УДК 519.6

Бевза М. В.¹, асп.,

M. V. Bevza¹, PhD stud.,

**Об'єднання вбудовувань для покращення
регуляризації нейронних мереж для задачі
розпізнавання іменованих сутностей**

**Tying of embeddings for improving
regularization in neural networks for named
entity recognition task**

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 і, базуючись на цьому, запропонували підхід для покращення архітектури існуючих нейронних мереж для задачі розпізнавання іменованих сутностей. Ми приводимо обґрунтування чому запропоновані методи покращують узагальненість мережі і як результат - якість її роботи.

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 new architecture 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 of them which we consider to be most likely to improve existing state of the art solutions for named entity recognition task. The architecture is inspired by recent developments in language modeling task. The suggested solution is based on a multi-task learning approach. We incorporate part of speech tags as input for the network. Part of speech tags to be yielded by some state of the art tagger and also ask the network to produce those tags in addition to the main named entity recognition tags. This way knowledge distillation from a strong part of speech tagger to our smaller network is happening. We hypothesize that designing neural network architecture in this way improves the generalizability of the system and provide arguments to support this statement.

Keywords: neural networks, named entity recognition, regularization, generalizability, 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 нашої нейронної мережі. У розділі “Огляд літератури” ми коротко узагальнимо існуючі підходи, що отримують найкращу якість визначення іменованих сутностей. У розділі “Архітектура нейронної мережі” ми опишемо нашу пропозицію щодо використання тегів частини мови для задачі NER, а також обґрунтуємо чому ми вважаємо, що цей підхід дасть покращення якості роботи мережі.

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) [8] використовуються для розрахунку ймовірності всієї послідовності $P(y_1, \dots, y_n)$. 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. [8] 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 $W = (w_1 w_2 \dots w_n)$.

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

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

Робота Ma & Novy [1] використовує LSTM як базову рекурентну одиницю. Їхня мережа навчається вбудовувань слів emb самостійно, а також додатково вбудовування

Рис. 1. Приклад маркування BIO-нотацією

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

отримуються на основі символьних репрезентацій. Для отримання символьної репрезентації згорткова мережа опрацьовує кожне слово на рівні символів і буде багатовимірний вектор який об'єднується зі стандартним вбудовуванням на рівні слів. Для 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 послідов-

ність 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. 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, кожне слово буде супроводжено також тегом частини мови. Ми побудуємо два словники: словник слів V_w та словник тегів частин мови V_{pos} . Матриця вбудовувань для V_w позначимо E_w , для V_{pos} - E_{pos} .

$$e_i = E_w(w_i)$$

$$e_{pos,i} = E_{pos}(pos_i)$$

Далі ми використаємо дві двосторонні рекурентні мережі. Одна буде отримувати на вхід послідовність вбудовувань слів, а друга - вбудовувань частин мови.

$$h_w = rnn_w(e_1, e_2, \dots, e_n)$$

$$h_{pos} = rnn_{pos}(e_{pos,1}, e_{pos,2}, \dots, e_{pos,n})$$

Отримані приховані стани використаємо для передбачення міток NER та тегів частин мови. Для цього використаємо повнозв'язні шари нейронної мережі, що перетворюють приховані стан в розподіл ймовірностей по словнику міток. out_{ner} та out_{pos} для міток NER та POS відповідно.

$$p_{ner} = softmax(out_{ner}(h_w))$$

$$p_w = softmax(out_{pos}([h_{pos}, h_w]))$$

$$softmax(x) = \frac{\exp(x_i)}{\sum_{i=1}^k \exp(x_i)}$$

Звернемо увагу, що для отримання розподілу ймовірностей рпег інформація про справжні мітки, що були подані на вхід, не використовується, в іншому випадку мережа навчалася б просто запам'ятовувати теги частин мови, а не передбачати їх, що призвело б до поганого рівня узагальнення навчання мережі, а отже - поганої якості роботи.

Рис. 3. Схематичне зображення запропонованої архітектури нейронної мережі

Звернемо увагу, що матриця E_{pos} та матриця out_{pos} мають однакові розмірності з точністю до операції транспонування:

$$E_{pos} = out_{pos}^T.$$

Ми пропонуємо ототожнити ці матриці і навчати їх одночасно. Ми вважаємо, що аналогічно до роботи Melis et. al. [6] ми отримаємо кращі результати, за рахунок того, що мережа буде мати меншу кількість степеней свободи і це буде запобігати перенавчанню на тренувальній вибірці.

Список використаних джерел

1. Ма К., Хові І., Розмічення послідовностей за допомогою двосторонньої LSTM-CNN-CRF мережі. // *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016. – 1064-1074 с.
2. Хошрейтер С., Шмідхубер Ю., Довга коросторокова пам'ять // *Neural Computation*. 9 – 1997. – 1735-1780 с.
3. Лекун Й., Бенжіо Й., Згорткові мережі для обробки зображень та звуку. // *The handbook of brain theory and neural networks*. – 1995. – 255-258 с.
4. Валлаш Х. Умовні випадкові поля: вступ. // *Foundations and Trends in Machine Learning archive*. Volume 4 Issue 4. – 2004. 267-273 с.
5. Лампл Г., Баллестерос М., Субраманіан С., Кавакамі К., Даєр К. Нейронні архітектури для розпізнавання іменованих сутностей. // *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. – 2016. – 260-270 с.
6. Меліс Г., Даєр К., Блунсом П. Про найкращі результати в оцінках нейронних мереж. // *International Conference on Learning Representations*. – 2007 – 214-216 с.
8. Лафферті Й., Мак-Каллум А., Перейра Ф. Умовні випадкові поля: моделі для сегментування і розмітки послідовностей. // *Proc. 18th International Conf. on Machine Learning*. Morgan Kaufmann. – 2001 – 282-289 с.

References

1. MA, X. & HOVY, E. (2016) End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 1064-1074.
2. HOCHREITER, S. & SCHMIDHUBER, J. (1997). Long short-term memory. *Neural Computation*. 9: pp. 1735–1780.
3. LECUNN, Y., BENGIO, Y. (1995). Convolutional Networks for Image, Speech, and Time-Series. In *The handbook of brain theory and neural networks*. pp. 255-258.
4. WALLACH, H. (2004). Conditional Random Fields: An Introduction. In *Foundations and Trends in Machine Learning archive*. Volume 4 Issue 4, pp. 267-273
5. LAMPLE, G. & BALLESTEROS, M. & SUBRAMANIAN, S. & KAWAKAMI, K. & DYER, C. (2016). Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 260–270.
6. MELIS, G. & DYER, C. & BLUNSOM, P. (2017). On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations*. pp. 214-216.
8. LAFFERTY, J., MCCALLUM, A., PEREIRA, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. 18th International Conf. on Machine Learning*. Morgan Kaufmann. pp. 282–289.

Надійшла до редколегії 29.01.2019