УДК 004.93:159.95

О. В. Бісікало
О. В. Яхимович

ЗМЕНШЕННЯ ВЕРБАЛЬНОГО ШУМУ ПРИ ВИЗНАЧЕННІ КЛЮЧОВИХ СЛІВ

Вінницький національний технічний університет

Анотація

Розглядається вербальний шум при визначенні ключових слів і можливі шляхи його зменшення, а також можливість покращити результати роботи методу визначення ключових слів використовуючи підходи до зменшення кількості шумових слів.

Ключові слова: вербальний шум, ключові слова, стоп слова, DKPro Core, словосполучення.

Abstract

Consider the verbal noise in determining the keywords and possible ways to reduce it, and the ability to improve the results of the method for determining keywords using approaches to reducing the count of noise words.

Keywords: verbal noise, keywords, stop words, DKPro Core, phrase.

Вступ

В даний час обсяги і динаміка інформації, яка підлягає обробці в бібліотечній справі, лексикографії та термінознавстві, а також в задачах інформаційного пошуку, роблять особливо актуальною задачу автоматичного визначення ключових слів, які можуть використовуватися для створення і розвитку термінологічних ресурсів, а також для ефективної обробки документів: індексування, реферування, кластеризації та класифікації [1].

Метою роботи є розробка підходів для зменшення кількості вербального шуму при визначенні ключових слів та їх застосування для покращення результатів роботи методу визначення ключових слів.

Результати дослідження

Вербальний шум або шумові слова – термін з теорії пошуку інформації за ключовими словами. Це такі слова, які не несуть смислового навантаження, тому їх користь та роль для пошуку не суттєва [2].

В процесі обробки проводиться виключення з досліджуваного тексту слів, які за визначенням не можуть бути значущими тому, що складають «шум». На відміну від ключових ці слова називаються нейтральними або стоповими (стоп-словами). Такими є слова, що відносяться до службових частин мови, а також займенники [3].

Для розробленого авторами методу визначення ключових слів англомовного тексту на основі інструментальних засобів пакету DKPro Core [4] було розроблено додаткові модулі (рис. 1.). У результаті зменшення кількості шумових слів досягнуто за допомогою підходів: заміна займенників на відповідні до них іменники; вилучення словосполучень із типами зв’язків, які не несуть суттєвого смислового навантаження; вилучення слів, які відносяться до неінформативних частин мови; вилучення слів, які відносяться до списку стоп слів.

Заміна займенників на відповідні до них іменники (replace pronouns): дозволяє зменшити кількість займенників, а також збільшити кількість іменників, які можуть бути ключовими словами. Для методу визначення ключових слів англомовного тексту, запропонованому в [4], заміна займенників здійснюється засобами DKPro Core [5].

Вилучення словосполучень із типами зв’язків, які не несуть суттєвого смислового навантаження. Для англомовних текстів, такими типами зв’язків є: визначник (DET): the, which; знаки пунктуації (PUNCT); словосполучення з there або it в експозиційних конструкціях (EXPL) [6].

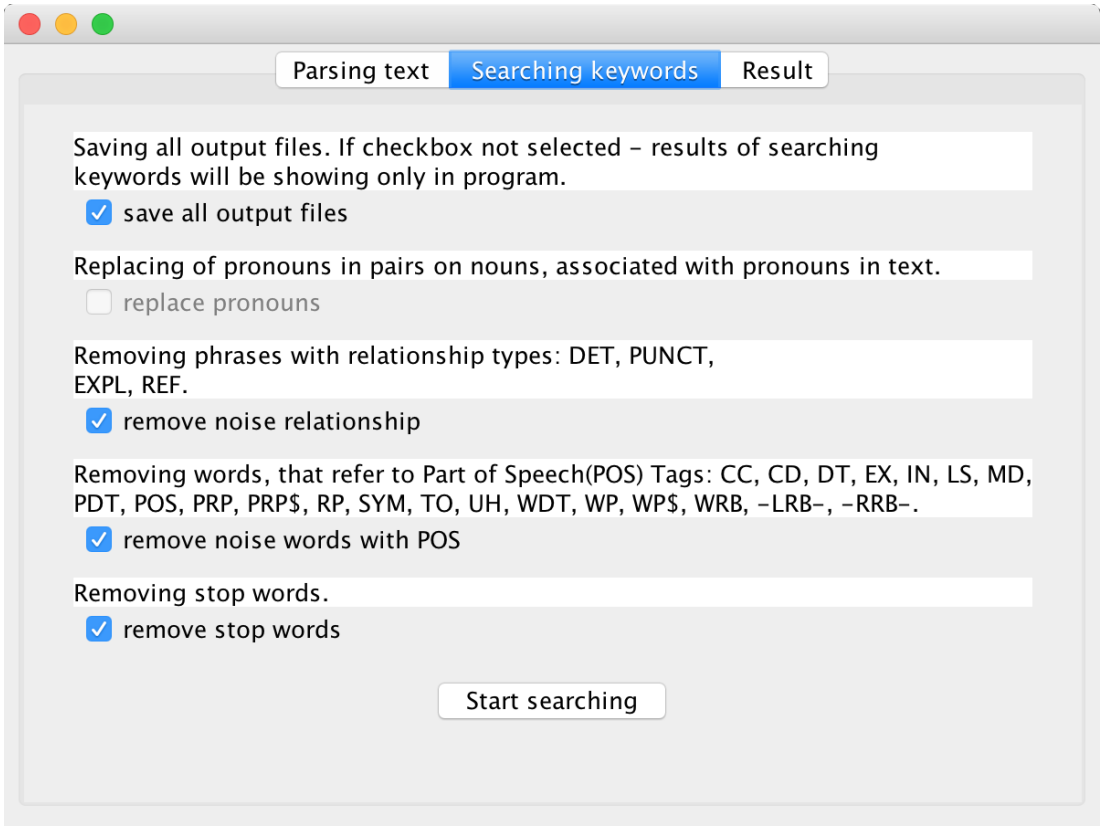

Рис. 1. Модулі зменшення кількості шумових слів

Вилучення слів, які відносяться до неінформативних частин мови (remove noise words with POS). Для англійської мови такими частинами мови є: CC, CD, DT, EX, IN, LS, MD, PDT, POS, PRP, PRP$, RP, SYM, TO, UH, WDT, WP, WP$, WRB, -LRB-, -RRB- [7].

Вилучення слів, які відносяться до списку стоп-слів. Список слів дня англомовних текстів описаний в [8].

Для апробації запропонованих підходів було проведено експеримент з текстом [9], який складається з 1588 слів, де ключові слова задані автором: Participatory, Historical geography at large, Archival activism, Canada, Aboriginal rights. Перші дев'ять ключових слів, без використання підходів до зменшення шуму, будуть: the, be, geography, have, work, geographer, participatory, to, history. З використання підходів до зменшення шуму, маємо такі перші дев'ять ключових слів: geography, geographer, participatory, research, history, learn, community, historical, issue.

Отже, використання підходів до зменшення шуму дозволило покращити результати визначення ключових слів – знайти три слова з заданих автором (participatory, historical, geography), тоді як без зменшення шуму було знайдено два слова заданих автором(participatory, geography)

Висновки

Використовуючи підходи до зменшення кількості шумових слів вдалося зменшити кількість шуму при визначення ключових слів англомовного тексту на основі інструментальних засобів пакету DKPro Core.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Ершов Ю. С. Выделение ключевых слов в русскоязычных текстах / Ю. С. Ершов // Молодежный научно-технический вестник. – М.: ФГБОУ ВПО "МГТУ им. Н.Э. Баумана", 2014. – № ФС77-51038. – С. 70-79.

2. Гращенко Л. А. О модельном стоп-словаре / Л. А. Гращенко // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук - 2013. - № 1 (150). - С. 40-46.

3. Андреев А. М. Модели и методы автоматической классификации текстовых документов / А. М. Андреев, Д. В. Березкин, В. В. Сюзев, Шабанов В.И. // Вестн. МГТУ. Сер. Приборостроение. – М.: МГТУ, 2003. – №3. – С. 64-94.

4. Bisikalo O.V. Method of determining of keywords in English texts based on the DKPro Core / Bisikalo, O.V., Wójcik, W., Yahimovich, O.V., Smailova, S. // Proceedings of SPIE 10031, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016. - Wilga, Poland 28 September 2016. - DOI:10.1117/12.2249225.

5. Natural Language Processing: Integration of Automatic and Manual Analysis [Електронний ресурс]. – Режим доступу: http://tuprints.ulb.tu-darmstadt.de/4151/1/rec-thesis-final.pdf – Назва з екрану.

6. English grammatical relations [Електронний ресурс]. – Режим доступу: http://universaldependencies.org/en/dep/ – Назва з екрану.

7. Word level. Bracketing Guidelines for Treebank II Style Penn Treebank Project [Електронний ресурс]. – Режим доступу: http://www.surdeanu.info/mihai/teaching/ista555-fall13/readings/PennTreebankConstituents.html – Назва з екрану.

8. Bougé K. Lists of stop words [Електронний ресурс]. – Режим доступу: https://sites.google.com/site/kevinbouge/stopwords-lists – Назва з екрану.

9. Cameron L. J. Participation, archival activism and learning to learn [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S0305748814001030 – Назва з екрану.

Олег Володимирович Бісікало — доктор технічних наук, професор, декан факультету комп'ютерних систем і автоматики, Вінницький національний технічний університет, Вінниця.

Олександр Вікторович Яхимович — аспірант кафедри автоматики та інформаційно-вимірювальної техніки, факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, Вінниця, e-mail: yahimovich.olexandr@gmail.com.

Oleg V. Bisikalo — Doctor of Engineering, Professor, Dean of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia

Alexander V. Yahimovich — Department Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, email: yahimovich.olexandr@gmail.com.