УДК 004.93:159.95

О. В. Бісікало
О. В. Яхимович

ФОРМАЛІЗАЦІЯ ЗАДАЧІ ВИЗНАЧЕННЯ КЛЮЧОВИХ СЛІВ ТЕКСТУ

Вінницький національний технічний університет

Анотація

Розглядається задача визначення ключових слів, формальне представлення тексту, зв'язків між словами, ключових слів.

Ключові слова: ключові слова, морфологічний аналіз, множина зв'язків тексту, словосполучення.

Abstract

Consider the problem of the determining keywords, formal representation of the text, relations between words, keywords.

Keywords: keywords, morphological analysis, set of text relations, phrase.

Вступ

Основний зміст документа (тексту) може бути представлений за допомогою певних слів, узятих безпосередньо з цього тексту. Як правило, до кожного розгорнутого тексту можна скласти цілий набір ключових слів різного обсягу (від 5 до 15 слів). Але взагалі кількість ключових слів може варіюватися в широких межах [1].

Метою роботи є формалізація тексту, зв'язків між словами, а також задачі визначення ключових слів.

Результати дослідження

Розглянемо довільний текст як множину синтаксично зв'язаних упорядкованих слів, які, в свою чергу, є підмножиною слів мови: $w \in \{W\} \subset \{\omega\}$, де w – слово, W – множина слів в тексті, ω – множина слів мови [2].

Відомо, що кожне слово має основну форму (канонічна форма, лема), до якої можна звести слово шляхом його морфологічного аналізу. Формально представимо це застосуванням до слова функції нормалізації m морфологічного аналізу: $m(w) = b$, $w \in \{W_i\}$, де $\{W_i\}$ – множина словоформ i-го слова, b – основна форма слова. Відомі дві основні властивості функції нормалізації [2]: в результаті нормалізації основної форми слова отримуємо основну форму слова, що показано у формулі (1); для будь-якого слова нормалізація дає основну форму слова, яка належить множині словоформ, що зазначено у формулі (2):

$$m(b) = b. \tag{1}$$

$$\forall w(w \in \{W\}) \Rightarrow m(w) = b,\ b \in \{W\}. \tag{2}$$

Отже, для задачі визначення ключових слів, текст T можна представити, як впорядкований набір слів w_i і символів c_i:

$$T = \{w_i, c_i\},\ w_i \in \{W\} \subset \{\omega\},\ c_i \in \{C\} \subset \{\varsigma\}, \tag{3}$$

де $\{W\}$ – множина слів в тексті, що є підмножиною $\{\omega\}$ – множини слів мови; c_i – знаки пунктуації, цифри, пробіли, переходи на новий рядок та інші символи, що не є буквами [3]; $\{C\}$ –

множина символів в тексті, що є підмножиною $\{\varsigma\}$ – множини всіх символів; i – порядковий номер слова або не буквенного символу в тексті.

Між словами w і символами c існують зв'язки R в тексті, які можуть бути трьох видів: зв'язок між словом і словом: $(w_i, w_l) \in R_j$, де j – порядковий номер зв'язку в тексті; зв'язок між словом і символом: $(w_i, c_l) \in R_j$; зв'язок між символом і символом: $(c_i, c_l) \in R_j$.

Будемо вважати ключовими словами K такі, які коротко представляють сутність тексту T:

$$K(T) = \{k_n\},\ k_n \in \{\omega\}, \tag{4}$$

де k_n – ключове слово, що належить множині слів мови $\{\omega\}$.

Розглянемо процес знаходження ключових слів як згортку текстової інформації за певними критеріями. У цьому випадку ключові слова не завжди можливо формально визначити в тексті. Також, в якості ключових, можуть використовуватися: синоніми, терміни з якими текст може бути пов'язаний логічно, власні назви, з якими асоціюється текст [4-5]. Найперше будемо аналізувати варіант, коли всі ключові слова знаходяться в тексті K_T. Наглядно такий випадок представлено спільним сектором на рисунку 1.

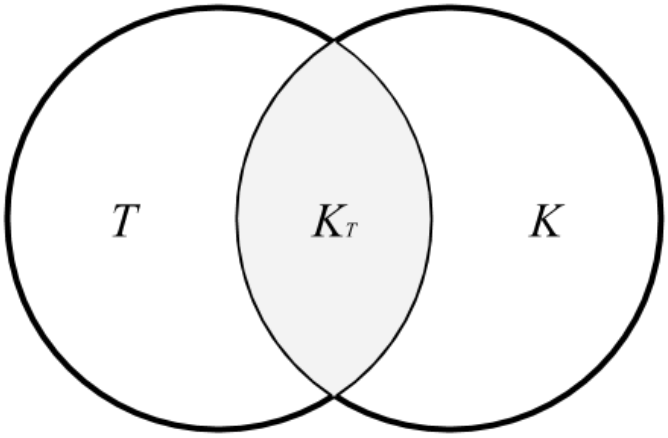

Рис.1. Множини слів в тексті і ключових слів

Отже, формально задачу визначення ключових слів K_T для тексту T можна описати як знаходження таких слів, які належать цьому тексту W_T і входять до множити слів мови:

$$K_T(T) = \{k_n\},\ k_n \in \{W_T\} \subset \{\omega\}. \tag{5}$$

Вважатимемо, що ключове слово k_n є словом w з тексту T, яке приведено до основної форми b:

$$k_n = b = m(w),\ w \in \{WF\}, \tag{6}$$

де $\{WF\}$ – множина словоформ одного слова; $m(w)$ – функція нормалізації морфологічної форми слова w.

Позначемо словосполучення як:

$$G - [DT] \rightarrow D, \tag{7}$$

де G – головне слово (Governor); $[DT]$ – тип зв'язку (Dependency Type); D – залежне слово (Dependent), при чому для зручності розгляду зв'язків слова з (7) краще представляти у формі (5), але для визначення ключових слів – у формі (6).

Представлення типових зв'язків Stanford Dependencies було розроблено таким чином, щоб забезпечити простий опис граматичних відношень у реченні [6-7]. Такий опис можна легко зрозуміти

і ефективно використовувати людьми без лінгвістичного досвіду, які хочуть програмно визначити текстові зв’язки. Зокрема, замість фразово-структурних уявлень, які давно домінують у співтоваристві комп’ютерної лінгвістики, Stanford Dependencies репрезентують всі зв’язки в реченні рівномірно, як типізовані залежності (фактично, як трійки відношень між парами слів). Це просте, однорідне представлення є цілком природним для не лінгвістів, які думають про завдання, пов'язані з вилученням інформації з тексту, а також ефективно для додатків, що забезпечують програмний доступ до зв’язків.

Висновки

В роботі, формалізовано задачу визначення ключових слів, а також представлено формально текст, зв'язки між словами, ключові слова. Представлення типових зв'язків Stanford Dependencies є цілком природним для не лінгвістів, які думають про завдання, пов'язані з вилученням інформації з тексту, а також ефективно для додатків, що забезпечують програмний доступ до зв’язків.

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. Кулешов С.В. Ассоциативно-онтологический подход к обработке текстов на естественном языке / С.В. Кулешов, А.А. Зайцева, В.С. Марков // Интеллектуальные технологии на транспорте. – 2015. – № 4. – С. 40-45.

3. A set of Unicode character values [Електронний ресурс]. – Режим доступу: http://www.unicode.org/reports/tr44/#General_Category_Values. – Назва з екрану.

4. Как составить список ключевых слов? [Електронний ресурс]. – Режим доступу: http://blog.creativeconomy.ru/2009/04/02/kak-sostavit-spisok-klyuchevyx-slov/. – Назва з екрану.

5. Абрамов Е. Г. Подбор ключевых слов для научной статьи / Е. Г. Абрамов // Научная периодика: проблемы и решения. – 2011. – № 1(2). – С. 35-40.

6. Universal Dependency Relations [Електронний ресурс] – Режим доступу до ресурсу: http://universaldependencies.org/u/dep/. – Назва з екрану.

7. Manning C. Stanford typed dependencies manual [Електронний ресурс] / C. Manning, M. de Marneffe. – 2016. – Режим доступу до ресурсу: https://nlp.stanford.edu/software/dependencies_manual.pdf. – 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 автоматизації та інформаційно-вимірювальної техніки, факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, Вінниця, e-mail: yahimovich.olexandr@gmail.com.

Oleg V. Bisikalo — Dr.Sc. (Eng.), Professor, Dean of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia

Alexander V. Yahimovich — Postgraduate student, Department Of Automation And Information Measuring Devices, Vinnytsia National Technical University, Vinnytsia, email: yahimovich.olexandr@gmail.com.