Институт информационных и вычислительных технологий МОН РК

Казахский Национальный Университет имени аль-Фараби

Университет Туран

Люблинский технический университет, Польша

«Ғылым ордасы»

МАТЕРИАЛЫ

IV международной научно-практической конференции "Информатика и прикладная математика", посвященной 70-летнему юбилею профессоров Биярова Т.Н., Вальдемара Вуйцика и 60-летию профессора Амиргалиева Е.Н. 25-29 сентябрь 2019, Алматы, Казахстан

Часть 1

Алматы 2019

Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", pp. 702-720. 2016.

2. Duc-Thuan Vo, Bagheri, E. Open information extraction. Encyclopedia with Semantic Computing and Robotic intelligence: 2016, Vol. 1, No. 1 (pp. 1630003). World Scientific Publishing Company.

3. Fader, A., Soderland, S., Etzioni, O. Identifying relations for open information extraction. Proceedings of the conference on empirical methods in natural language processing. Edinburgh, Scotland, UK , 2011, pp. 1535-1545

4. Etzioni, O., Banko, M., Soderland, S., Weld, D. Open information extraction from the web. Communications of the ACM, 2008. Vol. 51 No. 12 (pp. 68-74). New York, NY, USA.

5. Gamallo, P., Garcia, M., Fernandez-Lanza, S. Dependency-based open information extraction. Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP Avignon, France. 2012, (pp. 10–18).

6. Akbik, A., Loser, A. KrakeN: N-ary facts in open information extraction. Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction Montreal, Canada. 2012, (pp. 52–56).

7. Gashteovski, K., Gemulla, R., Del Corro, L. MinIE: Minimizing Facts in Open Information Extraction. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) Copenhagen, Denmark, 2017, (pp. 2630-2640).

8. Angeli, G., Premkumar, M. J., D Manning.C. D. Leveraging linguistic structure for open domain information extraction. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics Beijing, China. 2015, (pp. 344-354).

9. Gamallo, P., Garcia, M. Multilingual Open Information Extraction. In Portuguese Conference on Artificial Intelligence. Coimbra, Portugal. 2015, (pp. 711-722).

10. МФ Бондаренко, ЮП Шабанов-Кушнаренко. Мозгоподобные структуры: Справочное пособие. Том первый. Под редакцией акад. НАН Украины ИВ Сергиенко. К.: Наукова думкаа, 2011. – 460 с.

11. Khairova, N.F.,Petrasova, S.,Gautam, A.P. The logical-linguistic model of fact extraction from English texts. Information and Software Technologies. Volume 639 of the series Communications in Computer and Information Science, Springer, ISBN: 978-3-319-46253-0, 2016, pp. 625-635. doi> 10.1007/978-3-319-46254-7_51

12. Khairova, N., Lewoniewski, W., Wecel, K. Estimating the Quality of Articles in Russian Wikipedia Using the Logical-Linguistic Model of Fact Extraction. Conference proceedings. BIS 2017. Part of the Lecture Notes in Business Information Processing book series (LNBIP, volume 288). (pp. 28-40). Poland: Poznan.

ИЗВЛЕЧЕНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ ИЗ НОВОСТНЫХ ИСТОЧНИКОВ НА ОСНОВЕ BI-LSTM

Чикибаева Д.Ю., Мансурова М.Е., Нугуманова А.Б., Кыргызбаева М.Е.

e-mail: dashachikibaeva@gmail.com, mansurova.madina@gmail.com, yalisha@yandex.kz, marzhan.kyrgyzbaeva@gmail.com
Казахский Национальный университет им. аль-Фараби, Алматы, Казахстан

Аннотация. *Извлечение именованных сущностей является одной из важных задач в области обработки естественного языка. Оно может найти своё применение в таких сферах, как распознавание речи, информационный поиск, др. В последнее время хорошие результаты достигаются системами, основанными на методах глубокого обучения. В настоящей работе рассмотрен метод извлечения именованных сущностей на основе рекуррентных нейронных сетей.*

Ключевые слова. *Извлечение именованных сущностей, обработка естественного языка, машинное обучение, глубокое обучение, bi-LSTM.*

Введение

В настоящее время существуют различные подходы для извлечения информации. Они разнообразны и сложно сказать, что один лучше другого, так как тот или иной показывает хорошие результаты в разных ситуациях. Подходы для извлечения информации могут быть классифицированы на следующие категории:

–подходы, основанные на правилах. Эксперты составляют вручную наборы правил, необходимые для извлечения определённых данных.

–подходы, основанные на знаниях. Сюда относятся модели, основанные на онтологиях [1], модели, основанные на тезаурусах [2].

–статистические подходы. Включают скрытые марковские модели [3-5], условные марковские модели [6], условные случайные поля [7].

–подходы, основанные на машинном обучении.

Одной из подзадач извлечения информации является распознавание именованных сущностей. Из текста выделяются именованные объекты, такие как имена людей, названия организаций, локации, геополитические объекты, временные метки, расширенный вариант может включать специфичные для определённой предметной области термины, такие как медицинские, биологические, др. Существующие методы распознавания именованных сущностей можно разделить на категории:

–Основанные на правилах. Являются одними из ранних систем для распознавания именованных сущностей. Правила основываются на лексико-синтаксических шаблонах специфичных для определённого языка.

–Обучение с учителем. Данный метод требует большого количества тренировочных данных, размеченных вручную экспертами. Затем система на основе предоставленных данных извлекает правила для распознавания именованных сущностей.

–Обучение без учителя. Система использует небольшой начальный набор экземпляров именованных сущностей, который используется для дальнейшего обучения системы.

–Гибридные системы. Включают в себя две или более техники обучения на основе машинного обучения или на основе правил. [8, 9].

Модель

В данной работе используется подход обучения с учителем. Модель основана на bi-LSTM блоке с применением векторизации символов и слов.

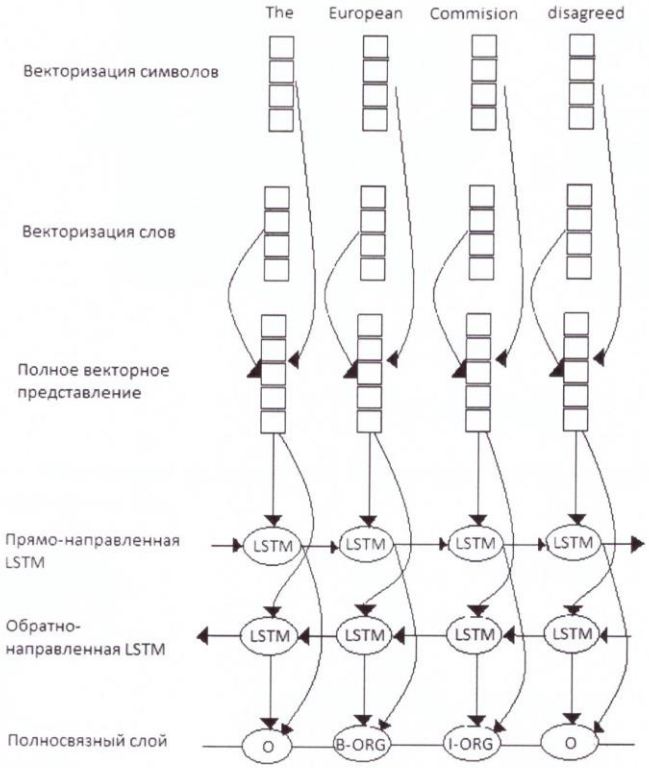

Рис. 1 Основная архитектура сети

LSTM. LSTM (Long Short-Term Memory) – разновидность рекуррентных нейронных сетей. Рекуррентные нейронные сети обладают способностью запоминания результатов прошлых итераций, однако они не способны запоминать их долгосрочно. Появляется проблема исчезновения градиента [10]. LSTM сети [11] разработаны с целью бороться с данной проблемой. Они содержат три основных блока, которые контролируют какая информация будет забыта, а какая будет передана на последующие итерации. Схематично LSTM блок может быть изображен как на рис. 2.

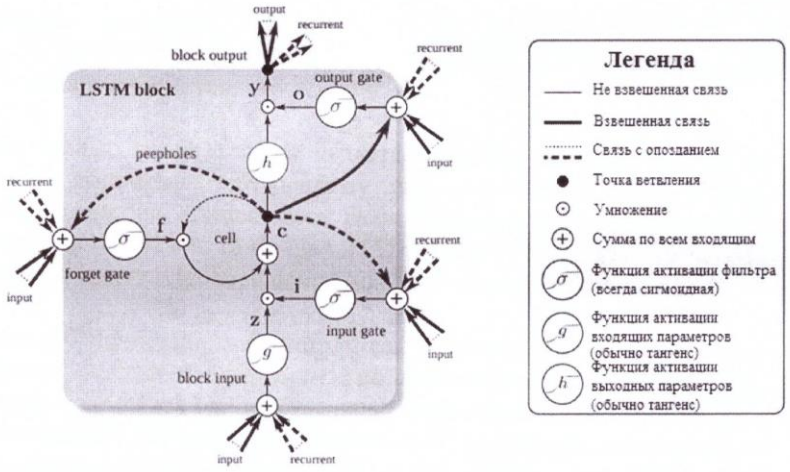

Рис. 2 LSTM блок [12]

Формально формулы для обновления LSTM блока в момент времени t можно представить, как:

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \quad (1)$$

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \quad (2)$$

$$\tilde{c}_t = tanh(W_c h_{t-1} + U_c x_t + b_c) \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (4)$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o) \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

где

x_t – входной вектор в момент времени t,
h_t – вектор скрытого состояния (вывода) в момент времени t,
σ- сигмоидная функция,
Ui, Uf, Uc, Uo – матрицы весов различных фильтров для входного вектора x_t,
Wi, Wf, Wc, Wo – матрицы весов для вектора скрытого состояния h_t,
$\odot$ - поэлементное умножение,
b_i, b_f, b_c, b_o – векторы смещения.

Bi-LSTM. Главной идеей bi-LSTM является учитывание последовательности символов не только до текущей, но и после [13]. Таким образом происходит рассмотрение полного окружения. Bi-LSTM состоит из прямо направленной LSTM,

которая рассматривает последовательность до и обратно направленной, рассматривающей последовательность после. Затем полученные последовательности конкатенируются.

Векторизация. Целью векторизации является извлечение информации из текстового корпуса и сопоставление каждому его элементу (слову/символу) уникального числового вектора. Существуют разнообразные методы реализации векторизации, среди них Word2Vec [14], LSA [15], GloVe [16]. В данной работе использовался один из простых способов векторизации – one-hot embedding.

Набор данных. Обучение построенной модели проводилось на данных на русском языке, в качестве набора данных были использованы новостные источники, собранные с различных Казахстанских официальных новостных интернет-порталов [17-23]. Объём набора данных 330 составил предложений. Количество именованных сущностей по категориям в наборе данных: LOC – 295, ORG – 380, PER – 214. Данные были размечены с применением семи тегов: B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG, O. Где PER – PERSON (персона, имена людей), LOC – LOCATION (локация), ORG – ORGANIZATION (организация), B – BEGINNING (начало, первый токен именованной сущности), I – INSIDE (последующие (внутренние) токены именованной сущности), O - OTHER, токен, не являющийся именованной сущностью.

Обучение. Тренировка модели происходила в течение 10 эпох.

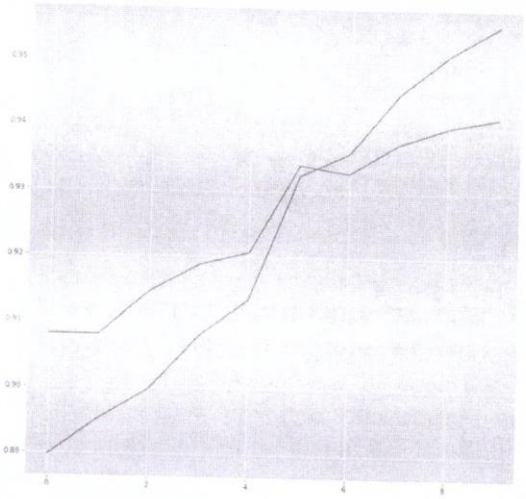

Рис. 3 Изменение точности (accuracy) в зависимости от эпохи, красной линией показана точность при обучении, синей линией – точность при валидации

Представитель	O	O
филиала	O	O
НДП	O	O
«	O	O
Нұр	B-ORG	B-ORG
Отан	I-ORG	I-ORG
»	O	O
ознакомили	O	O
участников	O	O
семинара	O	O
с	O	O
основными	O	O
требованиями	O	O
антикоррупционного	O	O
законодательства	O	O
Республики	B-LOC	B-LOC
Казахстан	I-LOC	I-LOC
	O	O

Рис. 4 Пример распознавания именованных сущностей в предложении, где во второй колонке показаны исходные теги, а в третьей – теги, предсказанные моделью

Анализ построенной модели осуществлялся с помощью F_1-меры (9). Пусть tp – количество правильно распознанных тегов, fp – количество ошибочно распознанных как именованные сущности тегов, fn – количество ошибочно пропущенных, не распознанных как именованные сущности, тегов. Тогда

$$Precision = \frac{tp}{tp+fp}, \tag{7}$$

$$Recall = \frac{tp}{tp+fp}, \tag{8}$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}, \tag{9}$$

Значение F_1-меры было вычислено со средним «макро» и «микро».

```
f1-score with average macro: 0.6862226877552394
f1-score with average micro: 0.919754180037684
```

Рис. 5 Значение F1-меры

Заключение

В данной работе был представлен метод извлечения именованных сущностей на русском языке с помощью алгоритма Bi-LSTM. Несмотря на то, что задача извлечения именованных сущностей имеет множество подходов к решению, наиболее популярными остаются подходы на основе машинного обучения с использованием контекстной информации. Алгоритм Bi-LSTM в качестве

контекстной информации использует двухстороннее окружение целевого слова (до и после него), и как показали проведенные эксперименты демонстрирует высокие результаты точности и полноты даже несмотря на то, что использовалась очень маленькая обучающая коллекция.

Исследование выполнено при финансовой поддержке МОН РК в рамках научного проекта № AP05132933 «Разработка системы извлечения знаний из гетерогенных источников данных для повышения качества принятия решений» (2018-2020).

Литература

1. D. W. Embley, D. M. Campbell, R. D. Smith, and S. W. Liddl. Ontology-based extraction and structuring of information from data-rich unstructured documents. In Conference on Information and Knowledge Management (CIKM), pages 52–59, 1998.
2. C. Cardie. A case-based approach to knowledge acquisition for domain-specific sentence analysis. In Proceedings of the Eleventh National Conference on Artificial Intelligence, pages 798–803. AAAI Press, 1993.
3. T. Scheffer, C. Decomain, and S. Wrobel. Active hidden Markov models for information extraction. In Proceedings of the International Symposium on Intelligent Data Analysis, 2001.
4. T. Scheffer, S. Wrobel, B. Popov, D. Ognianov, C. Decomain, and S. Hoche. Learning hidden Markov models for information extraction actively from partially labeled text. K¨unstliche Intelligenz, (2), 2002.
5. M. Skounakis, M. Craven, and S. Ray. Hierarchical hidden Markov models for information extraction. In IJCAI, 2003.
6. A. K. McCallum, D. Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. In ICML, 2000.
7. A. K. McCallum and D. Jensen. A note on the unification of information extraction and data mining using conditional-probability, relational models. In IJCAI'03 Workshop on Learning Statistical Models from Relational Data, 2003.
8. Saha, S. K., Sarkar, S., & Mitra, P. (2008, January). A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition. In Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I (pp. 343-349). Hyderabad, India.
9. Rocktäschel, T., Weidlich, M., & Leser, U. (2012). ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics, 28(12), 1633-1640.
10. Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty oftraining recurrent neural networks. 2012.
11. Sepp Hochreiter and J¨urgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780. 1997.
12. K. Greff, R. K. Srivastava, J. Koutník, Bas R. Steunebrink and J. Schmidhuber. LSTM: A Search Space Odyssey. 2017.
13. C. Dyer, M. Ballesteros, W. Ling, A. Matthews, and N. A. Smith.Transition-based dependency parsing withstack long short-term memory. InProceedingsof ACL-2015 (Volume 1: Long Papers), pages 334–343, Beijing, China, July. 2015.