УДК 519.6

Бевза М. В.¹, асп.,

M. V. Bevza¹, PhD stud.,

Розпізнавання іменованих сутностей в текстах за допомогою розмітки частин мови

Named entity recognition in texts with the help of part of speech tagging

¹ Київський національний університет імені
Тараса Шевченка, 01601, м. Київ, вул.
Володимирська 64/13,
e-mail: ¹ maksym.univ@gmail.com,

¹ Taras Shevchenko National University of Kyiv,
01601, Kyiv, Volodymyrska st. 64/13,
e-mail: ¹ maksym.univ@gmail.com,

У цій роботі ми аналізуємо архітектури нейронних мереж, що отримують найкращі результати у розв'язку задачі розпізнавання іменованих сутностей, та пропонуємо нову архітектуру для покращення цих результатів. Ми проаналізували літературу по новим підходам до створення нейронних мереж. В цій роботі ми презентуємо декілька архітектур які ми вважаємо найбільш перспективними для покращення результатів роботи алгоритму розмітки іменованих сутностей і частин мови. Ці архітектури базуються на останніх розробках в сфері нейронних мереж що навчаються одночасно розв'язувати декілька задач одночасно.

У цій роботі ми перевіряємо гіпотези, що задачі розмітки частин мови і іменованих сутностей є пов'язані один з одним, тобто додавання інформації про частини мови може допомогти розв'язати задачу визначення іменованих сутностей і навпаки.

Ця робота також містить реалізацію нейронної мережі і результати експериментів разом з висновками які були отримані внаслідок цих експериментів.

Ключові слова: нейронні мережі, розпізнавання іменованих сутностей, регуляризація, узагальненість, мульти-задачне тренування

We analyze neural network architectures that yield state of the art results on named entity recognition task and propose a number of new architectures for improving results even further. We have analyzed a number of ideas and approaches that researchers have used to achieve state of the art results in a variety of NLP tasks. In this work, we present a few architectures which we consider to be most likely to improve the existing state of the art solutions for named entity recognition task and part of speech tasks. The architectures are inspired by recent developments in multi-task learning.

This work tests the hypothesis that NER and POS are related tasks and adding information about POS tags as input to the network can help achieve better NER results. And vice versa, information about NER tags can help solve the task of POS tagging.

This work also contains the implementation of the network and results of the experiments together with the conclusions and future work.

Keywords: neural networks, named entity recognition, regularization, generalizability, part of speech tagging, multi-task learning.

Статтю представив д.ф-м.н., проф., Анісімов А. В.

Вступ

Стрімкий ріст кількості інформації у всесвітній мережі “Інтернет” призводить до необхідності розробки автоматичних методів роботи з нею. Більшість інформації в публічному доступі є неструктурованою і потребує інтелектуальних систем, які матимуть змогу обробляти її і отримувати корисні структуровані дані для подальшої обробки.

Однією з таких інтелектуальних систем є NER-система (Named Entity Recognition). Вона розв’язує задачу визначення іменованих сутностей. В літературі найчастіше розглядають наступні чотири їх типи: LOC, PER, ORG, MISC - місцевість, персона, організація, інше.

У цій роботі ми аналізуємо існуючі на даний момент роботи, що розв’язують задачу NER, архітектури їх нейронних мереж, а також пропонуємо альтернативну нейронну мережу, яка використовує теги частин мови (POS-tags) як додатковий вхід для покращення узагальненості мережі, а як результат - підвищення точності на тестовому наборі даних.

Сучасні методи розв’язку задачі розпізнавання іменованих сутностей використовують рекурентні (RNN) та згорткові (CNN) нейронні мережі, а також умовні випадкові поля (Conditional Random Fields).

Рекурентна нейронна мережа (RNN) може отримувати на вхід наперед невизначену кількість вхідних даних в послідовності і опрацьовувати їх послідовно. Щоправда, є практичні обмеження щодо кількості вхідних даних у одній послідовності, які зумовлені обмеженням розміром прихованого шару мережі. Рекурентні нейронні мережі мають велику популярність в сучасних методах обробки текстів, за рахунок їх можливості “пам’ятати” контекст речення, при подачі їм на вхід послідовності слів. Найбільш поширені реалізації рекурентних нейронних мереж - це LSTM [2] та GRU.

Згорткова нейронна мережа, вперше описана в роботі LeCun & Bengio [3], отримує на вхід послідовність фіксованого розміру і

працює в вузькому контексті (декілька елементів послідовності на вхід і один вихід). Згорткові нейронні мережі набули більшого поширення для задач обробки зображень, тому що інформація про конкретний піксель важлива не сама по собі, а в контексті сусідніх пікселів. Згодом згорткові нейронні мережі були використані і для задач, які пов’язані з текстом і послідовностями змінної довжини.

Умовні випадкові поля використовуються для розв’язку задач, що вимагають певної структури у вихідних даних. Для розв’язку задачі NER часто використовують BIO нотацію, що вимагає певну структуру у вихідних даних. У розділі “Система міток” буде детальніше описано необхідність використання умовних випадкових полів.

Наша робота має наступну структуру. У розділі “Система міток” ми опишемо систему класів, розпізнавання яких ми поставимо як задачу нашій нейронній мережі. У розділі “Огляд літератури” ми коротко узагальнимо існуючі підходи, що отримують найкращу якість визначення іменованих сутностей. У розділі “План експериментів” ми опишемо дані, що будуть використовувати для експериментів і самі гіпотези, що ми будемо перевіряти. У розділі “Деталі реалізації” ми опишемо структуру нейронної мережі на основі якої будуть проводити експерименти. У розділі “Результати” ми опишемо отримані емпіричні дані з експериментів. У розділі “Висновки” ми визначимо які з гіпотез підтвердилися.

2. Система міток

2.1 BIO-нотація

Задача розпізнавання іменованих сутностей - це задача маркування послідовності, тобто класифікації кожного її елемента на один з N класів. Для того, щоб мати можливість розмітити якусь сутність, яка складається з декількох слів часто використовують BIO нотацію. BIO - це

Beginning (початок), Inside (внутрішність), Outside (зовнішність). Правила маркування наступні:

- кожне слово, яке не є частиною жодної сутності помічається "O"
- кожне слово, яке є початком сутності помічається "B-<TYPE>", де <TYPE> - це тип сутності (LOC, PER, ORG, MISC).
- решта слів помічають "I-<TYPE>"

Для прикладу речення "Barak Obama is a president of United States" буде розмічено наступним чином: "Barak[B-PER] Obama[I-PER] is[O] a[O] president[O] of[O] United[B-LOC] States[I-LOC]". Таким чином наявні дві сутності: персона Barak Obama і місцевість United States.

В нашому випадку задача кожне слово послідовності буде прокласифіковане на одне з наступних класів:

1. O
2. B-PER
3. I-PER
4. B-LOC
5. I-LOC
6. B-ORG
7. I-ORG
8. B-MISC
9. I-MISC

2.2 Умовні випадкові поля

Умовні випадкові поля (CRF) [11] використовуються для розрахунку ймовірності всієї послідовності $P(y_1, y_2, y_3, \dots, y_n)$. Найчастіше їх використовують для структурного передбачення, тобто коли на вихід системи накладаються якісь обмеження. У нашому випадку обмеження є тільки одне. Мітка I-<TYPE> може слідувати тільки за міткою B-<TYPE> того ж типу <TYPE>.

Опис принципу роботи CRF виходить за межі цієї роботи і ми пропонуємо читачу звернутися до роботи Lafferty et al. [11] за детальним описом цього методу.

3. Огляд літератури

Найкращі результати у розв'язанні задачі NER отримують системи побудовані на основі нейронних мереж. У цьому розділі ми проаналізуємо існуючі архітектури і підходи до розв'язку цих задач.

У роботі Ma & Novy [1] автори пропонують використати двосторонню рекурентну мережу, що отримує на вхід вбудовування слів, які вони самі і навчають в процесі тренування. Нагадаємо, що двостороння рекурентна мережа відрізняється від звичайної односторонньої тим, що це - дві рекурентні мережі, одна з яких отримує послідовність на вхід в прямому порядку, а інша - в зворотному, а отримані виходи з цих мереж об'єднуються конкатенацією їх векторів, створюючи вектор вдвічі більшої розмірності.

Вбудовування слів отримують за допомогою матриці emb розміром $D \times E$, де D - розмір словника слів, а E - розмір вбудовувань. $e = emb(w_i)$, $i = 1..n$, де n - довжина вхідної послідовності слів речення $W = (w_1 w_2 .. w_n)$.

Двостороння рекурентна мережа $rnn = [rnn_{fw}, rnn_{bw}]$ отримує на вхід послідовність і результатом її роботи є послідовність довжини n векторів прихованих станів.

$$\begin{aligned} h_{fw} &= rnn_{fw}(x_1, x_2, \dots, x_n), \quad h_{fw} = (h_{fw,1}, h_{fw,n}) \\ h_{bw} &= rnn_{bw}(x_n, x_{n-1}, \dots, x_1), \quad h_{bw} = (h_{bw,n}, h_{bw,1}) \\ h &= [h_{fw}, h_{bw}] \end{aligned}$$

Робота Ma & Novy [1] використовує LSTM як базову рекурентну одиницю. Їхня мережа навчається вбудовувань слів emb самостійно, а також додатково вбудовування отримуються на основі символічних репрезентацій. Для отримання символічної репрезентації згорткова мережа опрацьовує кожне слово на рівні символів і будує багатовимірний вектор який об'єднується зі стандартним вбудовуванням на рівні слів. Для

Рис. 1. Приклад маркування BIO-нотацією

Рис. 2. Застосування згорткової мережі для побудови символних репрезентацій слова. Джерело: Ma & Novy [1]

отримання коректної послідовності результуючих міток, в роботі використано CRF [8] (див. секцію “Система міток”).

У роботі Lample et al. [5] архітектура мережі ідентична до тієї, що запропонована Ma & Novy [1] за виключенням способу отримання вбудовувань слів. Вони пропонують робити це наступним способом: обробити послідовність символів за допомогою двосторонньої LSTM мережі.

Ця мережа отримує на вхід послідовність вбудовувань символів і вихід з останнього прихованого шару слугує вбудовуванням на основі символів. Ця мережа також використовує CRF для забезпечення коректної послідовності результуючих міток.

4. План експериментів

4.1. Дані

Як джерело даних для задачі розмітки іменованих сутностей і тегів часнин мови ми використаємо широко поширений в літературі набір даних CoNLL 2003. Корпус CoNLL 2003 NER був створена на основі новинних текстів з корпусу Reuters RCV1, розмічений чотирма різними типами іменованих сутностей (PER, LOC, ORG, MISC). Кожне слово в наборі також має мітку частини мови до якої воно належить.

4.2. Гіпотези

В межах цієї роботи ми перевіряємо наступні гіпотези.

Гіпотеза 1. Додавання POS як входу додатково до слів дає можливість покращити результати для задачі NER.

Гіпотеза 2. Додавання NER як входу додатково до слів дає можливість покращити результати для задачі POS.

Гіпотеза 3. Додавання POS як виходу і включення його в функцію втрат дає можливість покращити результати для задачі NER.

Гіпотеза 4. Додавання NER як виходу і включення його в функцію втрат дає можливість покращити результати для задачі POS.

5. Деталі реалізації

5.1. Архітектура нейронної мережі

У цій роботі ми пропонуємо архітектуру нейронної мережі, що складається з шару вбудовувань, двох шарів рекурентної нейронної мережі (LSTM), а також двох повнозв'язних шарів. Перший повнозв'язний шар буде видавати розподіл ймовірностей по словнику NER міток (O, B-LOC, I-LOC, B-PER, I-PER, B-ORG, I-ORG, B-MISC, I-MISC), а другий повнозв'язний шар буде видавати розподіл ймовірностей по словнику POS-міток.

У архітектурі, що ми пропонуємо,