DOI: 10.25728/mlsd.2022.1237

СБОР, НАКОПЛЕНИЕ, ПЕРЕВОД С ЯЗЫКА НА ЯЗЫК И АНАЛИЗ БОЛЬШИХ КОРПУСОВ ТЕКСТОВ – ЛЕГИТИМНЫЙ ПОИСК ЗАКРЫТЫХ ДАННЫХ В ОТКРЫТЫХ ИСТОЧНИКАХ: ВЗГЛЯД ЗА ЯЗЫКОВОЙ ЗАНАВЕС

Харламов А.А.
Институт высшей нервной деятельности и нейрофизиологии РАН,
Россия, г. Москва ул. Бутлерова д.5А
kharlamov@analyst.ru

Аннотация: В работе представлен алгоритм формирования и анализа корпуса текстов на незнакомом эксперту языке с целью получения легитимной информации из открытых источников, использующий имеющиеся в настоящий момент технологии автоматического анализа неструктурированных текстов. Анализ текстов осуществляется на языке поиска, а процедуры манипулирования с текстовой информацией осуществляются на языке пользователя.

Ключевые слова: мультиязычный анализ текстов, автоматический смысловой анализ неструктурированных текстов, смысловой портрет текста, ключевые слова, реферат, гипертекст.

Введение

В настоящий момент бесконечно вырос объем публикуемых в социуме текстов, что привело, с одной стороны, к необходимости их эффективного сбора и анализа, а, с другой, – к возможности получения практически любой информации из этого источника. Однако, языковой барьер часто ставит непреодолимые препятствия для получения этой информации.

Для решения проблемы легитимного поиска нужной информации за языковым барьером необходимо решение нескольких разнородных задач, включая задачу сбора и накопления текстовой информации, разнообразного анализа собранной информации, в том числе – перевода текстов с языка на язык, и наконец, задачу визуализации информации в удобной для пользователя форме.

В настоящий момент сформировано большое число технологий для решения этих задач. Это технологии вторичного поиска, семантического анализа, автоматического перевода, визуализации результатов анализа как гипертекстовой структуры. Объединение этих технологий особенно актуально в случаях, когда пользователь не знает языка источника, то есть он не может организовать полноценный поиск. Для решения этой задачи используются имеющиеся в настоящий момент технологии, которые объединяются несколько необычным способом.

Большую роль в выборе подходов, алгоритмов и конкретных продуктов для анализа текстов играет степень доверия к результатам анализа. В этом случае использование искусственных нейронных сетей уступает традиционным алгоритмам, основанным на правилах, в силу отсутствия возможности заглянуть внутрь большей части нейросетевых структур. Степень доверия к результатам анализа формируется в процессе их использования. Объединение двух подходов – нейросетевого и основанного на правилах – дает удобный способ оценить корректность анализа.

Стандартный подход к анализу текстовой информации включает сбор текстовых материалов, анализ и визуализацию результатов анализа. В случае, если языки текстов и представления различаются, в процесс включается перевод текстов на язык анализа (см. Рис. 1).

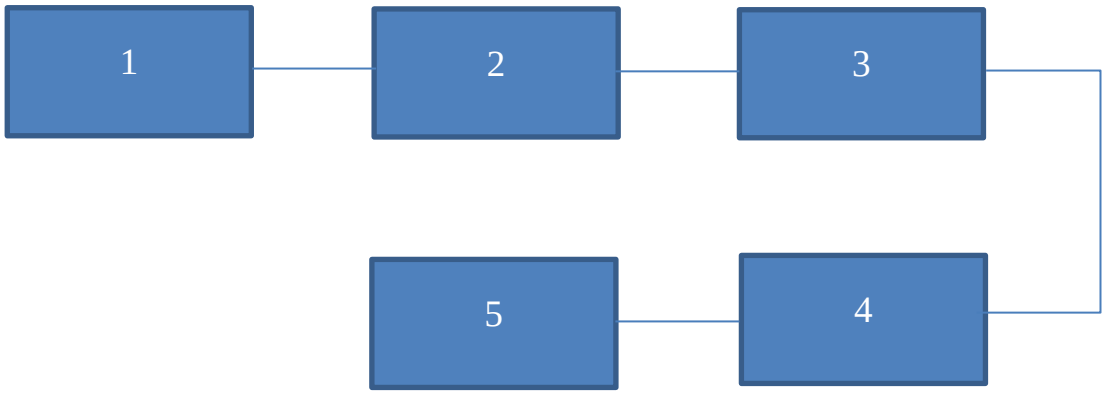

Рис. 1. Этапы сбора и обработки текстовой информации. (1) Запрос на поиск. (2) Сбор релевантной информации. (3) Перевод текстовой информации на язык анализа. (4) Анализ корпуса текстов. (5) Визуализация результатов анализа

Здесь возникают две сложности, и обе они касаются объемов обрабатываемой информации: (1) необходимо переводить весь текстовый материал, предназначенный для анализа; и (2) качество перевода должно быть хорошим, иначе результаты анализа могут сильно ухудшиться настолько, насколько снижено качество перевода.

Использование автоматических переводчиков не снимает этих сложностей, так как качество автоматического перевода оставляет пока желать лучшего: понять можно, но не более того.

1 Сбор, накопление, перевод с языка на язык и анализ больших корпусов текстов – поиск данных в открытых источниках

Избежать ухудшения результатов анализа можно поменяв два квадратика на схеме (Рис. 1): переводить не анализируемые тексты, а переводить результаты анализа (см. Рис. 2). Тогда качество анализируемых текстов остается высоким, что исключает влияние качества перевода на результаты анализа. И переводить уже результаты анализа, объем которых значительно меньше (по крайней мере, в случае реферирования текста – в 3-10 раз), и вместо трудностей перевода связного текста в этом случае необходимо переводить отдельные предложения (в случае реферирования с использованием подхода на основе выдержек).

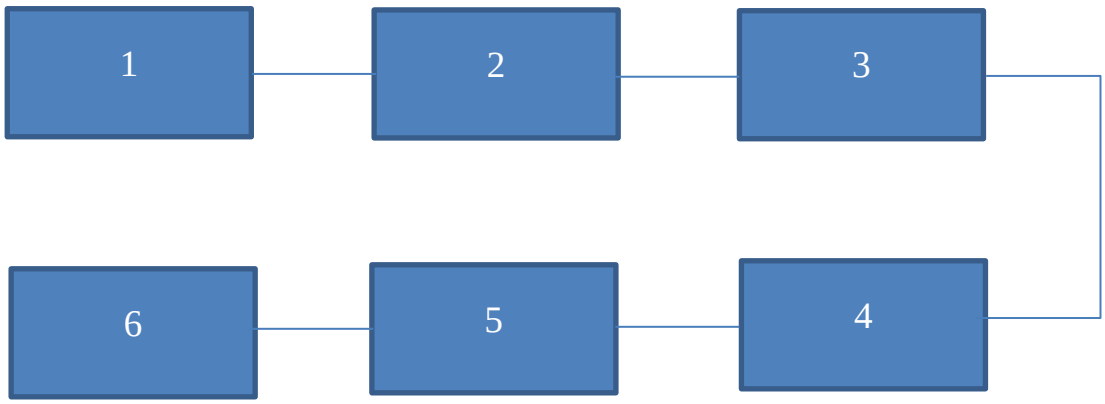

Рис. 2. Этапы сбора и обработки текстовой информации. (1) Запрос на поиск на языке пользователя. (2) Перевод запроса на поиск на язык поиска (качество перевода должно быть очень высоким). (3) Сбор релевантной информации на языке поиска. (4) Анализ корпуса текстов на языке поиска. (5) Перевод результатов анализа текстовой информации на язык пользователя. (6) Визуализация результатов анализа

Однако в этом случае особые требования по качеству предъявляются к анализу текстов на языке поиска.

Таким образом (см. Рис. 2), алгоритм легитимного поиска информации в открытых источниках включает в свой состав следующие этапы: (1) запрос на поиск на языке эксперта; (2) перевод запроса на поиск с языка эксперта на язык поиска; (3) поиск необходимой информации на языке поиска в соответствии с запросом; (4) анализ полученного корпуса текстов на языке поиска, в том числе, формирование семантического портрета корпуса текстов – семантической сети, выявление ключевых слов, реферирование текстов корпуса; (5) перевод результатов анализа с языка поиска на язык эксперта; и наконец (6) визуализация результатов анализа в виде гипертекстовой структуры, включающей в свой состав исходный текст, его семантическую сеть, ключевые слова (тематическую структуру), рефераты текстов, на языке эксперта.

1.1 Перевод запроса на поиск

Для сбора нужной текстовой информации на языке поиска необходимо сформировать запрос на поиск на языке поиска. Сейчас эта задача решается достаточно просто с использованием стандартного переводчика, например редактора Word. Он дает перевод достаточного качества. Однако, поскольку качество текста запроса влияет на результаты поиска, необходимо отнестись к этому этапу очень ответственно: вплоть до многократного перевода с языка на язык, чтобы удостовериться, что перевод осуществлен верно. Качество последующего перевода результатов анализа текстов, полученных в результате поиска, с языка поиска на язык эксперта не столь важно, так как смысл перевода восстанавливается в местах некорректного перевода за счет контекста.

1.2 Сбор и накопление

Наличие запроса на поиск на языке поиска является исходной точкой для поиска нужных текстов. Однако результат последующей автоматической обработки текстов на естественном языке крайне чувствителен к качеству исходных материалов. Подготовка текстовых материалов осложняется: (1) ограничением бесплатного доступа к материалам; (2) обширным тематическим охватом; (3) отсутствием в публичном доступе интерфейсов машинно-машинного взаимодействия для автоматизированной загрузки материалов по заданной теме в заданный диапазон времени на рабочую станцию исследователя; (4) большим количеством рекламных материалов и прочим информационным мусором, разжижающим содержательную часть материала.

Поэтому для эффективного решения задачи сбора и накопления текстовой информации необходимо использовать вторичные поисковики, снабженные механизмами маскировки целей пользователя. Подобное решение, например, реализовано фирмой NeoAge AI. Это мультиагентная система сбора и обработки электронных документов, поддерживающая свыше 3000 форматов и 200 естественных языков.

1.3 Анализ

Тем не менее, именно процедура анализа является самой важной частью процесса, так как точность, детальность и достоверность результатов гарантируют успех получения нужной информации.

Точность получаемых результатов определяется выбором подхода к выявлению значимой в предметной области информации, которых не так много. Дистрибутивная семантика дает результаты, хорошо интерпретируемые на текстах небольшого объема. Сетевая семантика, напротив, начинает работать на достаточно больших объемах обрабатываемой текстовой информации [1].

Детальность представления результатов зависит от постановки задачи: в одних случаях требуется взгляд сверху, в других – необходимо посмотреть на текст в лупу. Поэтому необходимо иметь возможность реализации анализа больших массивов текстовой информации на разных уровнях степени подробности при наличии механизма простого перехода в детальности представления с уровня на уровень. Эти уровни – это ключевые слова [2], реферат исходного текста [3], семантическая сеть [4] анализируемого текста – в рамках гипертекстовой структуры, включающей семантическую сеть текста, отдельные предложения текста, весь исходный текст.

Семантическая сеть – граф, вершинами которого являются ключевые понятия, а дугами – отношения этих понятий в тексте – это удобный способ представления содержания текста, так как в сознании человека знания представляются именно в виде семантических сетей, и поскольку и вершины, и связи взвешены их численными характеристиками – смысловым весом понятий и весом связей – ими удобно манипулировать в составе сети, например, выявлять наиболее важные предложения текста.

Анализ текстовой информации, основанный на сетевом подходе, предполагает выявление словаря слов-концептов, выявление их связности в тексте и оценку важности этих слов в рамках всего текста.

Словарь слов-концептов выявляется в тексте отображением F_n текстовой информации в многомерное пространство (см. (1)), порождающей в нем траекторию, самопересечения которой и формируют упомянутый словарь *{B}* (см. (2)) [5].

$$F_n : A \to \hat{A},\ F_n(A) = \hat{A}. \tag{1}$$

Здесь $A = (..., a_i, ... : a_i \in \{0,1\})$, а

$$\hat{A} = (..., \hat{a}_{-2}, \hat{a}_{-1}, ...\hat{a}_i, ...) = (..., (a_{-n-1}, a_{-n}, ..., a_{-2}), (a_{-n}, a_{-n+1}, ..., a_{-1}), ..., (a_{i-n+1}, a_{i-n+2}, ..., a_i ...), ...) .$$

$$\{\hat{B}_i\}_{k_1} = H_h RMF_n(\{A\}_{k_1}) . \tag{2}$$

Здесь $\{\hat{B}\}_{k_1}$ – словарь *k*-го уровня, *M* – функция записи, *R*- функция считывания, а H_h – порог по яркости траектории в многомерном пространстве.

Слова словаря формируют семантическую сеть, в которой учитываются связи слов в тексте (характеризуемые попарной встречаемостью z_{ij} в (3)).

Отдельно должны выставляться параметры по степени важности анализируемой информации: от ничего не значащих концептов до базовых концептов предметной области (учет смыслового веса). Это существенно, так как достоверность полученных результатов определяется степенью важности отдельных элементов текста, которые выявляются в результате анализа: чем выше семантический вес

анализируемых элементов, тем более вероятно их влияние на смысл целого текста.

Вычисление степени важности концептов этого словаря в анализируемом тексте осуществляется итеративной процедурой учета влияния предшествующих слов в цепочках на семантической сети.

$$w_i(t+1) = (\sum_{\substack{i \\ i \neq j}} w_i(t) w_{ij}) \sigma(\overline{E}) . \quad (3)$$

Здесь $w_i(0) = z_i$ – частота встречаемости слов в тексте, $w_{ij} = z_{ij} / z_j$, где z_{ij} – частота попарной встречаемости слов в предложениях текста, а $\sigma(\overline{E}) = 1/(1+e^{-k\overline{E}})$ – функция, нормирующая на среднее значение энергии всех вершин сети $\overline{E}$.

При этом ключевые слова можно извлекать разными способами, например, с использованием методов так называемого тематического моделирования. С рефератом сложнее, но тоже есть разные подходы, в основном основанные на извлечении выдержек. Наконец, комфортным способом навигации по тексту является гипертекстовая структура, в основе которой лежит семантическая сеть текста, которая отсылает пользователя к предложениям текста, содержащим ключевые понятия, а из них – непосредственно в текст. Гипертекстовая структура является удобным цитатником, который позволяет быстро находить нужные понятия навигацией по семантической сети в их контексте отдельных предложений или целого текста.

1.4 Визуализация

Гипертекстовая структура оказывается также удобным инструментом для визуализации результатов анализа. Визуализация результатов анализа важна с точки зрения возможности охвата содержания текста одним взглядом, с одной стороны, а с другой, возможности исследовать отдельные срезы текста по степени их смысловой важности. Поэтому удобно рассматривать одновременно все уровни представления текста на одном экране в разных окнах с возможностью перехода от одного окна к другому с сохранением смысловой связности этих переходов.

1.5 Разноязычный корпус текстов

Некоторые затруднения (иногда очень существенные) в анализ вносит разноязычность анализируемых текстов, а также необходимость использования результатов анализа на языке пользователя – эксперта. Поэтому вернемся к технологиям перевода с языка на язык еще раз. В настоящий момент автоматический перевод с широко распространенных языков не вызывает большой сложности. Специалист с легкостью читает автоматически переведенный программой Word текст, лишь иногда справляясь с так же автоматическим словарем Мультитран, например.

2 Алгоритм

Учитывая все вышеперечисленное, естественным является следующий алгоритм получения требуемой информации из открытых источников. (1) Запрос на поиск задается на языке пользователя (русском – для русско-язычного эксперта) и автоматически переводится на язык анализа (вордовским переводчиком на китайский, например). (2) Осуществляется поиск нужной информации на языке поиска (например, вторичным поисковиком NeoAge на китайском языке в китайско-язычной части Интеренета). (3) Осуществляется анализ собранной текстовой информации на языке поиска (например, с помощью технологии TextAnalyst – на китайском языке) с гипертекстовым представлением, включающим сформированную на полученном корпусе текстов семантическую сеть, выявленные ключевые слова, реферат, которые с помощью того же переводчика переводятся на язык пользователя (русский, в данном случае, так как объем предназначенного для визуализации, а следовательно, и для перевода – материала несоизмеримо меньше объема исходного корпуса анализируемых текстов). (4) Визуализированная посредством гипертекстового представления информация используется пользователем как экспертом, или используется для дальнейшего анализа с помощью других технологий (например, нечеткой кластеризации).

3 Пример анализа

Чтобы продемонстрировать вышеприведенные теоретические положения, попробуем получить информацию о положении в конкретной предметной области – в предметной области достаточно хорошо в настоящий момент исследованной – в предметной области автоматического смыслового анализа текстов. Промоделируем ситуацию анализа текста, представленного не на языке пользователя. В данном случае – это английский язык. В качестве инструмента используем

технологию для автоматического смыслового анализа текстов TextAnalyst [5]. Попробуем реализовать описанный выше алгоритм. В качестве исходного языка можно выбрать также немецкий и китайский – языки, которые обслуживаются упомянутой технологией. Английский язык выбран потому, что он знаком русскоязычным читателям лучше, чем немецкий и китайский.

3.1 Перевод запроса на поиск

Первым этапом этого алгоритма является запрос на поиск. Перевод запроса на язык поиска осуществляется с помощью стандартного переводчика программы Word. Переведем запрос «Автоматический смысловой анализ текстов». Перевод звучит так: «Automatic semantic text analysis».

3.2 Сбор и накопление

Запрос на поиск текстов в сети Интернет позволяет собрать корпус текстов на английском языке по указанной теме. Для последующего анализа выберем один текст (чтобы сэкономить время), и далее будем его анализировать. Это текст «Semantic Representations of Words and Automatic Keywords Extraction for Sentiment Analysis of Tourism Reviews», что означает «Что означает », взятый по адресу «restmex_paper5.pdf (ceur-ws.org)».

На самом деле сбор необходимой информации предваряется предварительным анализом (фильтрацией) текстов, отбираемых для формирования корпуса, на соответствие заданной теме запроса. И эта фильтрация осуществляется с помощью упомянутой технологии для автоматического смыслового анализа слабоструктурированной текстовой информации TextAnalyst путем сравнения семантической сети входного текста с семантической сетью корпуса текстов предметной области «Автоматический смысловой анализ текстов».

3.3 Анализ

Анализируем текст с помощью той же технологии TextAnalyst, функциональность которой включает автоматическое формирование семантической сети как смыслового портрета текста, выявление ключевых слов, формирование реферата текста, формирование гипертекстовой структуры текста в составе текста, семантической сети и предложений текста, относящихся к отдельным концептам семантической сети.

3.4 Обратный перевод результатов анализа

Обратный перевод результатов анализа выбранного текста с языка поиска на русский также может включать только те разделы (те результаты анализа), которые интересны пользователю. Ниже представлен только реферат с целью экономии места.

3.5 Визуализация

Для визуализации можно использовать многие технологии, но и в этом случае визуализация гиперетекстового представления текста оказывается наиболее удобной, так как позволяет менять детальность представления: «семантическая сеть (тематическая структура) текста-реферат (тематический реферат, фрагменты текста)-исходный текст».

3.6 Реферирование исходного текста

Для демонстрации работы алгоритма в качестве примера приведен результат реферирования (фрагмент реферата с целью экономии места) выбранного для анализа текста. В таком же представлении можно представлять результаты выявления ключевых слов, семантической сети, тематического реферата.

«In this **paper**, we describe the methods used to submit our results to the **Rest- Mex Sentiment Analysis task** of the **Iberian Languages Evaluation Forum** 2021.
For this competition edition, the **sentiment analysis** problem is defined as follows: "Given an opinion about a Mexican **tourist** place, the goal is to determine the **polarity**, between 1 and 5, of the **text**.
"The **sentiment analysis** sub-task is a **classification task** where the participating system has to **predict** the **polarity** of an opinion issued by a **tourist** who traveled to the most **representative** places of Guanajuato, Mexico.
Sentiment analysis task in **tourist texts** has gained relevance in the last decade.»

Автоматический перевод этого реферата с помощью встроенного переводчика программы Word имеет следующий вид.

«В этой статье мы описываем методы, используемые для представления наших результатов в задачу Rest-Mex Sentiment Analysis форума по оценке иберийских языков 2021 года.
Для этого конкурсного издания задача анализа настроений определяется следующим образом:
«Учитывая мнение о мексиканском туристическом месте, цель состоит в том, чтобы определить полярность, между 1 и 5, текста.
«Подзадача анализа настроений — это классификационная задача, в которой участвующая система должна предсказать полярность мнения, выданного туристом, который путешествовал по наиболее представительным местам Гуанахуато, Мексика.
Задача анализа настроений в туристических текстах приобрела актуальность в последнее десятилетие.»

Заключение

В работе представлен алгоритм применения имеющихся технологий автоматического анализа неструктурированной текстовой информации для их использования при поиске в иноязычной среде Интернет нужной информации при отсутствии у эксперта переводческих навыков в незнакомом ему языке. В результате использования этого алгоритма у эксперта, незнакомого с языком поиска, появляется возможность создания корпусов текстов на заданную тему с последующим анализом этих корпусов, включающим выявление ключевых слов, формирование смыслового портрета корпуса текстов в виде его семантической сети, реферирование текстов корпуса. И все это с учетом возможности изменения детальности анализа как по степени важности отдельных фрагментов текстов, так и по масштабу анализа – от ключевых слов до целого текста.

Литература

1. *Kharlamov Alexander, Gordeev Denis and Pantiukhin Dmitry* Distributional and Network Semantics. Text Analysis Approaches. //Neuroinformatics and Semantic Representations. Theory and Applications. Collective Monography. Chapter Four. – Newcastle upon Tyne: Cambridge Scholars Publishing. 2020. Pp. 83-139.
2. *Сухарева А.В., Воронцов К.В.* Построение полного набора тем вероятностных тематических моделей. //Интеллектуальные системы. Теория и приложения, том 23, выпуск 4. 2019. – С. 7–23.
3. *Осминин П.Г.* Современные подходы к автоматическому реферированию и аннотированию. //Вестник Южно-Уральского государственного университета. Серия: Лингвистика. № 25. 2012. – С. 134-135.
4. *Голенков В.В. с соавт.* Семантическая модель представления и обработки баз знаний. //Аналитика и управление данными в областях с интенсивным использованием данных: сборник научных трудов XIX Междунар. конф. DAMDID / RCDL’2017. Под ред. Л.А. Калиниченко, Я. Манолопулос, Н.А. Скворцова, В.А. Сухомлина. – М.: ФИЦ ИУ РАН, 2017. – С. 412 – 419.
5. *Харламов А.А.* Ассоциативная память — среда для формирования пространства знаний. От биологии к приложениям. – Дюссельдорф: Palmarium Academic Publishing, 2017. – 96 с.