УДК 004.912

Мазниченко Наталя Іванівна
Національний юридичний університет ім. Ярослава Мудрого
(Харків)

АВТОМАТИЧНИЙ АНАЛІЗ ТЕКСТІВ ЕЛЕКТРОННИХ ДОКУМЕНТІВ

Анотація. *Проаналізовані можливості застосування сучасних програмних засобів обчислювальної техніки для змістовного аналізу текстів електронних документів. Розглянуті можливі напрямки застосування відповідних комп'ютерних програм.*

Ключові слова: *аналіз електронних документів, лінгвістична обробка тексту, автоматичний аналіз текстів.*

Maznichenko Natalia
Yaroslav Mudryi National Law University
(Kharkov)

AUTOMATIC ANALYSIS OF TEXTS OF ELECTRONIC DOCUMENTS

Abstract. *Possibilities of application of modern programmatic facilities of the computing engineering are analyses for the rich in content analysis of texts of electronic documents. Possible directions of application of the proper computer programs are considered.*

Keywords: *analysis of electronic documents, linguistic text processing, automatic analysis of texts.*

Постановка проблеми. На даний час розвиток інформаційних комп'ютерних технологій призвів до різкого зростання обсягів інформації, яка має зберігатися, оброблятися, поширюватися за допомогою засобів комп'ютерних систем та мереж. Текст є однією з основних форм обміну інформацією в суспільстві. Текстова інформація в різних форматах складає значну долю інформаційних ресурсів комп'ютерних систем. Саме тому створення і розвиток технологій обробки тексту привертали велику увагу на усіх етапах розвитку інформаційних систем [1, с. 29].

Великі об'єми досліджуваних матеріалів практично унеможливлюють їх якісний «ручний» аналіз, істотно збільшують час, необхідний для проведення досліджень. Стало очевидним, що змістовний аналіз текстів документів в електронній формі неможливий без використання сучасної обчислювальної техніки. Виходом з ситуації, що склалася, є застосування автоматизованих систем аналізу текстів (на потрібній мові), які дозволять при проведенні досліджень текстів електронних документів значно підвищити якість отримуваних результатів, а також істотно скоротити терміни проведення досліджень.

Розвиток комп'ютерних технологій і збільшення ролі інформації на сьогоднішній день відвів методам аналізу тексту електронних документів особливу привілейовану роль. Методи аналізу тексту застосовуються при пошуку, систематизації, оцінці, відборі інформації, діагностиці, аналізі і

прогнозуванні подій або поведінки суб'єкта. Сучасний стан комп'ютерної техніки дозволив на цей час автоматизувати такі трудомісткі процеси як статистичну обробку текстів [2, с. 7].

Методам та моделям пошуку, обробки та аналізу текстової інформації в комп'ютерних системах та мережах присвячені праці таких авторів як П.І. Браславский, М.С. Агеев, Г.Г. Белоногов, Б.В. Добров, І.Є. Кураленок, Д.В. Ланде, Ю.М. Ліфшіц, І.С. Некрест'янов, О.В. Пескова, I. Dagan, S. Dumais, M. Halkidi, T. Joachims, T. Kohonen, D. Lewis, X. Liu, J. Platt, R. Schapire, H. Schutze, F. Sebastiani, Y. Yang, J.Hajic, E. Hajicova, P.Pecina і ряду інших.

За останнє десятиліття намітився прогрес в області обробки текстів електронних документів, проте багато завдань як і раніше залишаються невирішеними, також з'являються і нові задачі, пов'язані з моніторингом соціальних мереж і обробкою спотворених текстів. Усе це примушує інакше поглянути на системи аналізу та обробки текстів. Серед завдань, що вимагають перегляду методів обробки текстів, можна назвати витягування (добування) думок з тексту, визначення емоційного забарвлення текстів, аналіз реального впливу джерел інформації, обробка некоректних або навмисно спотворених текстів. У зв'язку з цим кількість систем для автоматичного аналізу тексту постійно зростає, з'являються нові алгоритми, нові підходи, переглядаються методики аналізу і системи, що використовуються.

Виклад основного матеріалу. Автоматична обробка текстів (АОТ) – це перетворення або аналіз тексту за допомогою ЕОМ.

Рішення практично будь-якої задачі АОТ так чи інакше включає проведення аналізу тексту на декількох рівнях представлення [3, с. 52]:

1. Графематичний аналіз. Виділення з масиву даних речень і слів (токенів).
2. Морфологічний аналіз. Виділення граматичної основи слова, визначення частин мови, приведення слова до словникової форми.
3. Синтаксичний аналіз. Виявлення синтаксичних зв'язків між словами в реченні, побудова синтаксичної структури речення.
4. Семантичний аналіз. Виявлення семантичних зв'язків між словами і синтаксичними групами, витягування (добування) семантичних відношень.

Кожен такий аналіз – самостійне завдання, що не має власного практичного застосування, але активно використовувана для вирішення загальних завдань.

Останнім часом часто згадують прагматичний аналіз (дискурс-аналіз) в якості ще одного рівню аналізу тексту. Його основне завдання – визначити мету, яку переслідує автор при викладі своїх думок. Але в процес побудови систем автоматичної обробки текстів даний етап не включений, тому що формалізація і автоматизація його доки залишається тільки на рівні обговорень [4, с. 9].

Кожен з перерахованих рівнів аналізу і обробки тексту представлений низкою відповідних комп'ютерних програм. Найбільш повний і актуальний список інструментів для автоматичного аналізу текстів в електронній формі приведений на сайті AskNet [5]. Деякі з представлених програм реалізують окремі рівні аналізу, а деякі дозволяють проводити одночасно аналіз текстів на декількох з перерахованих вище рівнях. Частина з цих програм поширюється

безкоштовно і зацікавлений користувач може встановити їх на своєму комп'ютері і перевірити у дії, інші мають демоверсии, які хоч і обмежені за функціональними можливостями, але цілком адекватно дозволяють представити можливості програми. Слід відзначити, що наведені приклади програм далеко не вичерпують усієї сфери програмних засобів щодо автоматичного дослідження текстів.

Якщо ж користувачів зацікавить інформація стосовно того, на яких мовах дозволяють обробляти тексти сучасні програми аналізу текстів електронних документів, доцільно звернутись до сайту NLPub [6]. Після дослідження представлених програм можна зробити сумний висновок, що систем, які обробляють тексти українською мовою, майже немає (я знайшла тільки одну – «Лингвистические компоненты»).

Безумовно, жодна програмна обробка тексту не може замінити собою аналіз, який може здійснити людина, – особливо експерт в тій або іншій області.

Актуальність теми дослідження можна обґрунтувати також наведенням можливих сфер застосування сучасних програм аналізу та обробки електронних документів.

Перерахуємо деякі прикладні завдання автоматичного аналізу текстів:

1. Завдання визначення авторства текстів. Для визначення автора тексту часто доводиться звертатися до експертів. Експерти можуть ідентифікувати автора невідомого тексту або визначити приналежність твору іншому авторові за допомогою характерних мовних особливостей, стилістичних прийомів.

2. Завдання визначення тематичної приналежності текстів.

Тематична приналежність текстів може бути корисна для:

– автоматичної рубрикації текстів по темах;

– визначення релевантна документів пошуковому запиту в пошукових системах;

– ігнорування спамдексінгу. Спамдексінг (пошуковий спам) – зловживання частотою ключових слів з метою маніпулювання пошуковими системами; загальна назва для методів неетичного просування сайтів, так звана «чорна оптимізація», або пошукова оптимізація (Search Engine Optimization, SEO).

3. Завдання аналізу тональності текстів.

Аналіз тональності текстів (визначення емоційного забарвлення текстів, сентимент-аналіз, Sentiment analysis, Opinion mining) – область комп'ютерної лінгвістики, яка займається вивченням думок і емоцій в текстових документах.

Аналіз тональності знаходить своє застосування в:

– соціології – збір різного роду даних з соціальних мереж (наприклад, про релігійні погляди);

– медицині і психології – виявлення психічних особливостей і особливостей поведінки конкретних користувачів і груп користувачів. Наприклад, також можна визначати тенденцію до виникнення депресії у користувачів соціальних мереж у зв'язку з політичними, економічними труднощами;

– маркетингу – наприклад, можна оцінити фільм, ресторан, готель і ін.;

– політології – збір свіжих даних з блогів про політичні погляди населення.

Висновки. Розглянувши програми для автоматичної обробки та аналізу текстів, можна зробити висновок, що аналіз, який може здійснити людина-експерт в різних областях, навряд чи зможе замінити програмна обробка тексту. Але ці програми можуть дозволити людині прийти до висновків, витративши на проведення дослідження меншу кількість часу. Також ці програми дозволяють випробувати гіпотези на набагато більшому об'ємі матеріалу і з великою часткою упевненості в об'єктивності отриманих даних.

Таким чином, завдяки комп'ютерам зараз вдається спростити або зробити непотрібними багато класичних операцій обробки і підготовки інформації. При цьому слід зауважити, що автоматичні системи аналізу текстів відіграють істотно підпорядковану і підготовчу роль для подальшої вдумливої роботи фахівців, оснащених перевіреними методиками якісного дослідження.

Нині в області автоматичної обробки текстів значна частина робіт присвячена перенесенню методів, розроблених для англійської мови, на російську, і, на жаль, оригінальних розробок дуже мало. Також слід відзначити, що комп'ютерних систем обробки та аналізу текстів в електронній формі на український мові, на жаль, майже немає, тому є простір для продуктивної наукової та дослідницької діяльності.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ ТА ЛІТЕРАТУРИ:

1. В.В. Диковицкий, М.Г. Шишаев. Обработка текстов естественного языка в моделях поисковых систем / Сборник научных трудов Кольского научного центра РАН. Выпуск № 3, 2010. С. 29-34.
2. Ланде Д.В. Елементи комп'ютерної лінгвістики в правовій інформатиці. – К.: НДІІП НАПрН України, 2014. – 168 с.
3. Дмитрий Ильвовский, Екатерина Черняк. Системы автоматической обработки текстов. // Открытые системы. – 2014. – № 01. – С. 51-53. Електронний ресурс. Режим доступу https://www.osp.ru/os/2014/01/13039687/
4. Батура, Т. В. Математическая лингвистика и автоматическая обработка текстов: учеб. пособие / Т. В. Батура; Новосиб. гос. ун-т. – Новосибирск: РИЦ НГУ, 2016. – 166 с.
5. Программы лингвистического анализа и обработки текста. Електронний ресурс. Режим доступу: http://www.asknet.ru/Analytics/programms.htm
6. Обработка текста. Електронний ресурс. Режим доступу: https://nlpub.ru/%D0%9E%D0%B1%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%BA%D0%B0_%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%B0