Information Technology for Data Selection in Natural Language

Serhiy Dyuh, Oleh Veres

Lviv Polytechnic National University, Lviv, Ukraine

The volume of data collected every day grows exponentially from the beginning of the new Millennium. Most of these data are stored in relational databases. In the past access to this data was interesting mainly large companies that have the ability to query data using structured query languages (SQL). With the increasing number of mobile phones grows and the number of personal data that is stored in various repositories. Thus, more and more people from different walks of life trying to seek and use their own data. However, despite the rapid growth in popularity of data science, most people do not have sufficient knowledge to write SQL queries to work with data. Even more, regular users do not have time to learn and use SQL. Even for those who are constantly working with SQL, writing repetitive queries again and again is a tedious task. Despite this, a huge amount of data available today, can not be effectively used. One solution to this problem is the creation of a system for processing natural speech and convert it to SQL queries.

There are two basic approaches to creating programs for natural language processing:

• Expert method, – compilation rules for processing text and highlight key features is carried out by man.

• Machine learning is the construction of rules is performed automatically based on pre-prepared data array.

To the task of natural language processing has been applied by many different classes of machine learning algorithms. These algorithms take in input a large set of "properties" that are generated from the source data. Some of the early algorithms such as decision trees, created a system of rigid rules "if-then", then such common systems of handwritten rules. Increasingly, however, research focus on statistical models that take fuzzy, probabilistic decisions based on assigning specific weights to each input property. Among the advantages of such models is that they can Express the relative certainty of many different possible answers rather than only one, giving more reliable results when such a model is included as an integral part of a larger system.

Systems based on machine learning algorithms have many advantages over rules, made by hand:

• The procedure used during machine learning automatically focus on the most common cases, whereas when writing rules manually is often it is not obvious where to focus efforts.

COLINS'2020, Volume II: Workshop. Lviv, Ukraine, April 23-24, 2020, ISSN 2523-4013
http://colins.in.ua, online

• Automatic learning procedures can use statistical inference algorithms to create patterns that are resistant to unfamiliar data (such as containing words or structures that have not previously been encountered) and misspellings (such as misspelled words or words accidentally omitted). Typically, handling such input using handwriting rules, or, in general, creating such handwriting rules systems that make fuzzy decisions is an extremely complex process and time-consuming.

• Systems based on automatic rule learning can be made more accurate by simply providing more input. While systems based on handwritten rules can only be made more accurate by increasing the complexity of the rules, this is a much more difficult task. In particular, there is a limit to the complexity of manual-based systems beyond which systems become increasingly unmanageable. However, creating more data to enter into machine learning systems simply requires a corresponding increase in the amount of hours spent by the person, usually without significantly increasing the complexity of the process.

Integration of this system into existing applications will allow non-technical users to work with databases without contacting analysts. As a result, it will not only increase the convenience of the system, but also reduce the latency of waiting for a response and increase the relevance of the data received.

References

1. Natural language and natural selection. - Access mode: https://www.semanticscholar.org/paper/Natural-language-and-natural-selection-Pinker-Bloom/dbe79abbbc1fc7df6ce90d263a363d99fabd3489
2. Kovaliuk, T., Kobets, N.: Semantic Analysis and Natural Language Text Search for Internet Portal. In: Computational linguistics and intelligent systems, COLINS, 277-287. (2019)
3. Zholtkevych, G., Polyakovska, N.: Machine Learning Technique for Regular Pattern Detector Synthesis: toward Mathematical Rationale. In: Computational linguistics and intelligent systems, COLINS, 254-265. (2019)
4. Shapo, V., Volovshchykov, V.: Cloud Technologies Application at English Language Studying for Maritime Branch Specialists. In: Computational linguistics and intelligent systems, COLINS, 243-253. (2019)
5. Kupriianov, Y., Akopiants, N.: Developing linguistic research tools for virtual lexicographic laboratory of the spanish language explanatory dictionary. In: Computational linguistics and intelligent systems, COLINS, 43-52. (2019)
6. Kovaliuk, T., Tielysheva, T., Kobets, N.: Method of Cross-Language Aspect-Oriented Analysis of Statements Using Categorization Model of Machine Learning. In: Computational linguistics and intelligent systems, COLINS, 32-42. (2019)
7. Burov, Y., Vysotska, V., Kravets, P.: Ontological Approach to Plot Analysis and Modeling. In: Computational linguistics and intelligent systems, COLINS, 22-31. (2019)
8. Shepelev, G., Khairova, N., Kochueva, Z.: Method" Mean-Risk" for Comparing Poly-Interval Objects in Intelligent Systems. In: Computational linguistics and intelligent systems, COLINS, 12-21. (2019)

9. Bisikalo, O., Ivanov, Y., Sholota, V. Modeling the Phenomenological Concepts for Figurative Processing of Natural-Language Constructions. Method" Mean-Risk" for Comparing Poly-Interval Objects in Intelligent Systems. In: Computational linguistics and intelligent systems, COLINS, 1-11. (2019)
10. Skopyk, K.: Language modelling and its use cases. In: Computational Linguistics and Intelligent Systems, COLINS, http://colins.in.ua/wp-content/uploads/2018/07/Language-modelling-and-its-use-cases.pdf, 1-11. (2018)
11. Grabar, N., Hamon, T.: Automatic Detection of Temporal Information in Ukrainian General-language Texts. In: Computational Linguistics and Intelligent Systems, COLINS, CEUR workshop proceedings, Vol-2136, 1-10. (2018)
12. Lytvyn, V., Sharonova, N., Hamon, T., Vysotska, V., Grabar, N., Kowalska-Styczen, A.: Computational linguistics and intelligent systems. In: CEUR workshop proceedings, Vol-2136. (2018).
13. Lytvyn, V., Sharonova, N., Hamon, T., Vysotska, V., Grabar, N., Kowalska-Styczen, A.: COLINS 2017. Kharkiv, Ukraine. (2017)
14. Kuprianov, Ye.: Semantic state superpositions and their treatment in virtual lexicographic laboratory for spanish language dictionary. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, P. 37–46. (2017)
15. Kotov, M.: NLP resources for a rare language morphological analyzer: danish case. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 31–36. (2017)
16. Lytvyn, V., Vysotska, V., Chyrun, L., Smolarz, A., Naum, O.: Intelligent system structure for Web resources processing and analysis. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 56–74. (2017)
17. Lande, D., Andrushchenko, V., Balagura, I.: An index of authors' popularity for Internet encyclopedia. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 47–55. (2017)
18. Hamon, T., Grabar, N.: Unsupervised acquisition of morphological resources for Ukrainian. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 20–30. (2017)
19. Grabar N., Hamon, T.: Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. In: 1st International Conference Computational Linguistics and Intelligent Systems, COLINS, 10–19. (2017)
20. Shestakevych T. Modelling of semantics of natural language sentences using generative grammars / Tetiana Shestakevych, Victoria Vysotska, Lyubomyr Chyrun, Liliya Chyrun // Computer Science and Information Technologies: Proc. of the IX-th Int. Conf. CSIT'2014, 18-22 November, 2014, Lviv, Ukraine..– Lviv: Publishing Lviv Polytechnic, 2014.– P.19-22.
21. Vysotska V. Generative regular grammars application to modeling the semantics of sentences in natural language / Victoria Vysotska // Комп'ютерні системи проектування. Теорія і практика, Вісник Національного університету "Львівська політехніка" № 808,- Львів 2014 – Стор.43-56.
22. Lytvyn V. Development of a method for determining the keywords in the slavic language texts based on the technology of web mining / V. Lytvyn, V. Vysotska, P. Pukach, O. Brodyak, D. Ugryn // Eastern-European Journal of Enterprise Technologies. – №2/2(86). – Харків, 2017. – P. 14-23. – ISSN 1729-3774.
23. Lytvyn V. Development of a method for the recognition of author's style in the ukrainian language texts based on linguometry, stylemetry and glottochronology / V. Lytvyn, V.

Vysotska, P. Pukach, I. Bobyk, D. Uhryn // Eastern-European Journal of Enterprise Technologies. – №4/2(88). – Харків 2017. – P. 10-18. – ISSN 1729-3774.
24. Vasyl Lytvyn, Victoria Vysotska, Petro Pukach, Zinovii Nytrebych, Ihor Demkiv, Roman Kovalchuk, Nadiia Huzyk. Development of the linguometric method for automatic identification of the author of text content based on statistical analysis of language diversity coefficients // Eastern-European Journal of Enterprise Technologies– Vol 5, No 2 (95) . – 2018. – ISSN 1729-3774. – P. 16-28. – http://journals.uran.ua/eejet/article/view/142451/142492.
25. Lytvyn V., Vysotska V., Pukach P., Nytrebych Z., Demkiv I., Kovalchuk R., Huzyk N.: Development of the linguometric method for automatic identification of the author of text content based on statistical analysis of language diversity coefficients, Eastern-European Journal of Enterprise Technologies, 5(2), 16-28 (2018)
26. Levchenko, O., Romanyshyn, N., Dosyn, D.: Method of Automated Identification of Metaphoric Meaning in Adjective + Noun Word Combinations (Based on the Ukrainian Language). In: CEUR Workshop Proceedings, Vol-2386, 370-380. (2019)
27. Шестакевич Т. В. Методи математичної лінгвістики у вирішенні мовознавчих завдань / Т. В. Шестакевич, В. А. Висоцька // Простір і час сучасної науки : Матеріали шостої всеукраїнської науково-практичної інтернет-конф., 22-24 жовтня 2009 р. : тези доп. – Інститут наукового прогнозування, Кримський інститут економіки та господарського права (Севастопольська філія), ТОВ „ТК Меганом”, 2009. – Ч.3. – С. 42-48.
28. Шестакевич Т.В. Застосування породжувальних граматик для генерування речень українською мовою / Т.В. Шестакевич, В.А. Висоцька // Східно-Європейський журнал передових технологій. – Харків, 2012. – № 3/2 (57). – С. 51-53.
29. Висоцька В.А. Особливості генерування семантики речення природною мовою за дапомогою породжувальних необмежених та контекстно-залежних граматик / В.А. Висоцька // Інформаційні системи та мережі. Вісник Національного університету “Львівська політехніка”. – № 783. – Львів, 2014. – Стор. 271-292.
30. Висоцька В.А. Концептуальна модель процесу формування семантики речення природною мовою / В.А. Висоцька // Інформаційні системи та мережі. Вісник Національного університету “Львівська політехніка”, № 805.- Львів 2014 – Стор. 258-278.
31. Висоцька В. А. Особливості моделювання синтаксису речення слов’янських та германських мов за допомогою породжувальних контекстно-вільних грамматик / В.А. Висоцька // Інформаційні системи та мережі. Вісник Національного університету “Львівська політехніка”, № 814.- Львів 2015 – Стор. 246-276.
32. Шестакевич Т.В. Моделювання семантики речення природною мовою за допомогою породжувальних грамматик / Т.В. Шестакевич, В.А. Висоцька, Л.В. Чирун, Л.Б. Чирун // Інформаційні системи та мережі. Вісник Національного університету “Львівська політехніка”, № 814.- Львів 2015 – Стор. 335-352.
33. Бісікало О.В. Експериментальне дослідження пошуку значущих ключових слів україномовного контенту / О.В. Бісікало, В.А. Висоцька // Інформаційні системи та мережі. Вісник Національного університету “Львівська політехніка”. – № 829. – Львів, 2015. – Стор. 255-272.
34. Бісікало О.В. Виявлення ключових слів на основі методу контент-моніторингу україномовних текстів / О.В. Бісікало, В.А. Висоцька // Науковий журнал «Радіоелектроніка. Інформатика. Управління.». – № 1(36). – Запоріжжя: ЗНТУ. –

2016/1. – С. 74-83. – ISSN 1607-3274 (print), ISSN 2313-688X (on-line). – http://ric.zntu.edu.ua/.

35. Бісікало О.В. Застосування методу синтаксичного аналізу речень для визначення ключових слів україномовного тексту / О.В. Бісікало, В.А. Висоцька // Науковий журнал «Радіоелектроніка. Інформатика. Управління.». – № 3(38). – Запоріжжя: ЗНТУ. – 2016/3. – С. 54-65. – ISSN 1607-3274 (print), ISSN 2313-688X (on-line). – http://ric.zntu.edu.ua/.
36. Бісікало О.В. Метод лінгвістичного аналізу україномовного комерційного контенту / О.В. Бісікало, В.А. Висоцька // Інформаційні системи та мережі. Вісник Національного університету "Львівська політехніка", № 854.- Львів 2016 – Стор. 185-204.
37. Висоцька В. Порівняння складності автоматичного опрацювання англійських та українських текстів з врахуванням семантики та синтаксису природних мов / Висоцька В. А., Наум О. // Інформаційні системи та мережі. Вісник Національного університету "Львівська політехніка", № 872.- Львів, 2017. – Стор. 149-162.
38. Чирун Л.Б., Чирун Л.В., Висоцька В.А. Метод визначення авторства текстового україномовного контенту. Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту» (ISDMCI'2018). – 21-27 травня 2018 р., Залізний Порт, Україна. – Стор. 287-289.
39. Davydov, M., Lozynska, O.: Information system for translation into Ukrainian sign language on mobile devices. In: International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 48-51. (2017)
40. Davydov, M., Lozynska, O.: Mathematical method of translation into ukrainian sign language based on ontologies. In: Advances in Intelligent Systems and Computing, 871, 89-100. (2018)
41. Lypak, O.H., Lytvyn, V., Lozynska, O., Vovnyanka, R., Bolyubash, Y., Rzheuskyi, A., Dosyn, D.: Formation of Efficient Pipeline Operation Procedures Based on Ontological Approach. In: Advances in Intelligent Systems and Computing, 871, 571-581. (2019)
42. Savchuk, V., Lozynska, O., Pasichnyk, V.: Architecture of the Subsystem of the Tourist Profile Formation. In: Advances in Intelligent Systems and Computing, 871, 561-570. (2019)
43. Victoria Vysotska. Computer linguistics for online marketing in information technology : Monograph. – Saarbrücken, Germany: LAP LAMBERT Academic Publishing, 2018. – 396 p. – ISBN-13: 978-613-9-84601-6, ISBN-10: 6139846013, EAN: 9786139846016. – Book language: English. – https://www.lap-publishing.com/catalog/details/store/gb/book/978-613-9-84601-6/computer-linguistics-for-online-marketing-in-information-technology?search=vysotska. – Published on: 2018-05-30
44. Vysotska V. Linguistic analysis and modelling semantics of textual content for digest formation / Victoria Vysotska, Lyubomyr Chyrun // MEST Journal (Management Education Science & Society Technologie). – Vol.3 No.1. – PP. 127-148 [Online]. – ISSN 2334-7171, ISSN 2334-7058 (Online), DOI 10.12709/issn.2334-7058. This issue: DOI 10.12709/mest.02.02.02.0. – Режим доступу: http://mest.meste.org/MEST_1_2015/Sadrzaj_eng.html http://mest.meste.org/MEST_1_2015/5_15.pdf.
45. Vysotska V. Features of Text Categorization of Commercial Content / Victoria Vysotska // Computer Science and Information Technologies: Proc. of the IX-th Int. Conf. CSIT'2014, 18-22 November, 2014, Lviv, Ukraine..– Lviv: Publishing Lviv Polytechnic, 2014.– P.5-8.

46. Berko A. Linguistic Analysis for the Textual Commercial Content / Andriy Berko, Victoria Vysotska, Lyubomyr Chyrun // Computer Science and Information Technologies: Proc. of the IX-th Int. Conf. CSIT'2014, 18-22 November, 2014, Lviv, Ukraine..– Lviv: Publishing Lviv Polytechnic, 2014.– P.11-14.
47. Shestakevych T. Modelling of semantics of natural language sentences using generative grammars / Tetiana Shestakevych, Victoria Vysotska, Lyubomyr Chyrun, Liliya Chyrun // Computer Science and Information Technologies: Proc. of the IX-th Int. Conf. CSIT'2014, 18-22 November, 2014, Lviv, Ukraine..– Lviv: Publishing Lviv Polytechnic, 2014.– P.19-22.
48. Vysotska V. Generative regular grammars application to modeling the semantics of sentences in natural language / Victoria Vysotska // Комп'ютерні системи проектування. Теорія і практика, Вісник Національного університету "Львівська політехніка" № 808,- Львів 2014 – Стор.43-56.
49. Vysotska V. Linguistic Analysis of Textual Commercial Content for Information Resources Processing / Victoria Vysotska // Proceedings of the XIIIth International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET'2016). - February 23–26, 2016. Lviv–Slavske, Ukraine.- P. 709-713.
50. Chyrun Lyubomyr. Specifics Informational Resources Processing for Textual Content Linguistic Analysis / Lyubomyr Chyrun, Victoria Vysotska, Vasyl Lytvyn // Proceeding of XIIth International Conference of Perspective Technologies and Methods in MEMS Design, MEMSTECH 2016. – 20-24 April, 2016, Lvi-Polyana, Ukraine. – Lviv Politechnic Publishing House. – P. 214-219.
51. Chyrun L. Informational resources processing intellectual systems with textual commercial content linguistic analysis usage constructional means and tools development / L. Chyrun, V. Vysotska, I. Kozak // Econtechmod : an international quarterly journal on economics in technology, new technologies and modelling processes. – Lublin ; Rzeszow, 2016. – Volum 5, nomber 2. – P. 85–94. – Bibliography: 60 titles.
52. Lytvyn Vasyl. Content Linguistic Analysis Methods for Textual Documents Classification / Vasyl Lytvyn, Victoria Vysotska, Oleh Veres, Ihor Rishnyak, Halya Rishnyak // Computer Science and Information Technologies: Proc. of the XI-th Int. Conf. CSIT'2016, 6-10 September, 2016, Lviv, Ukraine..– Lviv: Lviv Polytechnic Publishing House, 2016.– P.190-192.
53. Фольтович В.М. Метод контент-аналізу текстової інформації Інтернет газети / В.М. Фольтович, М.В. Коробчинський, Л.Б. Чирун, В.А. Висоцька // Комп'ютерні науки та інформаційні технології. Вісник НУ "Львівська політехніка". – № 864. – Львів 2017. – С.7-19.
54. Гасько Р.В. Особливості контент-аналізу користувацької Інтернет-діяльності для формування зрізу психологічного стану особистості / Р.В. Гасько, Л.В. Чирун, В.А. Висоцька // Комп'ютерні науки та інформаційні технології. Вісник НУ "Львівська політехніка". – № 864. – Львів 2017. – С. 221-238.
55. Lytvyn V. Intelligent System Structure for Web Resources Processing and Analysis / V. Lytvyn, V. Vysotska, L. Chyrun, A. Smolarz, O. Naum // 1st International Conference Computational Linguistics and Intelligent Systems, COLINS'2017. – 21 April 2017, Kharkiv. – P. 56-74.
56. Lytvyn V. A Method of Construction of Automated Basic Ontology / V. Lytvyn, V. Vysotska, W. Wojcik, D. Dosyn // 1st International Conference Computational Linguistics and Intelligent Systems, COLINS'2017. – 21 April 2017, Kharkiv. – P. 75-83.