13. Pardo A., Lacey A. L., Fernández V. M., Fan H.-J., Fan Y., Hall M. B. *J Biol Inorg Chem* 2006, 11, 286–306.
14. Wu H., Hall M. B. *Comptes Rendus Chimie* 2008, 11, 790–804.
15. Siegbahn P. E. M., Tye J. W., Hall M. B. *Chem. Rev.* 2007, 107, 4414–4435.
16. Perdew J. P., Burke J., Ernzerhof M. *Phys. Rev. Lett.* 1996, 77, 3865.
17. Becke A. D. *J. Chem. Phys.* 1993, 98, 5648.
18. Stephens P. J., Devlin F. J., Chablowski C. F., Frisch M. J. *J. Phys. Chem.* 1994, 98, 11623.
19. Tao J., Perdew J., Staroverov V., Scuseria G. *Phys. Rev. Lett.* 2003, 91, 146401.
20. Dolg M. et. al. *Chem. Phys.* 1987, 86, 866–872.
21. Rassolov V. A. et. al. *J. Chem. Phys.* 1998, 109, 1223–1229.
22. Chen S., Raugei S., Rousseau R., Dupuis M., Bullock R. M. *J. Phys. Chem.* 2010, 114, 12716.
23. Weigend F., Ahlrichs R. *PCCP.* 2005, 7, 3297.
24. Fan H. J., Hall M. B. *J. Am. Chem. Soc.* 2002, 124, 394–395.
25. Yson R. L., Gilgor J. L., Guberman B. A., Varganov S. A. *Chem. Phys. Lett.* 2013, 577, 138–141.

© Калякин, Д. С., Варганов С. А., Лыхин А. О.,
Кузубов А. А., Краснов П. О., 2013

УДК 004.021

ПОСТРОЕНИЕ ЧАСТОТНЫХ СЛОВАРЕЙ НА ОСНОВЕ ИСХОДНЫХ ТЕКСТОВ С ПРИМЕНЕНИЕМ ЛЕММАТИЗАЦИИ

И. В. Ковалев, А. И. Середин, М. В. Карасева, П. В. Зеленков, В. В. Храпунова

Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: alexzanderIV@yandex.ru

Рассматривается вопрос снижения трудоемкости изучения информационно-терминологического базиса за счет уменьшения объема частотного словаря, на основе которого строится базис. Рассматривается вопрос построения частотных словарей на основе исходных текстов с применением лемматизации для последующего формирования информационно-терминологического базиса. Приводится алгоритм построения частотного словаря на основе исходных текстов с применением лемматизации, а также модификация данного алгоритма с использованием проверки терминов формируемого частотного словаря по специализированному словарю.

Ключевые слова: частотный словарь, информационно-терминологический базис, лемматизация.

FREQUENCY DICTIONARY CONSTRUCTION BASED ON THE SOURCE TEXT USING LEMMATIZATION

I. V. Kovalev, A. I. Seredin, M. V. Karaseva, P. V. Zelenkov, V. V. Khrapunova

Siberian State Aerospace University named after academician M. F. Reshetnev
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660014, Russian Federation
E-mail: alexzanderIV@yandex.ru

The issue of reducing the complexity of the information-vocabulary basis study by decreasing the amount of the frequency dictionary (on which base the basis is constructed), is considered. The frequency dictionary construction based on the source text using lemmatization for the subsequent formation of information-vocabulary basis is considered. The algorithm for frequency dictionary construction based on the source text using lemmatization is presented, as well as the modification of this algorithm with checking the terms of the generated frequency dictionary by a specialized dictionary.

Keywords: frequency dictionary, information-vocabulary basis, lemmatization.

В соответствии с работами [1–2], занимающимися решением задачи формирования и анализа мультилингвистического информационно-терминологического базиса (ИТБ), основным фактором, влияющим на трудоемкость изучения ИТБ является объем этого базиса. Для существенного снижения трудоемкости

изучения ИТБ, целесообразно уменьшить его объем, но при этом важно не допустить снижения качества. Уменьшения объема ИТБ можно добиться за счет уменьшения частотного словаря, на основе которого строится ИТБ. Но при искусственном уменьшении частотного словаря (например, при его урезании до

определенного размера или выборки слов, принадлежащих только к одной части речи) может произойти снижение качества получаемого ИТБ, так как часть относительно важных слов для изучения может быть потеряна, а менее важные будут включены в базис.

Построение частотных словарей на основе оригинальных текстов. Очевидным решением вышеописанных проблемы является построение частотного словаря только на основе тех текстов, с которыми, предполагается работать обучаемому. Будем называть такие тексты исходными. Например, такой подход может быть применен для быстрой наработки словарного запаса сотрудников крупных межнациональных корпораций, где используется большое количество документации на разных языках. Так, для обучающихся, работающих с техническими инструкциями, для достижения основной цели обучения будет достаточно построить частотный словарь на основе этих технических инструкций. Таким образом, в ИТБ попадут только те понятия, которые будут непосредственно встречаться обучающимся в ходе их дальнейшей работы, что будет способствовать более качественному и быстрому обучению.

Кроме того, частотные словари для текстов одной предметной области целесообразно строить автоматически. Это позволит достаточно быстро и легко сгенерировать частотный словарь для какой-либо предметной области без необходимости долгого ручного построения такого словаря.

В данной статье рассматривается способ автоматического построения частотного словаря на основе исходных текстов с применением лемматизации.

Построение частотных словарей с применением лемматизации. Лемматизация – это процесс приведения словоформы к лемме – ее нормальной (словарной) форме [3].

Благодаря лемматизации в частотный словарь будут заноситься не все словоформы одного и того же термина, встречающиеся в исходных текстах, а только одна (нормальная) форма слова, что приведет к существенному снижению объема частотного словаря.

Элемент частотного словаря представляет собой структуру данных, состоящую из следующих полей:

- 1) слово на исходном языке;
- 2) абсолютная частота встречаемости слова;
- 3) относительная частота встречаемости слова.

Алгоритм построения частотного словаря с применением лемматизации:

1. Разбить исходный текст на слова и занести их в массив слов `word[]`.
2. Проверить, достигнут ли конец массива слов `word[]`:
 - а) если дошли до конца массива, то перейти к шагу 7;
 - б) иначе выполнить следующие действия.
3. Взять очередное слово из массива слов (`word[i]`).
4. Провести лемматизацию текущего слова `word[i]`.
5. Поместить полученную лемму в текущий терм (`term`).

6. Проверить, есть ли полученный терм (`term`) в частотном словаре:

а) если терм (`term`) уже есть в частотном словаре, то увеличить частоту встречаемости данного термина на 1;

б) иначе (полученного термина (`term`) еще нет в частотном словаре) занести `term` в частотный словарь с частотой встречаемости, равной 1.

7. Перейти к следующему слову из массива слов ($i = i+1$).

8. Рассчитать относительную частоту встречаемости для каждого термина из частотного словаря.

9. Отсортировать частотный словарь (либо в алфавитном порядке, либо по частоте встречаемости термов).

10. Сохранить частотный словарь.

11. Завершить работу алгоритма.

Однако при таком подходе в частотный словарь попадут все леммы из исходного текста, в том числе и общеупотребительная лексика, которая будет иметь высокую частоту встречаемости и может не соответствовать задачам обучения. Например, если стоит задача изучения терминологии предметной области, а не лексики в целом, то информационно-терминологический базис, полученный на основе такого частотного словаря, будет иметь больший объем и сравнительно низкое качество.

Кроме того, при лемматизации возможна потеря части терминов. Например, два термина `development` (разработка) и `developer` (разработчик) после лемматизации дадут один термин `develop` (разрабатывать), что опять же приводит к снижению качества информационно-терминологического базиса.

Поэтому для решения этих проблем можно применить метод сверки генерируемого частотного словаря с электронным специализированным словарем для конкретной предметной области. Таким образом, можно исключить из частотного словаря общеупотребительные выражения и лексику, оставив только термины и выражения, характерные для изучаемой предметной области, а также избежать потери терминов.

При таком подходе модифицированный алгоритм построения частотного словаря примет вид:

1. Разбить исходный текст на слова и занести их в массив слов `word[]`.
2. Проверить, достигнут ли конец массива слов `word[]`:
 - а) если дошли до конца массива, то перейти к шагу 7;
 - б) иначе выполнить следующие действия.
3. Взять очередное слово из массива слов (`word[i]`).
4. Проверить, есть ли текущее слово в специализированном словаре:
 - а) Если `word[i]` есть в специализированном словаре, то поместить текущее слово в текущий терм (`term = word[i]`);
 - б) иначе (текущего слова нет в специализированном словаре) провести лемматизацию текущего слова `word[i]`;

в) проверить, есть ли полученная лемма w_lem в специализированном словаре:

г) если w_lem есть в специализированном словаре, то поместить полученную лемму слова в текущий терм (term);

д) иначе перейти к шагу 6.

5. Проверить, есть ли полученный терм (term) в частотном словаре:

а) если терм (term) уже есть в частотном словаре, то увеличить частоту встречаемости данного термина на 1;

б) иначе (полученного термина (term) еще нет в частотном словаре) занести терм в частотный словарь с частотой встречаемости, равной 1.

6. Перейти к следующему слову из массива слов ($i = i+1$).

7. Рассчитать относительную частоту встречаемости для каждого термина из частотного словаря.

8. Отсортировать частотный словарь (либо в алфавитном порядке, либо по частоте встречаемости термов).

9. Сохранить частотный словарь.

10. Завершить работу алгоритма.

Таким образом, в генерируемый частотный словарь попадут только те слова, выражения и формы слов, которые представляют собой термины либо специализированную лексику конкретной предметной области. Это позволит сократить объем частотного словаря, а соответственно, и информационно-терминологического базиса, а также повысить их качество.

В данной работе на основе анализа способов построения частотных словарей авторами предложен модифицированный алгоритм автоматического формирования частотных словарей на основе исходных текстов с применением лемматизации. При использовании данного алгоритма можно обеспечить построение информационно-терминологического базиса

меньшего объема и с сопоставимым уровнем качества. Таким образом, удастся достичь снижения трудоемкости изучения ИТБ и упрощения процесса его формирования.

Библиографические ссылки

1. Огнерубов С. С. Формирование информационно-терминологического базиса в мультилингвистических системах обучения : дис. ... канд. техн. наук. Красноярск, 2008.

2. Ковалев И. В., Карасева М. В., Лесков В. О. Алгоритмизация процедур включения связанных лексем в структуру информационно-терминологического базиса // Программные продукты и системы. 2009. № 4. С. 28–32.

3. Атанов Г. А., Пустынникова И. Н. Структурирование понятий предметной области с помощью методов представления знаний // Искусственный интеллект. 1997. № 2. С. 29–40.

References

1. Ognrubov S. S. *Formirovaniye informatsionno-terminologicheskogo bazisa v multilingvisticheskikh sistemakh obucheniya*. Diss. kand. tekhn. Nauk [Formation of the informational and terminological basis in multilingual training systems. Thesis for candidate's degree]. Krasnoyarsk, 2008.

2. Kovalev I. V., karaseva M. V., Leskov V. O. *Programmnie produkti i sistemi*. 2009, no. 4, p. 28–32.

3. Atanov G. A., Pustinnikova I. N. *Iskusstvennii intellekt*. 1997 no 2, p. 29–40.

© Ковалев И. В., Середин А. И., Карасева М. В., Зеленков П. В., Храпунова В. В., 2013

УДК 62.501

О НЕПАРАМЕТРИЧЕСКОЙ ИДЕНТИФИКАЦИИ СИСТЕМЫ УПРАВЛЕНИЯ КОНВЕРТЕРНОЙ ПЛАВКОЙ

М. Е. Корнет

Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: maria4business@mail.ru

Рассматривается задача непараметрической идентификации дискретно-непрерывного процесса кислородно-конвертерной плавки стали. Отмечаются особенности идентификации в условиях параметрической и непараметрической неопределенности. Подчеркивается целесообразность использования непараметрических методов в условиях малой априорной информации. Построен ряд непараметрических моделей процесса кислородно-конвертерной плавки стали по важнейшим выходным переменным, определяющим химический состав металла на повалке: углероду (C), сере (S), фосфору (P) и марганцу (Mn). Приводятся некоторые результаты моделирования процесса кислородно-конвертерной плавки стали, которые показывают, что процесс ведется в рамках технологического регламента, однако этого не достаточно для качественного управления процессом плавки.

Ключевые слова: идентификация дискретно-непрерывных процессов, непараметрическая регрессия, непараметрическая идентификация кислородно-конвертерной плавки.