УДК 004.85

МАШИННО-НАВЧАЛЬНІ МЕТОДИ РОЗПІЗНАВАННЯ ІМЕНОВАНИХ СУТНОСТЕЙ ТЕКСТУ

О.О. Марченко

У статті розглянуто машинно-навчальні методи розпізнавання іменованих сутностей тексту. Розглянуто дві базові моделі машинного навчання – наївна модель Байєса та модель умовних випадкових полів, застосовані для вирішення задачі ідентифікації та аналізу іменованих сутностей. Також досліджено модель, в якій для мультикласифікації іменованих сутностей текстів використовуються корегуючі вихідні коди. В роботі описано процес навчання та результати експериментів з тестування побудованих класифікаторів. Умовні випадкові поля перевершили інші моделі за оцінками точності та надійності роботи методу.
Ключові слова: машинне навчання, обробка природної мови, розпізнання іменованих сутностей тексту.

В статье исследуются машинно-обучаемые методы распознавания именованных сущностей текста. Рассмотрены две базовые модели машинного обучения – наивная модель Байеса и модель условных случайных полей, которые были использованы для решения задачи идентификации и анализа именованных сущностей. Также исследована модель, в которой для мультиклассификации именованных сущностей текстов используются корректирующие выходные коды. В работе описаны процесс обучения и результаты экспериментов по тестированию построенных классификаторов. Условные случайные поля превзошли другие модели по оценкам точности и надежности работы метода.
Ключевые слова: машинное обучение, обработка естественного языка, распознавание именованных сущностей текста.

The article describes machine learning methods for the named entity recognition. To build named entity classifiers two basic models of machine learning, The Naïve Bayes and Conditional Random Fields, were used. A model for multi-classification of named entities using Error Correcting Output Codes was also researched. The paper describes a method for classifiers' training and the results of test experiments. Conditional Random Fields overcome other models in precision and recall evaluations.
Key words: machine learning, natural language processing, named entity recognition.

Вступ

Проблема визначення іменованих сутностей тексту не є новою, дослідження активно ведуться вже понад 20 років, і оприлюднені досить високі результати роботи прикладних систем (до 93 % точності у розпізнаванні іменованих сутностей машиною проти 96 % точності у розпізнанні 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 та інша 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.

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 структурно складається з наступних блоків:

- блок ідентифікації та аналізу іменованих сутностей на основі Байєсівської моделі;

- блок ідентифікації та аналізу іменованих сутностей на основі моделі умовних випадкових полів − Conditional random field (CRF).

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

– ідентифікація синтаксичних груп речень тексту, які містять іменовані сутності;

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

– визначення типів знайдених іменованих сутностей.

Підсистеми виконують дану обробку тексту з відповідною розміткою.

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

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

Ontonotes Type	**Description**	**Type in system**
PERSON	People, including fictional	PERSON
ORGANIZATION	Companies, agencies, institutions, etc.	ORGANIZATION
LOCATION	Locations, mountain ranges, bodies of water	LOCATION

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

Класифікатор на основі моделі умовних випадкових полів − Conditional random fields

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

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

- неорієнтований граф або фактор-граф $G = (V, E)$, де кожна вершина $v \in V$ – випадкова змінна X і кожне ребро $(u, v) \in E$ – залежність між випадковими величинами u и v;

- набір потенційних функцій (potential function) або факторів $\{\varphi_k\}$, одна для кожної кліки у графі (кліка − повний підграф G неорієнтованого графу). Функція φ_k ставить кожному можливому стану елементів кліки 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 потенційної функції.

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

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

де $\varphi_k (x_{\{k\}})$ – потенційна 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 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($Y \mid X$). Потенційна 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:

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

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 $p(y \mid x)$ відбувається як розв'язання оптимізаційної задачі з заданими обмеженнями [2] (різниця між спостереженням та його оцінкою має бути нульовою та має виконуватися умова

$$\sum_{y \in Y} p(y \mid x) = 1 \text{ по всім } x \in X .$$

На кожній ітерації заново обчислюються множники Лагранжа, обчислення проводиться з використанням традиційних алгоритмів – «forward-backward» та Вітербі.

Метод CRF, як і метод марковські моделі максимальної ентропії (MMME), є дискримінативним імовірнісним методом, на відміну від генеративних методів, таких як приховані марковські моделі HMM та модель Байєса (Naïve Bayes).

За аналогією з марковськими моделями максимальної ентропії, вибір факторів-ознак для завдання імовірності переходу між станами при наявності спостереження значення x_t залежить від специфіки конкретних даних, але на відміну від того ж MMME, CRF може враховувати будь-які особливості та взаємозв'язки у вхідних даних. Вектор ознак $\Lambda = \{\lambda_k\}$ обчислюється на основі навчальної вибірки та визначає вагу кожної потенційної 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 імовірностей та нормалізація (коефіцієнт Z(x)) виконується загалом, а не у рамках окремого стану. Це, безумовно, є перевагою метода: алгоритм не потребує припущення незалежності спостережних змінних. Крім того, використання довільних факторів дозволяє описати різноманітні ознаки об'єктів, що знижує вимоги до повноти та обсягу навчальної вибірки. При цьому точність буде визначатися не лише обсягом 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 обсягів інформації.

Навчання моделі

Для навчання моделі був обраний корпус текстів Ontonotes [3], який містить достатній обсяг 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 розмітку:

- задання меж іменованих сутностей (перше слово сутності – останнє слово сутності);
- задання 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 найкращих робіт за даною тематикою [4–6]. Побудовано набор базових ознакових функцій, наприклад:

$$f_i(x, y) = \begin{cases} 1, \_ \text{якщо } \_ y = \langle LOC \rangle, y \_ \text{починається\_з\_великої\_літери}, x = "City", \\ 0, \_ \text{інакше}. \end{cases}$$

Далі в процесі дослідження були проведені чисельні експерименти з навчання 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 і 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 функцій ознак були визначені оптимальні набори ознакових функцій $\{ f_i' \}$ та $\{ f_i'' \}$, на яких досягнуто максимальні оцінки точності ідентифікації та визначення типів іменованих сутностей тексту класифікатором Байєса та класифікатором на базі моделі умовних випадкових полів (CRF), відповідно.

Розпізнавання іменованих сутностей тексту з використанням корегуючих вихідних кодів (ECOC)

Для вирішення задачі визначення іменованих сутностей у тексті як альтернативний підхід були використані корегуючі вихідні коди (Error-Correcting Output Codes, ECOC). Даний підхід застосовують при вирішенні задач мультикласифікації, коли число класів перевищує два. У випадку визначення іменованих сутностей як класи маємо класи слів, такі як Person, Location, Organization, Event, Product та інші. Також в іншій серії експериментів використовувалась розмітка на класи з використанням boundary-тегів, в цьому випадку маємо наступні класи: Person-Begin, Person-Inside, Location-Begin, Location-Inside, Organization-Begin, Organization-Inside та інші.

Задача мультикласифікації полягає у знаходженні невідомої функції $f(x)$, область значень якої дискретна множина, що містить k значень (класів), $k > 2$. Дана функція $f(x)$ визначається у процесі навчання на основі навчальної вибірки прикладів виду (x_i, d_i), $i = \overline{1,n}$, де $d_i = f(x_i)$ – відоме значення класу для прикладу x_i.

Вирішення задачі мультикласифікації зводиться до розв'язання підзадач бінарної класифікації, а результатом мультикласифікації є поєднання отриманих розв'язків. Для поєднання розв'язків бінарних класифікаторів було застосовано підхід розподіленого вихідного представлення (Distributed Output Representation); як бінарні класифікатори використовуються класифікатори CRF.

Під *розподіленим вихідним представленням* розуміється задання кожного класу бінарним рядком довжини n – "кодовим словом". Кожен біт кодового слова відповідає окремому бінарному класифікатору, який навчається. Вирішення задачі мультикласифікації зводиться до обробки так званої матриці кодових слів, рядки якої – кодові слова, що відповідають класам, об'єкти яких розпізнаються, а стовпчики відповідають бінарним класифікаторам (це ті значення, що видають класифікатори на відповідних класах). Після навчання класифікаторів новий об'єкт x класифікується оцінюванням кожного з n бінарних класифікаторів для отримання n-бітового кодового слова. Отримане кодове слово об'єкта x порівнюється з кожним із k кодових слів матриці. Об'єкт x належить класу, чиє кодове слово є найближчим згідно вибраної метрики до його власного слова. Визначення мінімальної відстані від отриманого кодового слова об'єкта x, що класифікується, до одного з кодових слів матриці розглядається як процес *декодування*. Для реалізації процесу декодування використовується відстань Хемінга. Зокрема, мінімальна відстань між отриманим кодовим словом $f(x) = (f_1(x), f_2(x), \ldots, f_n(x))$ та кодовими словами матриці M визначається як число позицій, у яких відповідні значення бітів різні.

Від виду матриці кодових слів залежить скільки помилок здатен виправити даний розподілений вихідний код у процесі декодування. Коди, які дозволяють виправити помилки в процесі декодування називаються *корегуючі вихідні коди (*Error-Correcting Output Codes*)*. Мірою якості коду є мінімальна відстань Хемінга між парами кодових слів матриці. Якщо мінімальна відстань Хемінга дорівнює d, відповідний код гарантовано може виправити $(d-1)/2$ помилкових біт при декодуванні.

Моделі розподіленого вихідного коду будуються відповідно до різних представлень матриці кодових слів $M \in \{0,1\}^{k \times n}$, де k – кількість класів, n – кількість бінарних класифікаторів, тобто довжина кодового слова.

У рамках досліджень була використана модель корегуючих вихідних кодів *Exhaustive Code.* Згідно даної моделі рядками матриці кодових слів є кодові слова довжини $2^{k-1} - 1$. Перший рядок матриці заповнюється одиницями, далі i-й рядок матриці заповнюється 2^{k-i} нулями та 2^{k-i} одиницями, що чередуються, починаючи з нулів. Приклад матриці з вичерпним кодом для 4-х класів наведено у табл. 1.

Таблиця 1. Вичерпний код для 4-х класів

Клас	Кодові слова						
	f_1	f_2	f_3	f_4	f_5	f_6	f_7
C_1	1	1	1	1	1	1	1
C_2	0	0	0	0	1	1	1
C_3	0	0	1	1	0	0	1
C_4	0	1	0	1	0	1	0

При дослідженні було проведено наступні експерименти. Для розпізнавання іменованих сутностей класів Person, Location, Organization було залучено також класи слів NE (куди відносяться всі інші сутності, які не належать до Person, Location, Organization, наприклад, сутності класів Event, Product, WorkArt, Money тощо), а також OTHER, які включають всі інші лексеми, які не відносяться до переліку іменованих сутностей. Таким чином у даному експерименті було використано 5 класів. Матриця кодових слів у цьому випадку містить 15 стовпчиків (бінарних класифікаторів).

У другому експерименті були залучені спеціальні boundary-тегі для формування класів, таким чином використовувалися наступні 8 класів: Person-Begin, Person-Inside, Location-Begin, Location-Inside, Organization-Begin, Organization-Inside, NE та OTHER. Матриця кодових слів у цьому випадку складається з 127 бінарних класифікаторів.

Отримані результати дозволили зробити наступні висновки. У результаті надвеликої кількості лексем, які відносяться до класу OTHER, та мають велику частоту вживання у корпусі, а також у результаті того, що деякі ознаки сутностей (досить великий їх відсоток), що належать до основних класів (не до OTHER), при формуванні бінарного класифікатора потрапляють в один клас, то бінарні класифікатори типу CRF на таких нерівномірних вибірках показали невисоку якість роботи. Наприклад, при розділенні на два класи отримаємо, що до одного класу належать високочастотні лексеми з OTHER та низькочастотні з Event, Product, Location тощо, а до другого класу – виключно низькочастотні лексеми з Person та Organization. Тоді елементи першого класу мають і високу частоту, і значну частину ознак другого класу, що призводить до значного превалювання першого класу над другим. За рахунок великої кількості помилок, отриманих бінарними класифікаторами (слід врахувати, що кожен з 15, у випадку 5 класів, та кожен із 127, у випадку 8 класів, має великий відсоток помилок) застосування розподілених вихідних кодів, зокрема ECOC, не дозволило отримати бажані високі оцінки якості. Для покращення результатів роботи моделі з використанням ECOC необхідно мати навчальну вибірку з більш рівномірним розподілом лексем по класах. Проте, використання підходу ECOC (а саме матриці кодових слів та реалізації процесу декодування за допомогою обчислення відстаней Хемінга) виправдане як одного з потенційних способів, коли треба знайти лексеми, що можливо були віднесені не до свого класу.

Отримані результати

У таблицях 2–5 представлені оцінки роботи класифікатора Байєса та класифікатора на основі моделі умовних випадкових полів (CRF), навчених на оптимальних наборах ознакових функцій $\{f_i'\}$ та $\{f_i''\}$, відповідно. У таблиці 6 надані оцінки роботи мультикласифікатора, побудованого з використанням корегуючих вихідних кодів (ECOC).

В експериментах обчислювалися оцінки точності (Precision, P), повноти (Recall, R) та комбінована міра F_1:

$$F_1 = \frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})}.$$

Таблиця 2. Оцінки класифікатора Байєса на підкорпусі Broadcast News (100 файлів)

	Precision	**Recall**	**F1**
LOCATION	0,8242	0,7881	0,8057
ORGANIZATION	0,2552	0,4301	0,3203
PERSON	0,5188	0,9047	0,6594
Total	**0,5493**	**0,7868**	**0,6469**

Таблиця 3. Оцінки класифікатора Байєса 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**

Таблиця 4. Оцінки класифікатора Байєса 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**

Таблиця 5. 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

Таблиця 6. Оцінки мультикласифікатора з використанням корегуючих вихідних кодів (ECOC)

Корпус		
WB	**BN**	**NW**
LOC		
Precision: 0.8179	Precision: 0.7328	Precision: 0.8271
Recall: 0.6547	Recall: 0.8012	Recall: 0.8113
F1: 0.7273	F1: 0.7655	F1: 0.8191
ORG		
Precision: 0.5378	Precision: 0.7637	Precision: 0.7734
Recall: 0.3792	Recall: 0.6354	Recall: 0.7422
F1: 0.4448	F1: 0.6937	F1: 0.7575
PER		
Precision: 0.7473	Precision: 0.7804	Precision: 0.8530
Recall: 0.5509	Recall: 0.8567	Recall: 0.8037
F1: 0.6342	F1: 0.816772	F1: 0.8276
TOTAL		
Precision: 0.7253	Precision: 0.7590	Precision: 0.8178
Recall: 0.5420	Recall: 0.7644	Recall: 0.7857
F1: 0.6204	F1: 0.7617	F1: 0.8015

Оцінки точності та повноти, отримані в результаті тестування розробленої системи на базі моделі CRF (табл. 5), демонструють найвищі значення на рівні найкращих існуючих світових аналогів. На тестових текстах корпусу Ontonotes розроблена система змогла перевершити показники відомої системи Стенфордського університету для розпізнавання іменованих сутностей тексту Stanford Named Entity Recognizer [8]. 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 Байєса та умовних випадкових полів, – було побудовано систему ідентифікації та аналізу іменованих сутностей тексту. Результати дослідження та експериментів показали високу якість роботи класифікатора, реалізованого на основі моделі 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 реалізований алгоритм продемонстрував високу точність визначення типів іменованих сутностей тексту на рівні найкращих існуючих світових аналогів.

Також була досліджена модель, в якій для мультикласифікації іменованих сутностей текстів використовуються корегуючі вихідні коди (ECOC). Результати експериментів доводять наявність серйозних перспектив застосування даного підходу для вирішення класичних та прикладних задач комп'ютерної лінгвістики.

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.

3. *Linguistic* Data Consortium (2011) Text Corpus Ontonotes 4.0 – 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 // Труды международной научной конференции. 1–6 сентября 2013. – Кенигсберг; 2013. – С. 321–325.
8. *The* Stanford NLP Group (2006–2015) Stanford Named Entity Recognizer. – http://www-nlp.stanford.edu/software//CRF-NER.html

References

1. LAFFERTY J., MCCALLUM A., PEREIRA F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. in The 18th International Conference on Machine Learning. Williamstown, MA, USA. June 28-July 1, 2001. – Williamstown. 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.
3. Linguistic Data Consortium (2011) OntoNotes Release 4.0 [Online] Available from: https://catalog.ldc.upenn.edu/LDC2011T03
4. TURIAN J., RATINOV L., BENGIO Y. Word representations: a simple and general method for semi-supervised learning. in The 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden. July 11–16, 2010. Uppsala. – P. 384–394.
5. NADEAU D., SEKINE S. A survey of named entity recognition and classification. Lingvisticae Investigationes. 30 (1). – P. 3–26.
6. NADEAU D., TURNEY P., MATWIN S. Unsupervised Named Entity Recognition: Generating Gazetteers and Resolving Ambiguity. in Canadian Conference on Artificial Intelligence-2006. Quebec, Canada. June 7–9, 2006. Quebec. – P. 266–277.
7. ANTONOVA A.Y., SOLOVYOV A.N. Method of Conditional Random Fields in tasks of russian texts processing. in The International Conference on Information technologies and systems-2013. Königsberg. September 1-6, 2013. Königsberg. – P. 321–325.
8. The Stanford NLP Group (2006-2015) Stanford Named Entity Recognizer [Online] Available from: http://www-nlp.stanford.edu/software/CRF-NER.html

Про автора:

Марченко Олександр Олександрович,
доцент, доктор фізико-математичних наук,
доцент кафедри Математичної інформатики факультету кібернетики.
Кількість наукових публікацій в українських виданнях – 52.
Кількість наукових публікацій в іноземних виданнях – 10.
Індекс Гірша – 2.
http://orcid.org/ 0000-0002-5408-5279.

Місце роботи автора:

Київський національний університет імені Тараса Шевченка,
01601, Київ, вул. Володимирська, 64/13.
Факультет кібернетики, кафедра Математичної інформатики.
Тел.: (050) 440 7328.
Факс: (044) 259 0129.
E-mail: rozenkrans@yandex.ua