УДК 004.9

ВИЗНАЧЕННЯ КЛЮЧОВИХ СЛІВ АНГЛОМОВНОГО ТЕКСТУ З ВИКОРИСТАННЯМ ТЕХНОЛОГІЇ DKPRO CORE

Олександр Яхимович, студент групи 1КСУА-14б, Вінницький національний технічний університет (ВНТУ), Україна
Науковий керівник – **Олег Бісікало**, д-р техн. наук, професор, в.о. декана ФКСА, ВНТУ, Україна

Завдання виділення ключових слів з тексту виникає у бібліотечній справі, лексикографії та термінознавстві, а також в задачах інформаційного пошуку. В даний час обсяги і динаміка інформації, яка підлягає обробці в цих областях, роблять особливо актуальною задачу автоматичного визначення ключових слів, які можуть використовуватися для створення і розвитку термінологічних ресурсів, а також для ефективної обробки документів.

Ключове слово – це таке слово в тексті, яке здатне в сукупності з іншими ключовими словами представляти зміст тексту.

У всіх текстових документах, що створені людиною, можна виділити статистичні закономірності. Їх визначення базується на використанні законів Ципфа [1]. Найбільшою популярністю для визначення ключових слів користується модель TF*IDF [2]. Однак застосовуються й інші індексуючі функції, включаючи ймовірні способи індексування [3] та методики індексування структурованих документів [4]. Інші функції індексації можуть знадобитися в тих випадках, коли спочатку навчальну множину не дано і частоту для документа не вдається порахувати. У цих випадках TF*IDF змінюють на емпіричні функції [5]. При використанні цього підходу не виключена ймовірність попадання в ключові слова випадкових спеціальних термінів, рідкісних слів, власних імен та іншого «вербального шуму».

Проте результати парсерингу природних мов за допомогою сучасних лінгвістичних пакетів дозволяють на доступному програмному рівні [6] оперувати синтаксичними зв’язками між словами окремого речення. Одним з таких пакетів є DKPro Core – це набір програмних компонентів для обробки природної мови, заснований на Apache UIMA framework. Він був побудований з метою підвищення продуктивності дослідників, які займаються автоматичним аналізом мови. Колекція прагне досягти цієї мети, слідуючи таким принципам:

1. Вибір – для більшості кроків аналізу DKPro Core включає кілька різних інструментів від різних постачальників. DKPro Core охоплює такі завдання аналізу, як визначення мови, лематизація, морфологічний аналіз, синтаксичний розбір, розбір залежностей, сегментація, маркування смислових ролей, перевірка орфографії та морфології.
2. Покриття – DKPro Core об'єднує 94 моделі на 15 природних мовах.
3. Взаємозамінність – імена параметрів компонентів однакові, а де це можливо, компоненти приймають однакові параметри для різних ресурсів.
4. Переносимість – компоненти аналізу завантажуються та працюють на різних платформах системи або за допомогою віртуальної машини Java, або

шляхом використання бінарних файлів, скомпільованих для різних операційних систем.

5. Зручність – компоненти аналізу вимагають тільки мінімальної обов'язкової конфігурації і багато компонентів не потребують обов'язкового налаштування взагалі [7].

На основі DKPro Core було розроблено програму, що визначає ключові слова в англомовному тексті. Для проведення експерименту було взято текст тез [8], де ключові слова задані авторами: Software product line, Variability, Adaptive object model, Reflection, Brazilian Satellites Launcher. Власна розробка знаходить найбільше слів заданих автором – 6 слів. Аналоги знаходять по 5 слів. Власна розробка, так само як і аналоги, знаходить перших 4 слова заданих автором, але не знаходить п’яте – adaptive, проте вона знаходить такі ключові слова: model, reflection, чого не роблять аналоги.

Література

1. Андреев, А. М. Модели и методы автоматической классификации текстовых документов [Электронный ресурс] / А. М. Андреев, Д. В. Березкин, В. В. Сюзев, В. И. Шабанов // Вестник МГТУ им. Н. Э. Баумана. Сер. Приборостроение. – 2003. – №4. – Режим доступа: \www/URL: http://vestnikprib.bmstu.ru/articles/397/html/files/assets/basic-html/page1.html. – 21.01.2015.
2. Joachims, T. Text categorization with Support Vector Machines: Learning with many relevant features [Text] / T. Joachims // Machine Learning: ECML-98 Lecture Notes in Computer Science. – 1998. – Vol. 1398. – P. 137–142.
3. Jensen, R. A Rough Set-Aided System for Sorting WWW Bookmarks [Electronic resource] / R. Jensen. – The University of Edinburgh, 2000. – Available at: \www/URL: http://users.aber.ac.uk/rkj/research/mscthesis.pdf. – 21.01.2015.
4. Larkey, L. S. Combining classifiers in text categorization. [Text] / L. S. Larkey, W. B. Croft // Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’96. – ACM Press, 1996. – P. 289-297.doi:10.1145/243199.243276
5. Scott, S. Text Classification Using WordNet Hypernyms [Electronic resource] / S. Scott, S. Matwin. – University of Ottawa, 1998. – Available at: \www/URL: http://www.aclweb.org/anthology/W98-0706. – 21.01.2015.
6. Бісікало, О. В. Концептуальна модель системи образного аналізу і синтезу природно-мовних конструкцій [Текст] / О. В. Бісікало // Математичні машини і системи. – 2013. – № 2. – С. 184–187. – ISSN 1028-9763.
7. Natural Language Processing: Integration of Automatic and Manual Analysis [Electronic resource]. – Technischen Universität Darmstadt, 2014. – Available at: \www/URL: http://tuprints.ulb.tu-darmstadt.de/4151/1/rec-thesis-final.pdf. – 21.01.2015.
8. Burgareli, L. A. (2009, Jul.-Dec.). Variability management in software product lines using adaptive object and reflection. Journal of Aerospace Technology and Management, V. 1, № 2. Available: http://www.jatm.com.br/papers/vol1_n2/JA-TMv1n2_thesis_abstracts.pdf. Last accessed 21.01.2015.