Журнал обчислювальної та прикладної математики **2016, №3(120)** Journal of Computational & Applied Mathematics

УДК 681.3
MSC 68T50

DEVELOPMENT OF NAMED ENTITY RECOGNITION SYSTEM

Oleksandr Marchenko

Faculty of Cybernetics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, E-mail: omarchenko@univ.kiev.ua.

РОЗРОБКА СИСТЕМИ РОЗПІЗНАВАННЯ ІМЕНОВАНИХ СУТНОСТЕЙ ТЕКСТУ

О. О. Марченко

Факультет кібернетики, Київський національний університет імені Тараса Шевченка, Київ, Україна, E-mail: omarchenko@univ.kiev.ua

Abstract. The article describes the developed named entity recognition system. To build a named entity classifier the two basic models of machine learning, The Naïve Bayes and Conditional Random Fields, were used. The paper describes a method for classifiers' training and the results of test experiments. Conditional Random Fields have proved most suitable for the description and effective solution to this problem.
Keywords: Named Entity Recognition, Natural Language Processing, Machine Learning, The Naive Bayes Model, Conditional Random Fields.

Резюме. У статті розглянуто розроблену систему розпізнавання іменованих сутностей тексту. Для створення блоків класифікації іменованих сутностей було застосовано дві базові моделі машинного навчання — наївна модель Баєса та модель умовних випадкових полів. В роботі описано метод навчання та результати експериментів з тестування побудованих класифікаторів. Модель умовних випадкових полів показала себе з найкращої сторони для опису та ефективного розв'язання поставленої задачі.
Ключові слова: Розпізнавання іменованих сутностей тексту, обробка текстів природною мовою, машинне навчання, наївна модель Баєса, умовні випадкові поля.

Вступ

Проблема визначення іменованих сутностей тексту не є новою, дослідження активно ведуться вже понад 20 років, і досягнуто високі результати роботи прикладних систем (до 93% точності у розпізнаванні іменованих сутностей машиною проти 96% точності у розпізнаванні іменованих сутностей людиною). Незважаючи на заявлений високий відсоток правильності

розпізнавання, проблема досі вважається відкритою і за даною проблематикою активно ведуться дослідження.

Актуальність проблеми пояснюється специфічністю середовища, в якому отримані надвисокі результати: як правило таке середовище створюється штучно для тестування системи і не може бути відтворено в реальному світі. До штучного середовища можна віднести додаткові 100% коректні дані про текст (наприклад, завжди гарантовано правильні синтаксичні дерева речень, морфологічна, семантична і т. д. інформація), які є недоступними в реальних умовах. Також до таких умов можна віднести надвисокі потужності задіяного обладнання, коли задача вирішується в лабораторних умовах на суперкомп'ютерах, та специфіку корпусів тестування. Наприклад, на тестові корпуси часто накладається умова обмеження словника іменованих сутностей до розміру словника навчальної вибірки, в таких умовах задача NER (named entity recognition — розпізнавання іменованих сутностей) зводиться до задачі розпізнавання сутностей за словником.

Через це різниця між заявленими в теорії та отриманими на практиці результатами є досить значною. Проведена оцінка найбільш популярних систем на ринку показала їх низьку ефективність. Більшість типів іменованих сутностей розпізнаються з точністю близько 60%–65%, що є недостатнім для ефективного використання в задачах аналізу текстів. Лише в деяких випадках реальна точність розпізнавання певних типів сутностей сягає 70%.

Дане дослідження було проведено з метою розробки придатного для промислового використання класифікатора, здатного розрізняти основні базові типи іменованих сутностей та ефективно працювати з реальними текстами поза межами лабораторного середовища, і видавати результати на рівні найкращих існуючих аналогів — state-of-the-art систем.

1. Система розпізнавання іменованих сутностей тексту

Основною задачею системи є розпізнавання у тексті іменованих сутностей та визначення типу цих сутностей. Вхідними даними системи є текст, написаний правильною англійською мовою з мінімальним вживанням сленгу та відсутністю орфографічних і граматичних помилок.

Архітектурно система складається з кількох ключових блоків, кожен блок виконує функції певного етапу побудови розв'язку задачі. Усі модулі попередньої обробки тексту для перетворення його у необхідний системі вигляд винесено за межі системи.

Система включає в себе:

- Блок ідентифікації та аналізу іменованих сутностей на основі Баєсівської моделі;
- Блок ідентифікації та аналізу іменованих сутностей на основі моделі умовних випадкових полів — Conditional random fields (CRF).

Ці блоки є підсистемами, які паралельно і незалежно одна від одної виконують обробку вхідного тексту:

1. ідентифікація синтаксичних груп, які містять іменовані сутності;

2. визначення меж знайдених іменованих сутностей (перше слово сутності– останнє слово сутності);

3. визначення типів знайдених іменованих сутностей.

Результатом роботи системи є текст з відповідною розміткою іменованих сутностей (id сутності, границі сутності, тип сутності).

Система налаштована для розпізнання типів іменованих сутностей (Type in system), кожен тип інтерпретується у відповідності до його трактування у корпусі Ontonotes:

Ontonotes Type	Description	Type in system
PERSON	People, including fictional	PER
ORGANIZATION	Companies, agencies, institutions	ORG
LOCATION	Locations, mountains, water bodies	LOC

Вхідними даними для розроблених класифікаторів є текст англійською мовою, дерева виведення та залежностей речень вхідного тексту, а також всі дані стосовно лексичних значень слів речень тексту згідно розмітки GOLD у корпусі Ontonotes.

Навчання класифікаторів на основі моделі Баєса та на основі моделі умовних випадкових полів — Conditional random field (CRF) проводилося на базі розміченого текстового корпусу Ontonotes. Так як Баєсівські класифікатори є відомим, розповсюдженим та досить простим методом, автор утримується від безпосереднього опису самої моделі Баєса та переходить до методу класифікації на основі умовних випадкових полів — Conditional random fields (CRF) [1].

2. Класифікатор на основі моделі умовних випадкових полів — Conditional random field (CRF)

Метод умовних випадкових полів — Conditional random fields (CRF), є аналогом методу Марківських випадкових полів (Markov random fields). Даний метод користується широкою популярністю у різних областях штучного інтелекту. Зокрема його успішно використовують в задачах розпізнавання мовлення та образів, в обробці текстової інформації, в комп'ютерній графіці та в інших задачах.

Марковським випадковим полем називають графову модель, яка використовується для представлення сумісних розподілів набору декількох випадкових змінних. Формально Марківське випадкове поле складається з:

- неорієнтованого графу або фактор-графу $G = (V, E)$, де кожна вершина $v \in V$ є випадковою змінною і кожне ребро $(u, v) \in E$ є залежністю між випадковими величинами u и v.
- набору потенціальних функцій (potential function) або факторів $\{\varphi_k\}$, одна для кожної кліки у графі(кліка — повний підграф неорієнтованого графу G). Функція φ_k ставить кожному можливому стану елементів клики у відповідність деяке невід'ємне дійсне число.

Вершини, що не є суміжними, мають відповідати умовно незалежним випадковим величинам. Група суміжних вершин формує кліку, набір станів вершин є аргументом відповідної потенціальної функції.

Сумісний розподіл набору випадкових величин $X = \{x_k\}$ у Марківському випадковому полі обчислюється за формулою:

$$P(x) = \frac{1}{Z} \prod_k \varphi_k(x_{(k)}),$$

де $\varphi_k(x_{(k)})$ — потенціальна функція, що описує стан випадкових величин у k-ій кліці; Z — коефіцієнт нормалізації, що обчислюється за формулою:

$$Z = \sum_{x \in X} \prod_k \varphi_k(x_{(k)}).$$

Множина вхідних слів-лексем $X = \{x_t\}$ та множина відповідних їм типів $Y = \{y_t\}$ у сукупності формують множину випадкових змінних $V = X \bigcup Y$. Для розв'язання задачі виділення інформації з тексту достатньо визначити умовну ймовірність $P(Y|X)$. Потенціальна функція має вигляд:

$$\varphi_k(x_{(k)}) = \exp(\sum_k \lambda_k f_k(y_t, y_{t-1}, x_t)),$$

де $\sum\{\lambda_k\}$ — дійснозначний параметричний вектор (множники Лагранжа), $\sum\{f_k(y_t, y_{t-1}, x_t)\}$ — набір ознакових функцій. Тоді лінійним умовним випадковим полем називається розподіл виду:

$$p(y|x) = \frac{1}{Z(x)} \prod_k \exp(\sum_k \lambda_k f_k(y_t, y_{t-1}, x_t)).$$

Коефіцієнт нормалізації $Z(x)$ обчислюється за формулою:

$$Z(x) = \sum_{y \in Y} \prod_k exp(\sum_k \lambda_k f_k(y_t, y_{t-1}, x_t)).$$

Обчислення моделі $p(y|x)$ відбувається як розв'язання оптимізаційної задачі з заданими обмеженнями [2] (різниця між спостереженням та його оцінкою має бути нульовою, і має виконуватися умова $\sum_{y \in Y} p(y|x) = 1$ по всім $x \in X$). На кожній ітерації заново обчислюються множники Лагранжа, обчислення проводиться з використанням алгоритмів „forward-backward“ та Вітербі.

Метод CRF, як і метод MEMM (Марківські Моделі Максимальної Ентропії), є дискримінативним імовірнісним методом, на відміну від генеративних методів, таких як приховані марківські моделі HMM та модель Баєса (Naïve Bayes).

По аналогії з марківськими моделями максимальної ентропії, вибір факторів-ознак для завдання імовірності переходу між станами при наявності спостереження значення залежить від специфіки конкретних даних, але, на відміну від того ж MEMM, CRF може враховувати будь-які особливості

та взаємозв'язки у вхідних даних. Вектор ознак $\bigwedge = \{\lambda_k\}$ обчислюється на основі навчальної вибірки та визначає вагу кожної потенціальної функції.

В умовних випадкових полях відсутня так звана label bias problem — ситуація, коли перевагу мають стани з меншою кількістю переходів, так як будується один єдиний розподіл ймовірностей та нормалізація (коефіцієнт $Z(x)$) виконується загалом, а не у рамках окремого стану. Це, безумовно, є перевагою методу: алгоритм не потребує припущення незалежності спостережних змінних. Крім того, використання довільних факторів дозволяє описати різноманітні ознаки об'єктів, що відчутно понижує вимоги до повноти та об'єму навчальної вибірки. При цьому точність буде визначатися не лише об'ємом вибірки, але й обраними факторами. Недоліком підходу CRF є обчислювальна складність аналізу навчальної вибірки, що ускладнює постійне оновлення моделі при отриманні нових навчальних даних. Слід відзначити високу швидкість роботи алгоритму CRF, що є дуже важливою перевагою при обробці великих об'ємів інформації.

3. Навчання моделі

Для навчання моделі був обраний корпус текстів Ontonotes [3], який містить достатній об'єм текстів, розмічених вручну. Розмітка текстів повністю відповідає задачі ідентифікації та аналізу іменованих сутностей, та обраним моделям машинного навчання. В рамках задачі аналізу іменованих сутностей тексти корпусу містять:

1. розмітку меж іменованих сутностей (перше слово сутності–останнє слово сутності);

2. розмітку типів знайдених іменованих сутностей (Людина, Організація, Локація).

Розмічені тексти містять синтаксичні структури речень — дерева виведення та дерева залежностей. Тобто доступними є межі синтаксичних груп речення та відношення залежностей між словами. Доступними є також повні лексичні значення слів речень (частина мови, рід, число, час для дієслів і т.д.). Алгоритми використовують також спеціальні словники імен, географічних назв та типових назв організацій для залучення додаткових знань у систему. Для формування базової множини ознакових функцій було проведено дослідження та аналіз кращих робіт по даній тематиці [4]–[6]. Було побудовано набір базових ознакових функцій, наприклад:

$$f_i(x,y) = \begin{cases} 1 & \text{якщо } y = \langle LOC \rangle, y \text{ з великої літери та } x = \text{„City“}; \\ 0, & \text{інакше.} \end{cases}$$

Далі в процесі дослідження були проведені чисельні експерименти по навчанню моделей на розмічених текстах корпусу Ontonotes, після чого виконувалося тестування навченого алгоритму на точність ідентифікації та визначення типу іменованих сутностей на текстах з інших частин корпусу. Потім, згідно процедури кросвалідації, навчальна та тестова частини

корпусу мінялися місцями та процес навчання та тестування моделей повторювався з початку. Із всіх отриманих оцінок точності обиралися мінімальні, як найбільш об'єктивні та гарантовано досяжні.

Навчання та тестування моделей проводилось багато разів з різними наборами ознакових функцій. В результаті проведення багатьох ітерацій етапів навчання–тестування з перебором множини ознакових функцій були визначені оптимальні набори ознакових функцій $\{f'_i\}$ та $\{f''_i\}$, на яких було досягнуто максимальні оцінки точності ідентифікації та визначення типів іменованих сутностей тексту класифікатором Баєса та класифікатором на базі моделі умовних випадкових полів (CRF), відповідно.

4. Отримані результати

У таблицях нижче продемонстровано фінальні оцінки роботи класифікатору Баєса та класифікатору на основі моделі умовних випадкових полів(CRF), навчених на оптимальних наборах ознакових функцій $\{f'_i\}$ та $\{f''_i\}$, відповідно. Надано оцінки точності (Precision, P), повноти (Recall, R) та комбінована міра F1:

$$F_1 = \frac{2^* Precision^* Recall}{Precision + Recall}.$$

Проведене дослідження показало, що в розв'язанні даної задачі очевидною є значна перевага моделі умовних випадкових полів (CRF) над Баєсівською моделлю. Припущення незалежності ознак, що є визначною властивістю у наївній Баєсівській моделі, не задовільняє природі задачі класифікації іменованих сутностей тексту.

В комп'ютерній лінгвістиці є ряд задач, в яких метод Баєса демонструє високі оцінки точності та достатню ефективність. Часто в них використовується найпростіша модель представлення тексту „торба слів“(„bag of words“). Для задач, де достатньо найпростіших моделей представлення текстових документів, і за умови, що у вхідних текстах розподіл імовірностей дійсно наближений до рівномірного, метод Баєса підходить набагато краще, ніж для класифікації іменованих сутностей тексту. Це пов'язано із тим, що при аналізі іменованих сутностей потрібно враховувати складний контекст слів, який часто виражається одразу декількома складними взаємозалежними функціями ознак. Тому умовні випадкові поля (CRF), які є набагато більш складною та гнучкою моделлю, більш точно описують та ефективно розв'язують дану задачу.

Отримані результати дослідження, а також детальний аналіз задачі та існуючих програмних реалізацій для її розв'язання підводить до висновку про найкращу придатність саме моделі умовних випадкових полів (CRF) для розробки систем класифікації іменованих сутностей тексту. Наприклад, іноді ця модель навіть дещо поступається за точністю Марківським моделям максимальної ентропії MEMM [7], але при цьому значно переважає у повноті і, як результат, переважає за комбінованою мірою F1.

Табл. 1. Оцінки класифікатора Баєса на підкорпусі Web text (230 файлів)

	Precision	**Recall**	**F1**
LOCATION	0,5423	0,6527	0,5924
ORGANIZATION	0,0412	0,0350	0,0379
PERSON	0,3311	0,6127	0,4299
Total	**0,3450**	**0,4954**	**0,4067**

Табл. 2. Оцінки класифікатора Баєса на підкорпусі Newswire (1665 файлів)

	Precision	**Recall**	**F1**
LOCATION	0,6498	0,8501	0,7365
ORGANIZATION	0,5022	0,7482	0,6010
PERSON	0,6673	0,8388	0,7433
Total	**0,5813**	**0,8003**	**0,6734**

Табл. 3. Оцінки класифікатора на основі умовних випадкових полів(CRF)

Підкорпуси			
Web text	**Broadcast News**	**Newswire**	**Total**
LOC			
Precision: 0,8679	Precision: 0,9283	Precision: 0,9198	Precision: 0,9395
Recall:0,9323	Recall: 0,9530	Recall: 0,9190	Recall: 0,9369
F1: 0,8989	F1: 0,9405	F1: 0,9194	F1: 0,9382
ORG			
Precision: 0,7939	Precision: 0,8118	Precision: 0,8810	Precision: 0,8858
Recall: 0,7324	Recall: 0,7768	Recall: 0,8863	Recall: 0,8830
F1: 0,7619	F1: 0,7939	F1: 0,8836	F1: 0,8844
PER			
Precision: 0,9157	Precision: 0,8910	Precision: 0,9104	Precision: 0,9207
Recall: 0,9104	Recall: 0,9185	Recall: 0,8895	Recall: 0,9104
F1: 0,9130	F1: 0,9045	F1: 0,8998	F1: 0,9155
TOTAL			
Precision: 0,8647	Precision: 0,8909	Precision: 0,9008	Precision: 0,9140
Recall: 0,8638	Recall: 0,9029	Recall: 0,8974	Recall: 0,9092
F1: 0,8643	F1: 0,8968	F1: 0,8991	F1: 0,9116

Саме модель умовних випадкових полів була реалізована в проекті Стендфордського университету Stanford Named Entity Recognizer [8].

Отримані в результаті тестування розробленої системи оцінки точності та повноти (таблиця 3) демонструють найвищі значення на рівні найкращих існуючих світових аналогів. На тестових текстах корпусу Ontonotes розроблена система змогла перевершити показники Stanford Named Entity Recognizer. Це було досягнуто завдяки успішно проведеній оптимізації набору ознакових функцій, що дало змогу отримати максимально високі оцінки точності.

Висновки

На основі двох базових моделей машинного навчання — наївної моделі Баєса та умовних випадкових полів, — було розроблено систему ідентифікації та аналізу іменованих сутностей тексту. Результати дослідження показали високу якість роботи класифікатора, реалізованого на основі умовних випадкових полів. Досвід найкращих існуючих програмних реалізацій систем аналізу іменованих сутностей тексту приводить до висновку, що саме модель умовних випадкових полів (CRF) оптимально підходить для розробки класифікаторів іменованих сутностей.

В процесі тестування розроблена система продемонструвала високу точність визначення типів іменованих сутностей тексту на рівні найкращих існуючих світових аналогів.

Література

1. Lafferty J., McCallum A., Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data // The 18th International Conference on Machine Learning, June 28–July 1, 2001. Proceedings — Williamstown, MA, USA, 2001. — P. 282–289.
2. Klinger R., Tomanek K. Classical Probabilistic Models and Conditional Random Fields // Algorithm Engineering Report TR07-2-013, Department of Computer Science, Dortmund University of Technology, December 2007. ISSN 1864-4503.
3. Text Corpus Ontonotes. — https://catalog.ldc.upenn.edu/LDC2011T03
4. Turian J., Ratinov, L., Bengio, Y. Word representations: a simple and general method for semi-supervised learning // The 48th Annual Meeting of the Association for Computational Linguistics, July 11–16, 2010. Proceedings — Uppsala, Sweden, 2010. — P. 384–394.
5. Nadeau D, Sekine S. A survey of named entity recognition and classification // Lingvisticae Investigationes. — 2007. — 30 (1). — P. 3–26.
6. Nadeau D., Turney P., Matwin S. Unsupervised Named Entity Recognition: Generating Gazetteers and Resolving Ambiguity // Canadian Conference on Artificial Intelligence — 2006, June 7—9, 2006. Proceedings — Quebec, Canada, 2006. — P. 266–277.
7. Антонова А. Ю., Соловьев А. Н. Метод условных случайных полей в задачах обработки русскоязычных текстов // Информационные технологии и системы — 2013. Труды международной научной конференции. Калининград. Россия. — 2013. — С. 321–325.
8. Stanford Named Entity Recognizer. — http://www-nlp.stanford.edu/software/

Надійшла 02.10.2015