Лексическая основа словаря для системы квазиреферирования

Станкевич А.Ю.
Кафедра общего и славянского языкознания, филологический факультет
Гродненский государственный университет им. Янки Купалы
Гродно, Республика Беларусь
e-mail: a.stan.lab@gmail.com

***Аннотация* — Описана методика построения словаря строевых элементов научного текста, поддерживающего систему квазиреферирования. Дана характеристика лексической основы такого словаря.**

Ключевые слова: автоматическая обработка текста; квазиреферирование; словарь для АОТ

I. Введение

Квазилексемой называем объединенную общностью лексического и/или грамматического значения именованную коллекцию подстрок, наличие любой из которых на некотором фрагменте текста считаем входом квазилексемы в текст; вход квазилексемы возможен на участке текста, содержащим как одну, так и более одной словоформы. Множество именованных текстовых фрагментов называем телом квазилексемы, имя этого множества называем именем квазилексемы.

Словарем квазилексем называем программное отображение входной таблицы квазилексем, содержащейся в набранном пользователем текстовом файле. Словарь квазилексем содержит непустое неупорядоченное множество квазилексем и используется в реализованной на Python 2.7 программе обработки текста, где единицей словаря (объекта Dictionary) является кортеж формата (имя_квазилексемы, [список элементов тела квазилексемы]): например: (ИСПОЛЬЗОВАТЬ, [использу, использова(?:л|т|н), использован[^и]).

Предполагаем, что словарь строевых квазилексем научного текста (далее – словарь СК), единицы которого, во-первых, обслуживают подачу информации и, во-вторых, устойчиво присутствуют на массиве научных текстов безотносительно к предметной области, применим для поддержки автоматического реферирования научного текста.

Ниже описана методика построения такого словаря.

II. Эталонный текстовый массив

Текстовым массивом, на котором определялась основа словаря СК, стали группы REF_A и REF_B, состоящие каждая из 729 рефератов (источник: электронная реферативная БД ВИНИТИ рефератов «Медицина»; тематический охват: исследования в области экспериментальной и клинической медицины).

Группы REF_A И REF_B формировались так.

Шаг 1. Из электронной реферативной БД было загружено 1458 рефератов, соответствующих 113 149 словоупотреблениям. Выбор рефератов ВИНИТИ обусловлен основным назначением групп REF_A и REF_B, а именно: быть эталонным текстовым массивом, на котором, определяются, во-первых, словарь СК, и, во-вторых, типичные взаиморазмещения его единиц.

Шаг 2. Рефераты были пронумерованы; рефераты с четными номерами отобраны в группу REF_A, с нечетными – в группу REF_B. На этом шаге подтвердилось априорное предположение о равновесности реферативного продукта ВИНИТИ: количественные параметры текстов оказались весьма близки: REF_A: 57 776 словоупотреблений и 4 357 фразовых отрезков; REF_B: 55 369 словоупотреблений и 4 319 фразовых отрезков.

III. Основа словаря СК

Шаг 1. Рефераты групп REF_A и REF_B были лемматизированы (инструмент: парсер Mystem [1]). Число полученных лемм: 9 570 (REF_A) и 9 306 (REF_B) по-прежнему подтверждает исходное предположение о равновесности этих групп.

Шаг 2. Найдены пересечения множеств лемм групп REF_A и REF_B, из них автоматически исключены записи на латинице, леммы с метками «?» и «??» (сужение до 3 462 лемм).

Шаг 3. Исключены в ручном режиме предлоги, частицы, союзы, числительные и прочие элементы, с нашей точки зрения не несущие внутрисистемного значения (сужение до 1 506 лемм, принято за основу словаря СК).

Данные о покрытии основой словаря СК текстов разных типов (а именно: текстов рефератов трудов международной конференции по компьютерной лингвистике и ее приложениям «Диалог» за 2006–2010 гг. [2], далее – REF_D; текстов статей этой конференции, далее – Dialog; текстов заголовков статей этой конференции, далее – Head_D) приведены в таблицах 1–3.

Табл. 1.1. Покрытие основой словаря СК текстов REF_Dialog

Текст	*N*	*lemCov*	*textCov*
REF_D 2006	5 035	0.3542	0.4119
REF_D 2007	4 109	0.3794	0.4347
REF_D 2008	3 854	0.3704	0.4071
REF_D 2009	3 783	0.4035	0.4256
REF_D 2010	3 979	0.3509	0.3996

Табл. 1.2. Покрытие основой словаря СК текстов REF_Dialog (латиница исключена)

Текст	*N*	*lemCov*	*textCov*
REF_D 2006	4 999	0.3601	0.4149
REF_D 2007	4 079	0.3874	0.4379
REF_D 2008	3 817	0.38	0.4111
REF_D 2009	3 761	0.4101	0.4281
REF_D 2010	3 943	0.3585	0.4032

Табл. 2.1. Покрытие основой словаря СК текстов Dialog

Текст	*N*	*lemCov*	*textCov*
Dialog 2006	280 532	0.0666	0.2969
Dialog 2007	267 915	0.0774	0.3102
Dialog 2008	266 458	0.0675	0.2773
Dialog 2009	279 680	0.0649	0.2731
Dialog 2010	309 926	0.0596	0.2765

Табл. 2.2. Покрытие основой словаря СК текстов Dialog (латиница исключена)

Текст	*N*	*lemCov*	*textCov*
Dialog 2006	258 008	0.0911	0.3229
Dialog 2007	249 165	0.1014	0.3336
Dialog 2008	243 549	0.092	0.3034
Dialog 2009	246 763	0.0904	0.3095
Dialog 2010	272 412	0.0835	0.3146

Комментарий к таблицам

1. Покрытие текста (textCov) рассчитываем как отношение числа словоформ текста, покрываемых леммами основы словаря СК, к общему числу словоформ этого текста (N). Покрытие лемм (lemCov) рассчитываем как отношение числа лемм текста, покрываемых леммами основы , к общему числу лемм этого текста (N).

2. За длину текста (N) мы принимали накопленную частоту лемм, возвращенных парсером Mystem. Такое значение длины может несущественно отличаться от значения накопленной частоты, возвращенного программами создания частотных индексов. Разность частот объясняется спецификой работы парсера при обработке слов с дефисным написанием и некоторых других. К примеру, на тестовом файле, содержащем 2 226 примеров дефисного написания, извлеченных из текстов по медицине (1 879 примеров не содержат символов латиницы и цифр) парсером верно, т.е. без потери дефиса, лемматизированы 36: *сердечно-сосудистый (7), какой-либо (5), научно-исследовательский (3), масс-спектрометрия (3), физико-химический (3), темно-коричневый (2), научно-технический (2), по-разному (1), по-новому (1), юго-западный (1), агар-агар (1), какой-то (1), цифро-аналоговый (1), из-за (1), во-первых (1), во-вторых (1), аналого-цифровой (1), когда-либо (1)*.

Большая часть коллекции примеров содержит по понятной причине не распознаваемые морфологическим парсером единицы: аналоги сочинительных конструкций: *Ослера-Рендю-Вебера, октябре-ноябре*; сокращения (в т.ч. нетипичные): *к-рые, нек-рым, р-ра*; конструкции с аналитическими прилагательными: *Т-лимфоцитов-киллеров, сайт-адресованными, эритемо-маланина-метра*, меньшая часть коллекции достаточно тривиальна: сложные прилагательные: *молекулярно-генетических, научно-практической*, топонимы: *Рио-де-Жанейро, Санкт-Петербург* и т.п.

4. Значения lemCov близки на текстах рефератов и текстах заголовков. При этом значения lemCov заголовков и рефератов существенно выше значения lemCov полных текстов.

Табл. 3.1. Покрытие основой словаря СК текстов Head_D

Текст	*N*	*lemCov*	*textCov*
Head_D 2006	893	0.342	0.3561
Head_D 2007	805	0.3541	0.3329
Head_D 2008	823	0.3014	0.2868
Head_D 2009	827	0.3278	0.3337
Head_D 2010	808	0.331	0.3354

Табл. 3.2. Покрытие основой словаря СК текстов Head_D (латиница исключена)

Текст	*N*	*lemCov*	*textCov*
Head_D 2006	883	0.3489	0.3601
Head_D 2007	798	0.3601	0.3358
Head_D 2008	808	0.3127	0.2921
Head_D 2009	815	0.3374	0.3387
Head_D 2010	804	0.3341	0.3371

Иначе говоря, по значениям LemCov рефераты и заголовки противопоставляются полным текстам (содержательно и функционально заголовок ближе к реферату, чем к полному тексту). Это можно считать косвенным подтверждением того, что словарь СК пригоден для поддержки системы автоматического реферирования.

5. Значения TextCov близки на всех трех типах текстов, хотя, как и ожидалось, самым высоким стало среднее для текстов рефератов (0.423), самым низким - среднее для полных текстов (0.3168). Заметим, что средние длины для текстов каждого типа существенно различались.

Полученная основа словаря была отредактирована в полуавтоматическом режиме и приведена к формату словаря СК.

IV. Настройка словаря СК на предметную область

Как отмечается в [3], индикаторные словари (словарь СК можно отнести к этому типу), зависят «не только от жанра <…>, но и от специфики конкретной предметной области». Для снятия такой зависимости авторы процитированной работы предлагают «уже на этапе обучения формировать смешанную (из разных проблемных областей) подборку текстов».

В докладе будет представлена методика пополнения словаря СК, выполняемого с опорой на коэффициенты активности квазилексем, имеющих устойчивые взаиморазмещения на эталонных текстах.

[1] Mystem [Электронный ресурс].– Режим доступа: http://company.yandex.ru/technology/mystem/.

[2] Материалы международной конференции «Диалог» [Электронный ресурс].– Режим доступа: http://www.dialog-21.ru/dialog2010/materials.asp.– Дата доступа: 11.09.2011.

[3] Саломатина, Н.В. Уточнение и обогащение индикаторных словарей для автоматического извлечения информации из научных текстов / Н.В. Саломатина, В.Д. Гусев // Труды международной конференции «Диалог 2007» [Электронный ресурс].– Режим доступа: http://www.dialog-21.ru/dialog2007/materials/html/75.htm.– Дата доступа: 11.09.2011.