Dzendzik D., Serebryakov S. Semi-Automatic Generation of Linear Event Extraction Patterns for Free Texts // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. – 2013. – Т. 155, кн. 4. – С. 99–108.

UDK 004.912

SEMI-AUTOMATIC GENERATION OF LINEAR EVENT EXTRACTION PATTERNS FOR FREE TEXTS

D. Dzendzik, S. Serebryakov

Abstract

In this paper we describe semi-automatic approach to generating event extraction patterns for free texts. The algorithm is composed of four steps: we automatically extract possible events from a corpus of free documents, cluster them using dependency-based parse tree paths, validate random samples from each cluster and generate linear patterns using positive event clusters. We compare it with the system that uses handcrafted patterns.

Keywords: event extraction, linear patterns, regular expressions, TextMARKER, RUTA.

Резюме

Д.А. Дзендзик, С.В. Серебряков. Автоматизированное построение линейных правил для извлечения событий из неаннотированного текста.

В статье описывается автоматизированный подход к построению линейных правил для извлечения событий из неаннотированных текстов. Алгоритм состоит из четырех шагов: автоматическое извлечение потенциальных событий из корпуса неаннотированных документов, кластеризация их с использованием путей в дереве зависимостей, проверка случайно выбранных примеров из каждого кластера и построение линейных правил на основе кластеров, получивших положительную оценку. Проводится сравнение полученных правил с системой, использующей правила, построенные экспертом вручную.

Ключевые слова: извлечение событий, линейные правила, регулярные выражения, TextMARKER, RUTA.

References

1. *Soderland S.* Learning Information Extraction Rules for Semi-Structured and Free Text // Machine Learning. – 1999. – V. 34, No 1–3. – P. 233–272.
2. *Li Y., Krishnamurthy R., Raghavan S., Vaithyanathan S., Jagadish H.V.* Regular expression learning for information extraction // EMNLP'08 Proc. Conf. on Empirical Methods in Natural Language Processing. – Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. – P. 21–30.
3. *Agichtein E., Gravano L.* Snowball: extracting relations from large plain-text collections // DL'00 Proc. Fifth ACM Conf. Digital libraries. – N. Y., USA: ACM, 2000. – P. 85–94.
4. *Bach N., Badaskar S.* A Review of Relation Extraction. – 2007. – URL: http://www.cs.cmu.edu/ñbach/papers/A-survey-on-Relation-Extraction.pdf.

5. *McDonald R.* Extracting Relations from Unstructured Text. Technical Report: MS-CIS-05-06. – 2005. – URL: http://www.ryanmcd.com/papers/MS-CIS-05-06.pdf.

6. *Yangarber R., Grishman R., Tapanainen P.* Automatic Acquisition of Domain Knowledge for Information Extraction // COLING'00 Proc. 18th Conf. on Computational linguistics. – Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. – V. 2. – P. 940–946.

7. *Brin S.* Extracting Patterns and Relations from the World Wide Web // WebDB'98 Selected papers from the Int. Workshop on The World Wide Web and Databases. – London, UK: Springer-Verlag, 1999. – P. 172–183.

8. *Etzioni O., Banko M., Soderland S., Weld D.S.* Open information extraction from the web // Communications of the ACM. – 2008. – V. 51, No 12. – P. 68–74.

9. *Etzioni O., Cafarella M., Downey D., Kok S., Popescu A.-M., Shaked T., Soderland S., Weld D.S., Yates A.* Web-scale information extraction in knowitall: (preliminary results) // WWW'04 Proc. 13th Int. Conf. on World Wide Web. – N. Y., USA: ACM, 2004. – P. 100–110.

10. *Yates A., Banko M., Broadhead M., Cafarella M.J., Etzioni O., Soderland S.* TextRunner: Open Information Extraction on the Web // NAACL-Demonstrations'07 Proc. Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. – Stroudsburg, PA, USA: Association for Computational Linguistics, 2007. – P. 25–26.

11. *Kluegl P., Atzmueller M., Puppe F.* Integrating the Rule-Based IE Component TextMarker into UIMA // Proc. LWA. – 2008. – P. 73–77.

Поступила в редакцию
31.07.13

Dzendzik, Darya – Student, Saint-Petersburg State University; Intern, Hewlett-Packard Laboratories, Saint Petersburg, Russia.

Дзендзик Дарья Анатольевна – студент, Санкт-Петербургский государственный университет; практикант, Российское отделение исследовательской лаборатории "Hewlett-Packard Laboratories", г. Санкт-Петербург, Россия.

E-mail: *daryadzen@gmail.com*, *daria.dzendzik@hp.com*

Serebryakov, Sergey – PhD, Research Engineer, Hewlett-Packard Laboratories, Saint Petersburg, Russia.

Серебряков Сергей Валерьевич – кандидат технических наук, инженер-исследователь, Российское отделение исследовательской лаборатории "Hewlett-Packard Laboratories", г. Санкт-Петербург, Россия.

E-mail: *sergey.serebryakov@hp.com*