Технічні науки

УДК 004.912

Піпко Анна Сергіївна
студентка
Національний технічний університет України
«Київський Політехнічний Інститут»
Пипко Анна Сергеевна
студентка
Национальный технический университет Украины
«Киевский Политехнический Институт»
Pipko A.
student
National Technical University of Ukraine
«Kyiv Polytechnic Institute»

ДОСЛІДЖЕННЯ МЕТОДІВ АВТОМАТИЧНОЇ ЧАСТИНОМОВНОЇ РОЗМІТКИ ТЕКСТІВ
ИССЛЕДОВАНИЕ МЕТОДОВ АВТОМАТИЧЕСКОЙ ЧАСТЕРЕЧНОЙ РАЗМЕТКИ ТЕКСТОВ
INVESTIGATION OF PART-OF-SPEECH TAGGING METHODS

Анотація: досліджено деякі ймовірнісні методи автоматичної частиномовної розмітки тексту та порівняно якість їх роботи.

Ключові слова: автоматична частиномовна розмітка тексту, алгоритм Вітербі, прихована марковська модель, марковська модель максимальної ентропії.

Аннотация: исследовано некоторые вероятностные методы автоматической частеречной разметки текста и проведено сравнение качества их работы.

Ключевые слова: автоматическая частеречная разметка текста, алгоритм Витерби, скрытая марковская модель, марковская модель максимальной энтропии.

Summary: some of methods of part-of-speech tagging were investigated and compared.

Key words: part-of-speech tagging, Viterbi algorithm, Hidden Markov model, Maximum-entropy Markov model.

Автоматична обробка текстів (обробка природної мови) – загальний напрямок штучного інтелекту та математичної лінгвістики, що вивчає проблеми комп'ютерного аналізу та синтезу природних мов. Ці проблеми дуже актуальні, адже їх розв'язання буде означати створення зручнішої форми взаємодії людини та комп'ютера, а саме у задачах інформаційного пошуку, виділення фактів, машинного перекладу, розпізнавання та синтезу мовлення, створення систем «питання-відповідь» [1].

Частиномовна розмітка тексту (автоматична морфологічна розмітка, POST, POS-tagging, part-of-speech tagging) – один з перших етапів комп'ютерного аналізу тексту, метою якого є визначення частини мови, до якої відноситься слово у тексті (корпусі), з врахуванням контексту слова у словосполученні, реченні та тексту в цілому. Методи частиномовної розмітки поділяються на дві групи: засновані на правилах та ймовірнісні [2].

Метою даної роботи є дослідження, реалізація та порівняння практичних результатів роботи деяких методів ймовірнісного POS-tagging.

В якості першої моделі було взято очевидний POST (part-of-speech tagger, теггер), в якому для визначення відповіді обирається тег, який найчастіше зустрічався з заданим словом у тренувальному корпусі:

$$tag(w) = arg\max_{i\epsilon 1\ldots|Tags|} P(tag_i|w).$$

Суттєвим недоліком даного методу є те, що якщо слово не зустрічалося у тренувальному корпусі, то визначити для нього тег не вдасться. У цьому випадку можна співставляти слову якийсь фіксований тег чи розглядати вкорочене слово, яке, можливо, співпаде з відомими словами.

Друга модель реалізовує безконтекстний POST, який максимізує ймовірність слова, вважаючи, що на це впливає лише його тег:

$$\mathrm{tag}(w) = \arg \max_{i \epsilon 1 \ldots |Tags|} P(w|tag_i)\, P(tag_i).$$

Для визначення тега невідомого слова використовується згладжування $P(word|tag)$ для позбавлення від нульових ймовірностей.

Третя модель використовує алгоритм Вітербі – динамічний алгоритм пошуку найбільш ймовірного ланцюга станів (так званого шляху Вітербі), який у контексті прихованої марковської моделі отримує найбільш імовірну послідовність подій [2]. Тут в якості прихованої змінної виступає тег, а в якості спостережуваної – слово з тексту. Метою алгоритму Вітербі є визначення найбільш імовірної послідовності прихованих змінних $x_1, \ldots, x_T$, що визначається рекурентними співвідношеннями

$$V_{1,k} = P(y_1|k) \cdot \pi_k,$$

$$V_{t,k} = P(y_t|k) \cdot \max_{x \epsilon S}\{a_{x,k} \cdot V_{t-1,x}\},$$

де $V_{t,k}$ – найбільша ймовірність послідовності станів довжини t, що закінчуються в стані k;

π_k – початкові ймовірності знаходження у стані k;

S – простір станів;

$a_{x,k}$ – ймовірність переходу зі стану x в стан k;

y_k – спостережувана змінна.

Оскільки кожний наступний стан залежить тільки від попереднього, то достатньо пам’ятати найбільші ймовірності потрапляння в кожний стан на попередньому кроці. Завдяки цьому ж можна і відновити найбільш

імовірний шлях. Аналогічно можна побудувати модель оберненого POST на основі алгоритму Вітербі, вважаючи, що не попередній тег впливає на наступний, а навпаки.

П'ята розглянута модель є простою композицією безконтекстного POST, прямого та оберненого POST на основі алгоритму Вітербі, яка обирає тег, який видала найбільша кількість моделей.

Моделі були реалізовані мовою Python за допомогою відповідних класів, що містять необхідні умовні та безумовні ймовірності. Кожен теггер отримує на вхід речення та в результаті роботи повертає послідовність пар «слово-тег».

Для тренування та тестування моделей було обрано корпус CoNLL-2000, що створений з розділів корпусу Wall Street Journal [3]. Тренувальна вибірка містить 211727 токенів, тестова – 47377.

Якість роботи моделей оцінювалась на основі accuracy (точність) – відношення кількості правильно встановлених тегів до загальної кількості оброблених слів.

Для порівняння було обрано готові POST з бібліотеки автоматичної обробки текстів мовою Python NLTK (Natural Language Toolkit), що використовують уніграми чи біграми, та їх композиція [4]. З результатів оцінювання (табл. 1) можна зробити висновок, що власноруч реалізовані POST допускають майже в 2 рази менше помилок, ніж стандартні інструменти. Врахування контексту слова (точніше тегів слів контексту) дозволяє значно підвищити якість класифікації. Комбінація кількох методів зменшує кількість помилок, специфічних для кожної з моделей. Використання біграм у очевидному POST дає поганий результат за рахунок того, що дуже велика кількість біграм є унікальною та не зустрічалася у навчальному корпусі.

Таблиця 1

Результати роботи POST

POST	Accuracy
Очевидний POST	90.72%
Безконтекстний POST	91.62%
POST на основі алгоритму Вітербі	94.37%
Обернений POST на основі алгоритму Вітербі	94.40%
Композиція	94.46%
Уніграмний POST (NLTK)	89.57%
Біграмний POST (NLTK)	20.99%
Композиція (NLTK)	90.71%

В подальшому варто дослідити інші способи згладжування для уникнення нульових ймовірностей у випадку незнайомих слів та методи комбінації алгоритмів, які б враховували типи помилок, характерні та специфічні для кожного з них. Також варто зазначити, що алгоритм Вітербі можна адаптувати для використання марковської моделі максимальної ентропії, що буде моделювати залежність прихованої змінної від попереднього її значення та поточного значення спостережуваної:

$$V_{t,k} = \max_{x \epsilon S}\{V_{t-1,x} \cdot P(k|x, y_t)\}.$$

Висновки. В результаті роботи було реалізовано очевидний, безконтекстний POST, прямий та обернений POST на основі алгоритму Вітербі та їх композицію, яка перевершила в якості класифікації стандартні інструменти з бібліотеки NLTK. Безконтекстні методи, а особливо з використанням біграм, поступаються методам на основі ланцюгів Маркова. Можливим шляхом підвищення якості класифікації теггера є дослідження різноманітних способів комбінації методів та згладжування.

Література:

1. *Segeran T. Programming Collective Intelligence / T.Segeran. – O`Reilly Media, Inc., 2007. - 362p.*
2. *Manning, Christopher D. Foundations of Statistical Natural Language Processing / Christopher D., Hinrich Schultze. – The MIT Press, 1999. – 680 p.*
3. *Chunking [Електронний ресурс]. – Режим доступу: http://www.cnts.ua.ac.be/conll2000/chunking.*
4. *Bird S. Natural Language Processing with Python / S.Bird, Klein E., Loper E. – O`Reilly Media, Inc., 2009. - 504p.*