функції:

- слово
- суфікс слова, 3 літери

- суфікс слова, 2 літери
- чи перша літера у верхньому регістрі
- чи слово складається з цифр
- чи початок речення
- чи кінець речення
- довжина слова
- слово зі всіма буквами в нижньому регістрі
- чи всі букви у верхньому регістрі
- чи слово в нижньому регістрі
- POS
- POS префікс, 2 літери

Було реалізовано перебір деяких комбінацій features.

Features	Len	Lower	isSupper	POS-tag	POS-tag 2	F1	Prec	Recall
1						0.71	0.74	0.69
2						0.71	0.73	0.69
3						0.71	0.73	0.69
4						0.72	0.75	0.71
5						0.71	0.74	0.70
6						0.72	0.74	0.71
7						0.72	0.74	0.71
8						0.73	0.75	0.72
9						0.71	0.73	0.70
10						0.72	0.74	0.70
11						0.72	0.74	0.70
12						0.71	0.74	0.70
13						0.73	0.75	0.71
14						0.73	0.75	0.72
15						0.74	0.75	0.73
16						0.74	0.75	0.73
17						0.72	0.74	0.70
18						0.71	0.73	0.70
19						0.72	0.74	0.70
20						0.72	0.74	0.70
21						0.74	0.75	0.73
22						0.73	0.75	0.72
23						0.74	0.76	0.73
24						0.74	0.75	0.73
25						0.72	0.74	0.71
26						0.72	0.74	0.70
27						0.72	0.75	0.71
28						0.72	0.75	0.71
29						0.73	0.75	0.72
30						0.73	0.74	0.72
31						0.74	0.76	0.73
32						0.74	0.75	0.73
						0.724	0.744	0.711

**Рис. 1.** Результати на різних наборах feature функцій на іспанській мові використовуючи BIO тегування.

Переміг тест, у якому використовуються всі функції. Проте це не завжди так, як ми можемо переконатися дивлячись на графік. Навчання і тестування проводиться на однаковому наборі feature функцій.

Проведемо більш детальний аналіз використовуючи графіки F1, Precision та Recall.

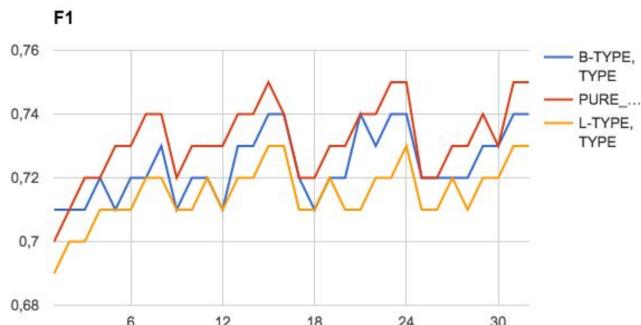


Рис. 2. Spain F1

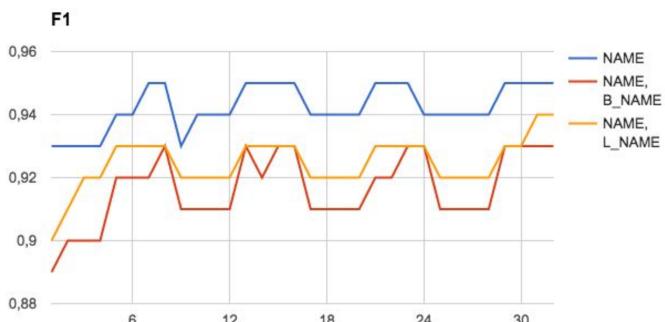


Рис. 3. Spain F1

Як бачимо зменшення кількості класів (перехід на чисті типи іменованих сущностей) покращує F1. Також більш ефективним є тегування Begin-Type, Inside-Type ніж Inside-Type, Last-Type, що пояснюється мовними особливостями побудови складених іменованих сущностей.

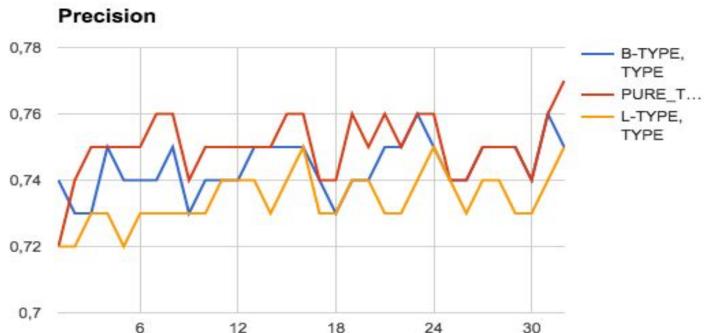


Рис. 4. Spain Precision

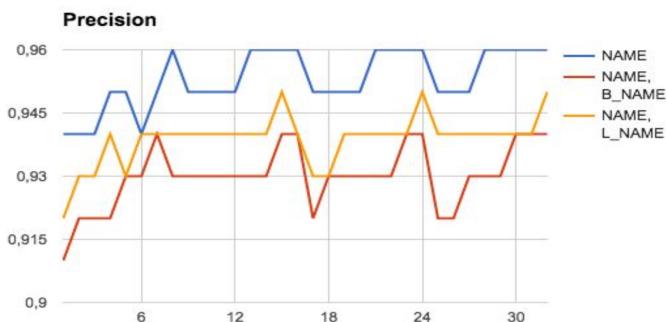


Рис. 5. Spain Precision

Схожу ситуацію можна спостерігати і на графіках точності та повноти.

Не можна не відмітити що в загальному найкращі результати дають тести з найбільшою кількістю feature функцій. Це можна пояснити доволі невеликою кількістю feature функцій - адже не виникає перенавчання моделей на несуттєвих змінних.

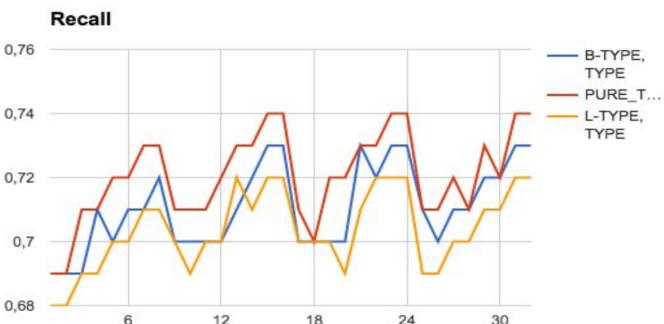


Рис. 6. Spain Recall

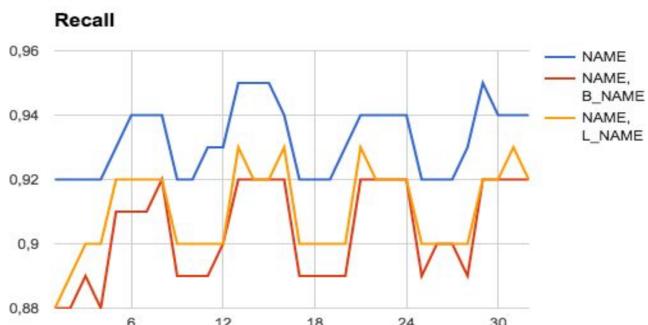


Рис. 7. Spain Recall

### Порівняння стабільності features на різних мовах

Можна помітити, що на тих самих наборах feature функцій зберігається характер функції F1 на різних мовах, тобто дані набори features є дійсно стабільними.

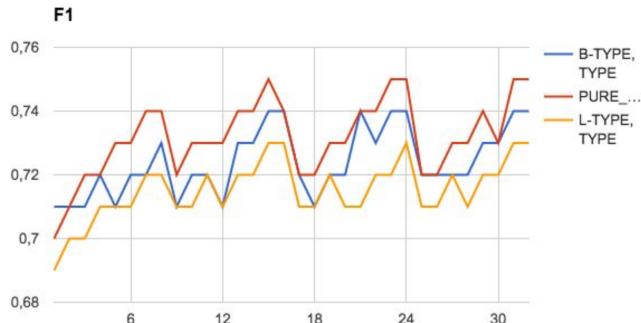


Рис. 8. Spain F1

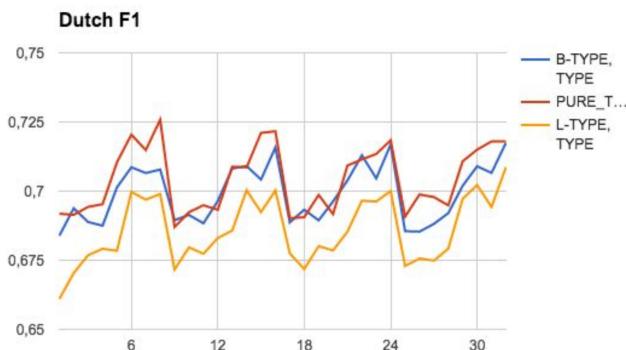


Рис. 9. Dutch F1

---

## Результати

---

В результаті дослідження було побудовані моделі, які дають змогу отримати F1 0.75 та 0.72 для іспанської та нідерландської мови відповідно.

Порівнюючи цей результат з результатами учасників змагання CoNLL-2002 [2] ми б входили в топ - 5 кращих результатів. Загалом було досліджено як комбінації feature функцій впливають на модель. Було показана їх стабільність на двох мовах, та проаналізовано вплив тегування на кінцевий результат.

---

## Автори

---

**Врублевський Віталій** — студент 4 курсу факультет кібернетики;  
E-mail: [vitaliyvrublevskiy@gmail.com](mailto:vitaliyvrublevskiy@gmail.com)

---

## Список літератури

---

- [1] J. Laerty, A. McCallum, F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the 18th International Conference on Machine Learning. 282-289. 2001.
  - [2] <http://www.cnts.ua.ac.be/conll2002/ner/>.
- 

## Додатки

---

Features	Len	Lower	isSupper	POS-tag	POS-tag 2	F1	Prec	Recall
1						0.70	0.72	0.69
2						0.71	0.74	0.69
3						0.72	0.75	0.71
4						0.72	0.75	0.71
5						0.73	0.75	0.72
6						0.73	0.75	0.72
7						0.74	0.76	0.73
8						0.74	0.76	0.73
9						0.72	0.74	0.71
10						0.73	0.75	0.71
11						0.73	0.75	0.71
12						0.73	0.75	0.72
13						0.74	0.75	0.73
14						0.74	0.75	0.73
15						0.75	0.76	0.74
16						0.74	0.76	0.74
17						0.72	0.74	0.71
18						0.72	0.74	0.70
19						0.73	0.76	0.72
20						0.73	0.75	0.72
21						0.74	0.76	0.73
22						0.74	0.75	0.73
23						0.75	0.76	0.74
24						0.75	0.76	0.74
25						0.72	0.74	0.71
26						0.72	0.74	0.71
27						0.73	0.75	0.72
28						0.73	0.75	0.71
29						0.74	0.75	0.73
30						0.73	0.74	0.72
31						0.75	0.76	0.74
32						0.75	0.77	0.74

**Рис. 10.** Результати на різних наборах feature функцій на іспанській мові використовуючи Type, O тегування.

Features	Len	Lower	isSupper	POS-tag	POS-tag 2	F1	Prec	Recall
1						0.93	0.94	0.92
2						0.93	0.94	0.92
3						0.93	0.94	0.92
4						0.93	0.95	0.92
5						0.94	0.95	0.93
6						0.94	0.94	0.94
7						0.95	0.95	0.94
8						0.95	0.96	0.94
9						0.93	0.95	0.92
10						0.94	0.95	0.92
11						0.94	0.95	0.93
12						0.94	0.95	0.93
13						0.95	0.96	0.95
14						0.95	0.96	0.95
15						0.95	0.96	0.95
16						0.95	0.96	0.94
17						0.94	0.95	0.92
18						0.94	0.95	0.92
19						0.94	0.95	0.92
20						0.94	0.95	0.93
21						0.95	0.96	0.94
22						0.95	0.96	0.94
23						0.95	0.96	0.94
24						0.94	0.96	0.94
25						0.94	0.95	0.92
26						0.94	0.95	0.92
27						0.94	0.95	0.92
28						0.94	0.96	0.93
29						0.95	0.96	0.95
30						0.95	0.96	0.94
31						0.95	0.96	0.94
32						0.95	0.96	0.94
					Середнє:	0.9425	0.953	0.932

**Рис. 11.** Результати на різних наборах feature функцій на іспанській мові використовуючи Name, O тегування.

Features	Len	Lower	isSupper	POS-tag	POS-tag 2	F1	Pre	Recall
1						0.89	0.91	0.88
2						0.90	0.92	0.88
3						0.90	0.92	0.89
4						0.90	0.92	0.88
5						0.92	0.93	0.91
6						0.92	0.93	0.91
7						0.92	0.94	0.91
8						0.93	0.93	0.92
9						0.91	0.93	0.89
10						0.91	0.93	0.89
11						0.91	0.93	0.89
12						0.91	0.93	0.90
13						0.93	0.93	0.92
14						0.92	0.93	0.92
15						0.93	0.94	0.92
16						0.93	0.94	0.92
17						0.91	0.92	0.89
18						0.91	0.93	0.89
19						0.91	0.93	0.89
20						0.91	0.93	0.89
21						0.92	0.93	0.92
22						0.92	0.93	0.92
23						0.93	0.94	0.92
24						0.93	0.94	0.92
25						0.91	0.92	0.89
26						0.91	0.92	0.90
27						0.91	0.93	0.90
28						0.91	0.93	0.89
29						0.93	0.93	0.92
30						0.93	0.94	0.92
31						0.93	0.94	0.92
32						0.93	0.94	0.92
					Середнє:	0.9165	0.93	0.94

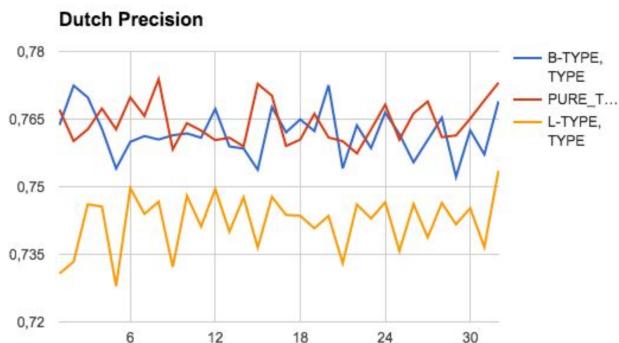
**Рис. 12.** Результати на різних наборах feature функцій на іспанській мові використовуючи Name, B-Name,O тегування. .

Features	Len	Lower	isSupper	POS-tag	POS-tag 2	F1	Pre	Recall
1						0.90	0.92	0.88
2						0.91	0.93	0.89
3						0.92	0.93	0.90
4						0.92	0.94	0.90
5						0.93	0.93	0.92
6						0.93	0.94	0.92
7						0.93	0.94	0.92
8						0.93	0.94	0.92
9						0.92	0.94	0.90
10						0.92	0.94	0.90
11						0.92	0.94	0.90
12						0.92	0.94	0.90
13						0.93	0.94	0.93
14						0.93	0.94	0.92
15						0.93	0.95	0.92
16						0.93	0.94	0.93
17						0.92	0.93	0.90
18						0.92	0.93	0.90
19						0.92	0.94	0.90
20						0.92	0.94	0.90
21						0.93	0.94	0.93
22						0.93	0.94	0.92
23						0.93	0.94	0.92
24						0.93	0.95	0.92
25						0.92	0.94	0.90
26						0.92	0.94	0.90
27						0.92	0.94	0.90
28						0.92	0.94	0.90
29						0.93	0.94	0.92
30						0.93	0.94	0.92
31						0.94	0.94	0.93
32						0.94	0.95	0.92
					Cереднє:	0.925	0.939	0.91

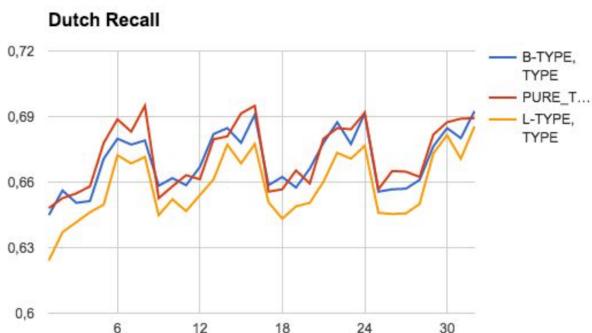
**Рис. 13.** Результати на різних наборах feature функцій на іспанській мові використовуючи Name, L-Name, O тегування.

Features	Len	Lower	isSupper	POS-tag	POS-tag 2	F1	Pre	Recall
1						0.69	0.72	0.68
2						0.70	0.72	0.68
3						0.70	0.73	0.69
4						0.71	0.73	0.69
5						0.71	0.72	0.70
6						0.71	0.73	0.70
7						0.72	0.73	0.71
8						0.72	0.73	0.71
9						0.71	0.73	0.70
10						0.71	0.73	0.69
11						0.72	0.74	0.70
12						0.71	0.74	0.70
13						0.72	0.74	0.72
14						0.72	0.73	0.71
15						0.73	0.74	0.72
16						0.73	0.75	0.72
17						0.71	0.73	0.70
18						0.71	0.73	0.70
19						0.72	0.74	0.70
20						0.71	0.74	0.69
21						0.71	0.73	0.71
22						0.72	0.73	0.72
23						0.72	0.74	0.72
24						0.73	0.75	0.72
25						0.71	0.74	0.69
26						0.71	0.73	0.69
27						0.72	0.74	0.70
28						0.71	0.74	0.70
29						0.72	0.73	0.71
30						0.72	0.73	0.71
31						0.73	0.74	0.72
32						0.73	0.75	0.72
					Середнє:	0,715	0,734	0,704

**Рис. 14.** Результати на різних наборах feature функцій на іспанській мові використовуючи L-Type, О тегування.



**Рис. 15.** Dutch Precision



**Рис. 16.** Dutch Recall