519.769

МЕТОД ТЕМАТИЧЕСКОЙ СЕГМЕНТАЦИИ ТЕКСТОВ НА ОСНОВЕ ГРАФА ЗНАНИЙ

А.Ф. Шарафиев
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: whiskeydudev@gmail.com

М.С. Гаврилов
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: cobraj@yandex.ru

Ключевые слова: тематическая сегментация, граф знаний, NLP.

Аннотация: Тематическая сегментация — это задача разделения неструктурированного текста на тематически связные сегменты (такие, в которых речь идет об одном и том же). Граф знаний — графовая структура, вершинами которой являются различные объекты, а ребрами — отношения между ними. Обе задачи не являются новыми, потому существует множество алгоритмов для их решения. Однако, методы решения задачи тематической сегментации с использованием графов знаний до сих пор исследованы мало. Более того, пока еще нельзя сказать, что задача тематической сегментации решена в общем виде, то есть существуют алгоритмы, способные при должной настройке, решить задачу с требуемым качеством на конкретном наборе данных. В данной работе предлагается новый метод решения задачи тематической сегментации на основе графов знаний. Применение графов знаний при сегментации позволяет использовать больше информации о словах в тексте: помимо того чтобы основываться на co-occurrance и семантических расстояниях (как классические алгоритмы), методы на основе графов знаний могут применять расстояние между словами на графе, инкорпорируя тем самым фактологическую информацию из графа знаний в процесс принятия решений о биении текста на сегменты. В данной работе предлагается метод решения задачи тематической сегментации на основе графов знаний.

1. Введение

Основная цель данной работы заключается в разработке алгоритма, который полностью неструктурированный текст, представляющий собой набор предложений, преобразует в текст, разделенный на абзацы, таким образом, чтобы каждый из них был тематически связным.

Актуальность данного направления обусловлена с одной стороны тем, что человеку в век больших объемов данных и всеобщей компьютеризации нередко приходится сталкиваться с такими текстами (например, потому что соответствующая структура документа была потеряна в следствии сбоев в системе хранения документов), а работать со структурированной информацией намного проще, чем с неструктурированной. С другой стороны, использование текстов (даже не очень больших размеров), разделенных на тематические сегменты, потенциально может повысить качество результатов методов, решающих другие задачи NLP, такие как поиск схожих документов, поиск требуемой информации в документах, вопросно-

ответные системы и т. д., за счет предоставления им дополнительной мета-информации о внутренних связях внутри текстов.

Актуальность данной работы следует из отсутствия на данный момент простых, доступных и быстрых алгоритмов, способных решить данную задачу с требуемым качеством, а также низкой исследованности методов решения задачи сегментации с использованием графов знаний.

2. Основная часть

2.1. Связанные работы

В [1] предлагается алгоритм сегментации, основанный на сравнении лексического состава смежных единиц документа. Сначала документ предобрабатывается (токенизация, лемматизация, удаление стоп-слов) и разбивается на блоки фиксированной длины. Затем для каждой пары соседних блоков подсчитывается оценка тематического сдвига. На основе сравнения значений полученных оценок принимается решение о выставлении границ сегментов теми или иными соседними блоками.

В работе [2] предлагается метод сегментации видео-лекций на основе графов знаний. Сначала каждому слайду лекции сопоставляется подграф общего графа знаний. После ряда предобработок над подграфами производится сравнение смежных подграфов, на основе чего формируется оценка их схожести и принимается решение об объединении соседних графов в один большой, соответствующий тематическому сегменту.

2.2. Используемые данные

В работе задействуется датасет, представляющий собой 18000 статей ПРНД. Для непосредственного использования из этих статей были автоматически отфильтрованы и предобработаны 6000 русскоязычных статей.

2.3. Описание метода

В данной работе предлагается алгоритм, основанный на работах [1, 2]. Предлагаемый метод представляет собой адаптацию алгоритма сегментации на основе графов знаний из [2] для сегментирования произвольных научных статей посредством механизмов из работы [1]. Показываются недостатки прямого подхода к адаптации, а также способы их устранения. Предлагаются способы построения сопутствующих графов знаний, а также влияние выбора этого способа на итоговый результат сегментации.

В общих чертах предлагаемый алгоритм сегментации выглядит следующим образом. Считается, что текст на входе представляет собой набор предложений. Каждое предложение отображается в подграф заранее построенного графа знаний. Сравнивая соседние подграфы и считая оценку прироста тематической связности при объединении их в один подграф, принимается решение о соответствующем объединении. В рамках работы рассматриваются несколько способов задания функционала для оценки прироста тематической связности. На рисунке 1 изображена схема работы предлагаемого алгоритма сегментации.

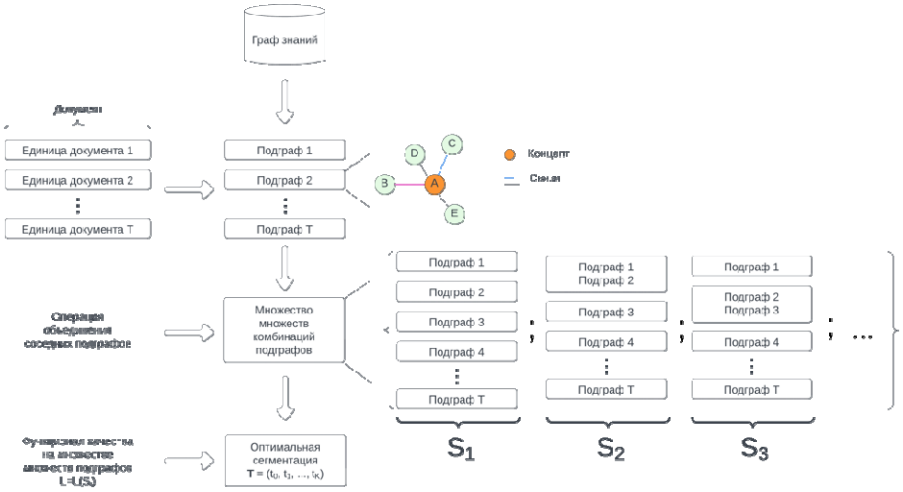

Рис. 1. Схема предлагаемого метода.

2.4. Результаты экспериментов

Были проведены эксперименты по валидации предлагаемого метода. Для непосредственной оценки качества предлагаемого алгоритма часть датасета была вручную размечена под задачу тематической сегментации. Было проведено сравнение результатов работы алгоритма с эталонной разметкой с помощью метрик Pk и WinDiff. Были получены соответствующие значения: Pk = 0.476, WinDiff = 0.524

Также были проведены эксперименты по применению метода отображения текста в подграф к задаче поиска тематически схожих текстов. Для этого, из вручную размеченного датасета выбирались случайным образом пары тематически схожих и тематически различных текстов в равном соотношении. Каждый текст из пары отображался в подграф и для пары подграфов считалась оценка прироста тематической связности. Оценка f1 для такого метода составила 0.69 в среднем по 5 различным случайным выборкам из 80 текстов, что сопоставимо с оценками классических алгоритмов.

3. Заключение

Представленный в данной работе метод тематической сегментации текстов показал высокое оценки при тестировании и может выступать в качестве доступного алгоритма решения задачи тематической сегментации научных текстов. Сам метод оценки тематической схожести фрагментов текста при помощи отображения в подграф был также протестирован на задаче определения похожих текстов.

Список литературы

1. Hearst A.M. TextTiling: segmenting text into multi-paragraph subtopic passages // Computational Linguistics. 1997. Vol. 23, No. 1. P. 33-64.
2. Das A., Das P.P. Incorporating Domain Knowledge To Improve Topic Segmentation Of Long MOOC Lecture Videos // arXiv:2012.07589. 2020.