ЛИНГВИСТИЧЕСКАЯ СОСТАВЛЯЮЩАЯ В ЗАДАЧЕ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ ЗАИМСТВОВАННЫХ ФРАГМЕНТОВ

Ю. Б. Крапивин

Брестский государственный технический университет
Брест, Беларусь
e-mail: ybox@list.ru

Сформулированы требования к необходимому уровню автоматического лингвистического анализа текста с целью автоматического распознавания заимствованных фрагментов текстовых документов.

Ключевые слова: естественный язык; автоматическая обработка текстов; заимствованный фрагмент.

LINGUISTIC COMPONENT IN THE PROBLEM OF THE AUTOMATIC IDENTIFICATION OF THE ADOPTED FRAGMENTS

Y. B. Krapivin

Brest State Technical University
Brest, Belarus

The article presents the level requirements of the automatic linguistic text analysis to automatic identification of the adopted fragments of the text documents.

Keywords: natural language; automatic text processing; adopted fragment.

Постоянно увеличивающийся объем информации, представленной на различных языках как в полнотекстовых базах данных, так и в сети интернет, обостряет проблему ее оперативной и качественной обработки с целью удовлетворения информационной потребности пользователей. Под информационным поиском (ИП) обычно понимают непосредственно процесс поиска и представления пользователю информации в соответствии с запросом, который, в свою очередь, и отражает эту информационную потребность. Одним из основных типов ИП является поиск релевантных документов, т. е. тех, которые схожи по содержанию с заданным документом-образцом, который и выступает в качестве запроса пользователя. Существенным достоинством такого запроса является его большая информативность, что способствует эффективному решению задачи. Оно обычно основывается [1] на некоторой процедуре индексирования, в результате которого с использованием, как правило, определенной лингвистической обработки строится формальное представление (поисковый образ) и запроса (ПОЗ), и документов (ПОД) из поисковой базы, а также процедуры сравнения поисковых образов запроса и документов с определением согласно некоторому правилу степени их

соответствия (релевантности). На основе получаемых оценок принимается решение о выдаче или невыдаче того или иного документа. Наиболее распространенными к настоящему времени моделями поиска, на основе которых система ИП принимает такое решение, являются модели, основанные на классификаторах, теоретико-множественные, алгебраические и вероятностные модели [2, 3]. Существует достаточно много систем индексирования и поиска документов из различных информационных источников. Среди наиболее известных и широко используемых следует указать такие поисковые службы сети интернет, как Google [4], Яндекс [5], Bing [6] и др., а также такие корпоративные системы поиска, как mnoGoSearch [7], dtSearch [8] и др. Как показали проведенные исследования, эти системы ориентированы на статистический анализ лексических единиц текста, учет позиций слов в документах по отношению к запросам пользователей, морфологический анализ и синонимию лексических единиц, преобразование запроса в набор ключевых слов и использование логических операций. Таким образом, данные системы еще далеки от использования серьезного лингвистического анализа текста (т. е. решают вопрос релевантности документов фактически на уровне их «лексического подобия»). Многие из них работают с несколькими языками, т. е. являются многоязычными, но не включают cross-language функциональность. Их недостаточно высокая точность информационного поиска в ряде случаев вызвана ограничениями, накладываемыми на длину ПОЗ, которые в свою очередь являются следствием требования приемлемой скорости реакции системы.

Самое непосредственное отношение к задаче поиска документов, релевантных данному, имеет актуальная задача автоматического распознавания плагиата. Существует достаточно много его определений [9, 10], но чаще всего под ним понимают умышленное присвоение авторства чужого произведения в науке или искусстве, чужих идей или изобретений [11]. В данном контексте речь идет о распознавании плагиата применительно к информации, представленной в виде текстовых документов на естественном языке (ЕЯ). При этом надо иметь в виду, что плагиат может быть явным и неявным. В первом случае речь фактически идет об одном и том же фрагменте текста, принадлежащем разным текстовым документам (могут допускаться минимальные расхождения, например за счет использования вводных слов, синонимов и т. п.). Здесь для простоты под фрагментом понимается одно и более предложений вплоть до целого документа, т. е. речь идет об одной и той же цепочке символов алфавита ЕЯ, включая цифры, знаки препинания, пробелы и т. п. (полное совпадение). Во втором случае речь идет о фрагментах различных текстовых документов, имеющих одинаковый по отношению к заданной системе знаний, но выраженный разными цепочками символов смысл.

Таким образом, задача распознавания плагиата сводится прежде всего к поиску таких фрагментов данного (входного) текстового документа, которые релевантны (от полного совпадения до одинакового смысла в соответствии с заданной системой знаний) фрагментам других текстовых документов, представленных в некоторой заданной коллекции – в закрытой и относительно статичной, как, например, полнотекстовая база данных, или открытой и постоянно изменяющейся – web-документы в сети интернет. Далее о релевантных в указанном выше смысле фрагментах текстовых документов будем говорить соответственно как о лексически и семантически заимствованных фрагментах (ЗФ). Такой поиск, а это основная составляющая решения задачи, может быть осуществлен по-разному:

1) вручную экспертом или группой экспертов, что может быть эффективно, например, если поисковая база включает буквально несколько документов;

2) путем выполнения предварительного автоматического поиска наиболее релевантных текстовых документов в поисковой базе (современный уровень развития средств ИП позволяет решать эту задачу достаточно эффективно) и далее с ними работает, как отмечалось в 1), эксперт/группа экспертов;

3) в случаях 1) и 2) поиск ЗФ (в случае 2) на втором этапе решения) осуществляется автоматически; учитывая момент автоматизации, в этом случае есть возможность достижения более высоких показателей полноты решения задачи. Безусловно, третья схема является наиболее перспективной, но она основывается на качественном решении задачи автоматического лингвистического анализа текста. Это особенно важно при переходе к проблеме автоматического распознавания семантически заимствованных фрагментов текстовых документов.

Понятно, что окончательное (не)признание плагиата требует в дополнение последующего экспертного анализа – в первом случае с целью установления факта цитирования (это задача по причине несоблюдения его правил является очень трудоемкой для автоматизации решения), во втором – еще и анализа тождественности смысла в контексте излагаемых фактов, идей и т. п., которое, как правило, лежит за рамками анализа, определяемого заданной системой знаний.

Проведенный анализ [12] показал, что существующие решения задачи автоматического распознавания ЗФ фактически ориентированы на распознавание лексически заимствованных фрагментов с учетом, в лучшем случае, простейших морфологических преобразований и отношений синонимии, не используют развитого лингвистического анализа текстовых документов и, следовательно, не ориентированы на решение задачи с учетом более сложных преобразований текста и автоматического распознавания семантически заимствованных фрагментов.

Задача автоматического распознавания ЗФ в целом включает в совокупности следующие задачи: 1) поиска релевантных входному текстовых документов в полнотекстовой БД и в сети интернет; 2) автоматического распознавания языка текстового документа; 3) машинного перевода текстовых документов и их поисковых образов во множестве заданных ЕЯ; 4) автоматического распознавания лексически и семантически заимствованных фрагментов текстовых документов; 5) автоматического построения отчета; 6) автоматического лингвистического анализа текстовых документов.

С учетом приведенного перечня основных, подлежащих решению задач в рамках общей задачи автоматического распознавания ЗФ, их анализа, а также основываясь на обоснованном тезисе об усилении лингвистической составляющей при построении решений этих задач, могут быть сформулированы требования к необходимому уровню автоматического лингвистического анализа текста (задача 6). Безусловно, что в этом плане наиболее серьезные требования выдвигает задача автоматического распознавания семантически ЗФ, поскольку здесь речь идет об уровне семантического анализа текста. А для этого, очевидно, необходима вся функциональность так называемого базового ЛП [13]: форматирование текста, лексический анализ, лексико-грамматический анализ, синтаксический анализ, семантико-синтаксический анализ. Эта функциональность, в свою очередь, обеспечивается соответствующей лингвистической базой знаний (ЛБЗ).

Основными компонентами ЛБЗ являются следующие:

– классификаторы лексико-грамматических, синтаксических и семантических свойств ЕЯ; их состав зависит от конкретных свойств ЕЯ и от характера приложения, определяющего степень детализации лингвистического анализа текста;

– базовый (эталонный) словарь; он реализуется в виде словаря словоформ ЕЯ и включает максимально возможное их количество; при этом для каждой словоформы указаны все ее возможные вне контекста лексико-грамматические классы (ЛГК); классические базовые словари (БС), например русского и белорусского языков – более одного миллиона словоформ;

– базовый (эталонный) корпус текстов (БКТ); реализуется в виде определенным образом подобранных текстов, причем как минимум каждому слову текста указан его единственный с точки зрения контекста ЛГК; БКТ предназначен прежде всего для получения количественных оценок языка, тестирования лингвистических гипотез и отдельных алгоритмов и систем автоматической обработки текста;

– лингвистические правила анализа (ЛПР) текста на различных уровнях глубины ЕЯ; такие правила, получаемые лингвистами-экспертами, являются основой разработки машинных алгоритмов для большинства этапов автоматического лингвистического анализа текста; совокупность этих правил, например, для лексико-грамматического и синтаксического анализа, составляет грамматику ЕЯ.

С целью машинной обработки ЛП должна быть разработана некоторая нотация для формального описания этих правил, в которой они обычно и представляются в ЛБЗ. Причем предлагаемый формализм должен быть максимально соотнесен с требованиями его доступности для использования экспертами, возможностью обобщения разрабатываемых правил и оптимизации скорости их обработки. В [13] в качестве такого формализма разработан и успешно внедрен в промышленные приложения так называемый язык расширенных регулярных выражений (WRE). В дополнение к основным перечисленным ресурсам в состав ЛБЗ входят различного рода словари (словари идиом, аббревиатур, имен собственных, слов, параметров и т. п.), специальные лексические базы данных, например типа WordNet [14], отображающие синонимические, иерархические и ассоциативные отношения концептов и т. д. Безусловно, ЛБЗ, ориентируясь на обработку текстов на нескольких языках, является многоязычной.

Таким образом, функциональности представленного выше БЛП, очевидно, тем более достаточно для обеспечения лингвистической составляющей при решении задачи (1) – для автоматического распознавания ключевых слов с целью построения для текстовых документов их ПОДов, задачи (2) – для автоматического построения поискового образа языка, например в виде словаря грамматических слов и т. д. Этот же БЛП обеспечивает и требуемый для решения всех задач уровень лингвостатистического анализа текста, который традиционно сводится прежде всего к подсчету частот, распознаваемых на приведенных выше этапах обработки текста лексических единиц и отношений.

ЗАКЛЮЧЕНИЕ

Применение методов анализа текстовой информации, опирающихся на знания о ЕЯ, обеспечит качественное решение задачи автоматического распознавания ЗФ, в том числе в части обнаружения лексически и семантически заимствованных фрагментов.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. 1 ed. Cambridge University Press, 2008.

2. Сегалович И. В. Как работают поисковые системы // Мир интернета. 2002. № 10. С. 24–32.
3. Цукерт А. Г. Проблемы и перспективы информационного поиска // Изв. Таганрог. гос. радиотехн. ун-та. 2001. Т. 21, № 3. С. 194–201.
4. Google [Электронный ресурс]. 2016. URL: https://www.google.com/ (дата доступа: 22.05.2016).
5. Яндекс [Электронный ресурс]. 2016. URL: https://www.yandex.ru/ (дата доступа: 22.05.2016).
6. Bing [Электронный ресурс]. 2016. URL: https://www.bing.com/ (дата доступа: 22.05.2016).
7. mnoGoSearch [Электронный ресурс]. 2016. URL: https://www.mnogosearch.org/ (дата доступа: 22.05.2016).
8. dtSearch [Электронный ресурс]. 2016. URL: https://www.dtsearch.com/ (дата доступа: 22.05.2016).
9. TEXNOLOGIA [Электронный ресурс]. 2016. URL: http://www.textologia.ru/slovari/-literaturovedcheskie-terminy/plagiat/?q=458&n=163 (дата доступа: 19.05.2016).
10. Большой энциклопедический словарь [Электронный ресурс]. 2016. URL: http://mirslovarei.com/bes_a/ (дата доступа: 22.05.2016).
11. Plagiarism [Electronic resource]. 2016. URL: https://en.wikipedia.org/wiki/Plagiarism (date of access: 22.05.2016).
12. Крапивин Ю. Б. К задаче автоматического распознавания воспроизведенных фрагментов текстовых документов // Вестн. БрГТУ. Сер.: Физика, математика, информатика. 2009. № 5(59). С. 120–123.
13. Чеусов А. В. Разработка алгоритмов и технологии построения многоязычного базового лингвистического процессора : дис. … канд. техн. наук. Минск, 2013.
14. WordNet [Electronic resource]. 2016. URL: http://wordnet.princeton.edu/ (date of access: 17.05.2016).