Оценка алгоритмов извлечения ключевых слов: инструментарий и ресурсы

Ванюшкин А.С., Псковский государственный университет
alexmandr@mail.ru
Гращенко Л.А., Академия ФСО России
graschenko@mail.ru

Аннотация

В работе описываются разработанный программный инструментарий, предназначенный для исследования алгоритмов извлечения ключевых слов, и вспомогательные лингвистические ресурсы. Приведен обзор 11 текстовых корпусов, востребованных в работах по выделению ключевых слов, показаны особенности их разметки. Ставится задача разметки создаваемого англо-русского корпуса аналитических текстов.

1 Введение

Автоматическое извлечение ключевых слов (КС) и фраз является одной из базовых процедур обработки текстов на естественных языках, вследствие чего данная задача на протяжении последних шестидесяти лет всесторонне рассматривалась многими исследователями. Но существенный рост числа исследований пришелся на последние 20 лет из-за повышения доступности вычислительных и информационных ресурсов.

Для известных алгоритмов извлечения ключевых слов (КС) из текстов на естественных языках наблюдается существенный разброс опубликованных значений показателей эффективности. Это обусловлено различиями в методологии и составе лингвистических ресурсов, используемых при их тестировании различными авторами. Кроме того, в отношении этих алгоритмов имеется недостаток информации об их практическом применении, так как они испытывались в основном исключительно разработчиками на ограниченном наборе языков, куда, как правило, русский язык не входил [Ванюшкин, Гращенко, 2016].

Для практического тестирования и верификации передовых алгоритмов и их адаптации к русскому языку необходима разработка программного исследовательского стенда. Такой инструментарий должен удовлетворять ряду требований:

– поддерживать гибкую настройку содержания этапов извлечения КС (предобработка, распознавание, постобработка);

– обладать дружественным графическим интерфейсом пользователя;

– реализовывать все продуктивные алгоритмы извлечения КС;

– поддерживать использование разнообразных лингвистических ресурсов;

– вычислять, визуализировать и хранить настраиваемые наборы разнородных данных - значения показателей эффективности исследуемых алгоритмов и характеристики обрабатываемых текстовых массивов;

– позволять варьировать схемы эксперимента и использовать их шаблоны;

– обладать модульностью и расширяемостью, в том числе языковой.

Кроме того, для объективного сравнения алгоритмов извлечения КС, созданных разными авторами в рамках различных подходов, необходимо проводить тестирование на одном и том же корпусе текстов. Именно так поступают организаторы соревнований разработчиков, предлагая всем участникам единый тестовый корпус. Однако если для оценки применимости решений к англоязычным текстам можно выбрать вариант из ранее апробированных размеченных ключевыми словами тестовых корпусов, то применительно к русскому языку такой корпус в открытом доступе не представлен. При этом целесообразно использовать, по крайней мере, двуязычный параллельный корпус, вследствие чего требуется обоснование состава и характеристик, а также формирование и разметка такого модельного корпуса.

Для решения указанных задач в настоящей работе на основе обсуждения предварительных результатов применения разработанного программного стенда выполним обзор существующих коллекций текстов, для которых актуальна задача извлечения КС, а также тестовых корпусов, апробированных для задачи извлечения КС в ранее выполненных иссле-

дованиях, с точки зрения особенностей их разметки.

2 Текущие программные разработки

Для реализации трехэтапной схемы извлечения КС программный стенд реализован в виде инструмента построения диаграммы процессов, рис. 1. Основной язык разработки – C# 5.0, визуальная часть программы использует компоненты Syncfusion[1].

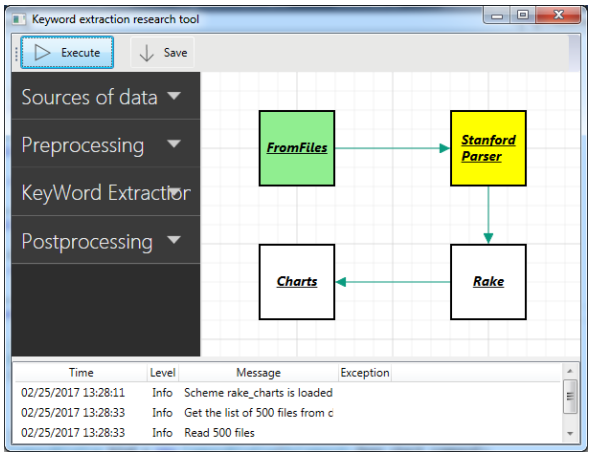

Рис. 1. Экранная форма главного окна исследовательского стенда

Работа с программой осуществляется путем визуального проектирования схемы обработки из доступных в коллекции элементов - процедур обработки текста, обладающих рядом настроек и представляющих собой отдельные динамически подключаемые библиотеки. В целях расширяемости приложения элементы обработки реализуют один из четырех абстрактных классов, рис. 2.

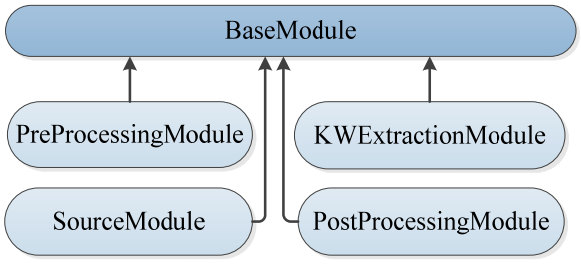

Рис. 2. Иерархия наследования подключаемых компонентов

На этапе загрузки программы при помощи механизма Reflection в стенд загружаются все доступные обработчики. Таким образом, для расширения функционала, например, добавления нового вида источника данных, или алгоритма извлечения КС, необходимо реализовать один из базовых классов обработчиков без изменения существующего кода.

На первом этапе в исследовательский стенд загружается текст или текстовый корпус. Здесь же происходит предварительная языкозависимая обработка: токенизация, частеречевая разметка (POS-tagging), стемминг. В текущей версии стенда для токенизации английских текстов используется широко применяемый Stanford Log-linear Part-Of-Speech Tagger[2], а для русских - библиотека Syncfusion.DocIO.DLS.

На втором этапе происходит непосредственно извлечение КС на основе одного из множества встроенных алгоритмов. На третьем этапе возможно выполнение следующих процедур: запись списка КС в файл, расчет точности (Precision), полноты (Recall), F-меры (F-measure), визуализация результатов и распределения длин текстов в корпусе на графиках.

Первоначально в состав доступных алгоритмов были включены три графовых алгоритма, реализующих структурные методы распознавания КС без обучения: Rake [Rose & etc., 2010], TextRank [Mihalcea, Tarau, 2004] и Palshikar [Palshikar, 2007]. Такой выбор обусловлен рядом преимуществ граф-ориентированных алгоритмов: языконезависимостью, относительной простотой программной реализации (относительно обучаемых) и отсутствием необходимости использования дополнительных лингвистических ресурсов. Их реализации прошли верификацию – были получены идентичные авторским результаты извлечения КС на тех же текстах при тех же настройках.

Дальнейшие эмпирические исследования подтвердили известное положение о возрастании сложности извлечения КС с увеличением длины входных текстов [Hasan, Ng, 2011]. Так, при испытаниях на различных тестовых корпусах было обнаружено, что значения показателей качества работы граф-ориентированных алгоритмов снижаются (или, по крайней мере, меняются в широких пределах) с увеличением размеров текстов (в словах). Примеры таких изменений результативности приведены ниже. На рисунке 3 представлены изменения F-меры для англоязычных текстов из корпуса Hulth-2003 при обработке по алгоритму Rake.

[1] www.syncfusion.com

[2] nlp.stanford.edu

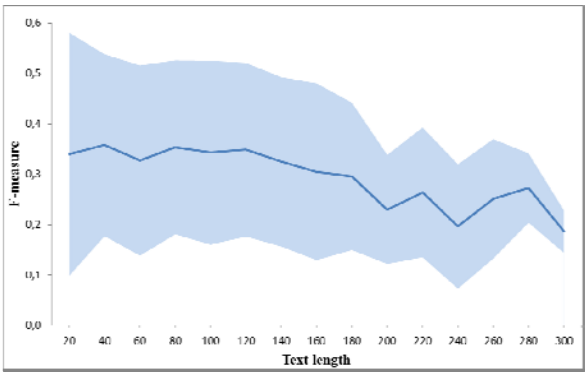

Рис. 3. Результативность работы алгоритма Rake для корпуса Hulth-2003

На рисунке 4 представлен пример изменения точности работы алгоритма TextRank для текстов из англоязычного корпуса 500N-KPCrowd-v1.1 [Marujo & etc., 2012].

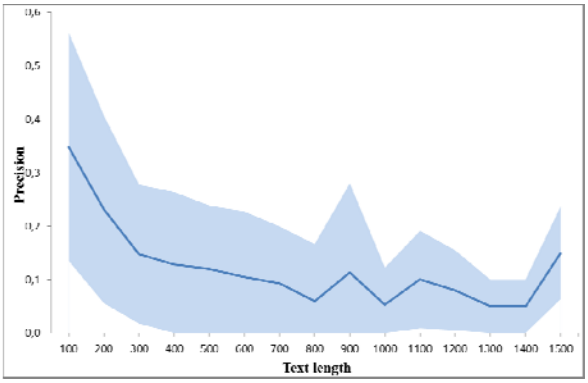

Рис. 4. Пример зависимости точности извлечения ключевых слов от длин текстов

Таким образом, наблюдается чувствительность графовых алгоритмов выделения ключевых слов к длинам входных текстов. Следовательно, вид и параметры закона распределения длин текстов (в словах), составляющих конкретный исследовательский корпус, определяют значения показателей эффективности, получаемые различными авторами. Вероятно, некоторое влияние также оказывает гомогенность корпуса по жанру и сложности текстов. Этим, вероятно, объясняется наблюдаемый разброс в публикуемых данных. Следовательно, необходимо выполнить обзор доступных лингвистических ресурсов и обосновать выбор оптимальных в контексте решаемой задачи текстовых корпусов.

3 Обзор доступных лингвистических ресурсов

3.1 Естественные коллекции текстов

Вопрос естественного распределения длин текстов рассматривался многими исследователями. Больше всего работ посвящено размерам постов в блогах, для которых длины текстов описываются распределениями с толстыми хвостами. Это справедливо как по отношению к пользовательским комментариям, сообщениям электронной почты, так и к длинам текстов, хранящихся на компьютерах пользователей. В работе [Blumenstock, 2008] предлагается считать длину текстов статей из Википедии показателем их качества, а в целом длины английских статей в ней распределены по логнормальному закону.

Для исследования существующих коллекций текстов в качестве информационно-статистической базы выбраны шесть веб-сайтов, содержащих объемные собрания аналитических статей различных тематик на английском языке. Данный выбор обусловлен предположением, что именно тематические тексты с элементами аналитики являются одним из основных объектов приложения задачи извлечения КС. Выбор ресурсов также определялся доступностью массовой загрузки веб-страниц с сайтов, так как доступ к архивам зачастую не предоставляется, а сайты имеют защиту от роботизированной загрузки.

После загрузки коллекций и корпусов автоматически производились парсинг страниц и извлечение открытого текста (plain text), далее выполнялась токенизация и подсчет количества слов в каждой статье.

Результаты анализа показали, что длины текстов в большинстве коллекций аналитических статей могут быть описаны логнормальным законом распределения. При этом большинство текстов лежит в диапазоне от 400 до 2500 слов, рис. 5.

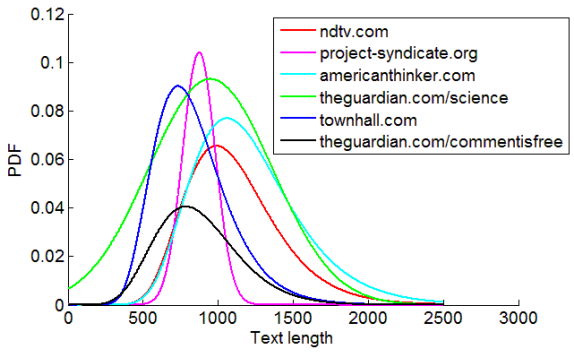

Рис. 5. Аппроксимирующие кривые распределений длин текстов в естественных коллекциях

Отклонения от формы кривой, описываемой функцией логнормального распределения, можно объяснить всевозможными редакционными ограничениями на содержание

и формат, а также длину публикуемых статей. Например, на ресурсе Project Syndicate авторам рекомендовано придерживаться длины статей около 1000 слов, вследствие чего распределение длин в данной коллекции близко к нормальному.

3.2 Корпусы для исследований в области извлечения ключевых слов

Обзор корпусов, апробированных в исследовательских работах по извлечению ключевых слов, показал их существенную дифференциацию по объему, составу, тематике и способу разметки КС. Так содержимое корпусов может представлять собой аннотации, аннотации с текстом или основной текст без аннотаций. Поэтому зачастую наблюдается разброс длин текстов для некоторых пар документов (в словах) в три порядка. А наличие текстов из десятков тысяч слов ставит под сомнение возможность и смысл использования алгоритмов выделения КС на всей их длине, без предварительного деления на смысловые части. Напротив, аннотации по своей сути содержат больший процент ключевых слов, чем тексты длиной в несколько тысяч слов. На рисунке 6 приведено сравнение распределений длин текстов в большинстве рассмотренных корпусов. Хорошо видно, что они значительно отличаются от параметров модельного распределения, полученного на основе исследования естественных коллекций текстов.

Также значительным недостатком доступных корпусов является «засоренность», так как многие из них содержат в текстах библиографию, таблицы, подрисуночные подписи и нетекстовые объекты. Кроме того, все рассмотренные корпуса являются моноязычными и не позволяют вести кросс-языковые исследования в области извлечения ключевых слов. Более подробный обзор существующих коллекций текстов и корпусов на данный момент готовится к публикации.

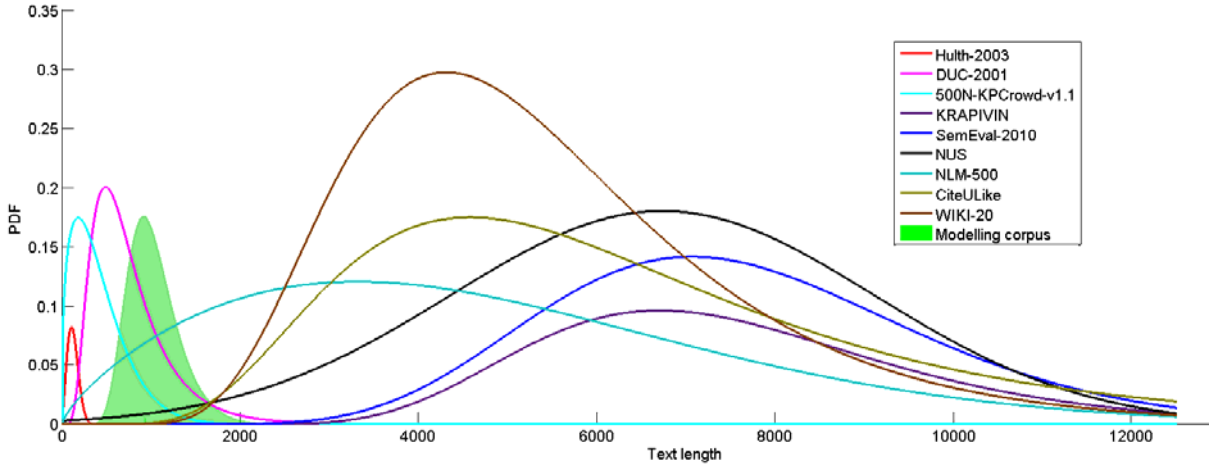

Рис. 6. Кривые распределений длин текстов в исследовательских корпусах

3.3 Разметка ключевыми словами

Разметка текстов ключевыми словами может осуществляться непосредственно их авторами, группой экспертов по тематике текста или при помощи широкой общественности.

Авторские КС могут хорошо отражать смысл документа, однако являются субъективным взглядом, так как набор КС здесь выделен одним человеком, и может не содержать многих подходящих терминов. Исследователи отмечают, что оценка алгоритмов не должна основываться только на данных наборах КС [Nguyen, Kan, 2007]. Поэтому при оценке алгоритмов в качестве модельного набора КС разумнее использовать либо выделенные несколькими экспертами, либо комбинировать ключевые фразы, выделенные разными способами. Однако ряд вопросов остается открытым. Во-первых, кого относить к категории экспертов. Часто из удобства к экспертам относят аспирантов, студентов и специалистов по индексации документов. Во-вторых, какую длину списка КС считать оптимальной. Существующие корпуса демонстрируют значительный разброс данного показателя. В таблице 1 приведена сравнительная

характеристика тестовых корпусов по особенностям разметки. Приводятся диапазоны (КС/документ) и средние (КС среднее) длин наборов КС на документ в рамках корпуса. Рассмотрим подробнее разметку КС существующих исследовательских корпусов.

Корпус **DUC-2001** изначально был подготовлен для задачи реферирования текстов в рамках конференции Document Understanding Conferences. Он состоит из 308 новостных статей и был размечен двумя аспирантами, при этом максимальное количество КС для одного текста ограничивалось 10. Разногласия при разметке решались путем обсуждения [Wan, Xiao, 2008].

Hulth-2003 содержит 2000 аннотаций по специальностям Computers and Control и Information Technology и изначально создавался для тестирования обучаемого алгоритма извлечения КС, поэтому разделен на три части: первая - для обучения, вторая – контрольная выборка и третья – тестовая [Hulth, 2003].

Корпус **NLM-500** размечен экспертами, при этом КС ограничены тезаурусом Medical Subject Headings [Gay, Kayaalp, Aronson, 2005]. В корпусе **NUS** насчитывается 211 статей, имеющих как авторскую разметку, так и разметку выполненную студентам, однако не для всех документов. При этом количество размечающих лиц одну статью здесь варьируется [Nguyen, Kan, 2007].

WIKI-20 размечен 15 парами студентов, обучающихся по дисциплинам информационных технологий, которым было рекомендовано присваивать около пяти терминов на статью. КС являются заголовками статей из энциклопедии Википедия [Medelyan, 2009].

Табл. 1. Характеристики доступных корпусов, размеченных ключевыми словами

№	Название	Год	Состав	КС/документ	КС среднее
1	DUC-2001	2001	Новостные статьи	3-14	8,1
2	500N-KPCrowd-v1.1	2012		5-254	39,7
3	Hulth-2003	2003	Аннотации статей по ИТ базы Inspec	1-31 (u); 1-15 (c)	9,6 (u); 4,5 (c)
4	NLM-500	2005	Документы базы PubMed	2-29	14,2
5	NUS	2007	Научные статьи конференций	2-9 (a); 3-20 (r)	4,3 (a); 9,1 (r)
6	WIKI-20	2008	Технические статьи	2-12	5,7
7	FAO-30	2008	Документы Продовольственной и Сельскохозяйственной Организаций при ООН	4-52 (r)	10,4
8	FAO-780	2008		2-23 (r)	8,0
9	KRAPIVIN	2009	Статьи конференций ACM	1-24 (a)	5,3
10	SemEval-2010	2010		1-13 (a); 7-34 (r)	3,8 (a); 12,5 (r)
11	CiteULike	2009	Научные статьи по биоинформатике	2-73	4,1

Примечание: a – авторские КС (author assigned keywords), r – экспертные КС (reader assigned keywords), c – контролируемые тезаурусом КС (controlled terms), u – неконтролируемые тезаурусом КС (uncontrolled terms).

KRAPIVIN размечен КС только авторами текстов [Krapivin, Autayeu, Marchese, 2009]. А разметку 180 документов **CiteULike** выполняли независимо 322 добровольца, при этом каждый обработал от одного до 25 текстов. Общее количество выделенных терминов 4638, но только 946 из них присутствуют в наборах хотя бы двух размечающих [Medelyan, 2009].

Экспертная разметка **SemEval-2010** выполнялась 50 студентами, каждый из которых размечал пять статей (с расчетом затрат времени на одну статью 10-15 мин). Таким образом, несмотря на значительное число участников процесса выделения КС, их результаты не комбинировались. В результате из 387 авторских КС 167 соответствуют выделенным экспертами. Авторы отмечают, что около 15% приписанных экспертами и 19% приписанных авторами КС не содержатся в текстах. Отличительной особенностью данного корпуса является представление эталонных списков КС – в виде основ слов [Kim et al., 2010].

Наборы ключевых слов корпусов **FAO-780** и **FAO-30** ограничены тезаурусом Agrovoc[1]. Экспертами здесь выступают сотрудники Продовольственной и сельскохозяйственной организаций при ООН [Medelyan, 2009].

Для разметки корпуса **500N-KPCrowd-v1.1** исследователи использовали краудсорсинговую платформу Amazon's Mechanical Turk[2]. Отмечается, что некоторые размечающие выделяли бессмысленные наборы слов. Для выбора только значимых результатов использовались несколько эвристических процедур,

[1] www.fao.org/agrovoc
[2] www.mturk.com

среди которых наличие в КС стоп-слов, слишком длинные последовательности (более десяти слов) и быстрое выполнение задания (менее 30 секунд). В результате 10 % выбранных КС были исключены [Marujo et al., 2012].

Из представленного обзора видно, что большинство доступных корпусов содержат в себе научные статьи. При этом длина списка КС, сопоставленного тексту, варьируется от нескольких единиц до нескольких десятков. Объемы и тематики содержимого корпусов также значительно различаются.

4 Создание и разметка корпуса

4.1 Создаваемый корпус

Подготавливаемый корпус включает аналитические статьи на английском языке и их переводы на русский, опубликованные на различных интернет ресурсах в отрытом доступе. Все тексты представлены в формате «plain text», в кодировке UTF-8 и не содержат в себе заголовков. Сам корпус является квотной выборкой из 203 статей.

Исходя из понятия текстового корпуса в корпусной лингвистике, основными его свойствами являются [Николаев, Митренина, Ландо, 2016]: электронный формат; репрезентативность; прагматическая ориентированность и размеченность.

Первое условие выполняется для любого современного корпуса. Для обеспечения репрезентативности набор текстов отбирался так, чтобы соблюсти пропорции в распределении длин текстов относительно условной генеральной совокупности, в качестве представления о которой берутся усредненные параметры распределений рассмотренных выше естественных коллекций. Прагматическая ориентированность подготавливаемого корпуса подразумевалась изначально. Для выполнения основного условия – разметки корпуса (в данном случае экстралингвистической), ведется разработка соответствующего проблемно-ориентированного программного обеспечения. Планируется двухэтапная схема автоматизированной разметки представленного исследовательского корпуса ключевыми словами: сначала экспертами, а затем добровольцами.

4.2 Обзор средств разметки

Ручная разметка (аннотирование) текстов - дорогостоящая и трудоемкая задача. Так, исходя из исследования, проведённого на студентах, только для чтения с усвоением 50% русскоязычной части подготавливаемого корпуса одному эксперту потребуется около двух часов [Вормсбехер, Кабин, 1980]. Однако разметка КС подразумевает продуктивное чтение и обдумывание содержимого. Как указано выше, наборы КС, выделенные различными людьми, значительно отличаются, поэтому предлагается размечать каждый текст минимум пятью экспертами. Таким образом, для минимизации временных затрат размечающих следует использовать программные средства позволяющие выполнять разметку максимально просто и эффективно. В то же время исследователю необходимо контролировать процесс разметки: ход выполнения, количество участников. На данный момент в открытом доступе представлены различные программные средства для разметки корпусов (*annotation tools*)[1]. Условно их можно разделить на обособленные *(stand-alone)* и веб-ориентированные *(web-based)*. При этом акцент разработчиков в последние годы сместился в сторону веб-приложений. Данные системы обладают рядом преимуществ:

- возможность одновременной разметки одного документа несколькими людьми;
- не требуют установки дополнительных программных средств, кроме браузера;
- гибкое разграничение прав доступа;
- отображение текущего прогресса процесса разметки;
- возможность модификации размечаемого корпуса.

Среди недостатков веб-решений стоит отметить необходимость использования сервера приложений, что достаточно сложно, т.к. каждая из систем обладает рядом специфических требований к системному и программному окружению. Также аренда серверных мощностей требует определенных затрат.

Современные веб-ориентированные средства разметки поддерживают реализацию трёх последовательных этапов. На первом этапе необходимо создать проект, добавить пользователей, загрузить/конвертировать в систему данные, добавить необходимые тэги. При этом некоторые системы позволяют разграничивать права пользователей по каждому документу.

[1] annotation.exmaralda.org,
omictools.com/text-annotation-category,
knot.fit.vutbr.cz/annotations/tool_comparison

На втором этапе приглашенные эксперты или представители общественности выполняют разметку в соответствии с заданием. Данный процесс в большинстве рассмотренных средств осуществляется схожим образом по принципу WYSIWYG[1]: выделение фрагмента текста мышкой и его маркировка доступными тэгами, рис. 7.

Рис. 7. Типовая последовательность этапов разметки текстового корпуса

В некоторых системах возможно создание отношений. На рис. 8 приведен пример использования одной из популярных средств аннотирования BRAT[2].

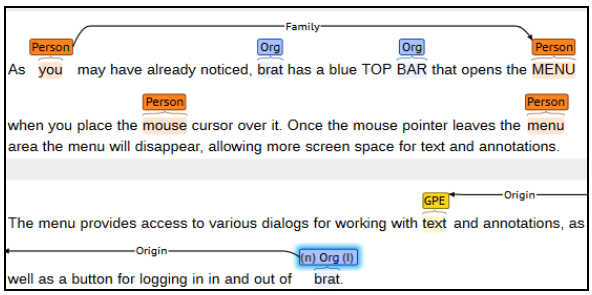

Рис. 8. Разметка текста в системе BRAT

Здесь же возможны корректировка администратором текущей коллекции, например добавление или удаление документов, просмотр текущего состояния выполнения задания.

На заключительном этапе исследователю предоставляются обобщенные полученные результаты. Здесь могут быть доступны такие функции как экспорт данных, сравнение и редактирование разметки, получение информации о работе каждого размечающего.

Несмотря на то, что разметка корпуса ключевыми словами является более узкоспециализированной задачей относительно возможностей, предоставляемых существующими средствами аннотирования, тем не менее, обладает рядом особенностей. Например, исходя из доступных сопровождающих документаций, удалось установить, что только WebAnno позволяет привлекать для разметки общественность [Yimam et al., 2014]. А учет затраченного на обработку одного документа времени присутствует лишь в системе BRAT [Stenetorp, Pyysalo, Topi, 2012]. Однако, наличие таких возможностей необходимо для качественной разметки предлагаемого корпуса.

Таким образом, вопрос выбора инструмента аннотирования является открытым, и дальнейшая исследовательская работа будет направлена на испытания существующих решений, а при необходимости создание нового программного средства разметки текстовых данных.

5 Выводы

Текущие возможности исследовательского стенда позволяют производить анализ алгоритмов извлечения КС и текстовых корпусов. Первые результаты испытаний графовых алгоритмов показали значительный разброс точности извлечения ключевых слов при варьировании размеров текстов.

Анализ существующих естественных коллекций текстов показал, что большинство документов обладают длиной в диапазоне от 400 до 2500 слов, а распределение длин близко к логнормальному. Однако разброс длин текстов внутри некоторых корпусов, используемых для тестирования алгоритмов извлечения КС, достигает трех порядков. Кроме того, все рассмотренные корпуса являются моноязычными. С учетом приведенных недостатков, сформирован и ожидает разметки с последующим введением в научный оборот новый англо-русский корпус.

Дальнейшая исследовательская работа будет направлена на увеличение числа реализованных в экспериментальном стенде алгоритмов и качественную разметку ключевыми словами создаваемого текстового корпуса.

Список литературы

Ванюшкин А.С. *Методы и алгоритмы извлечения ключевых слов* / А.С. Ванюшкин, Л.А. Гращенко // Новые информационные технологии в автоматизированных системах. – 2016. – №. 19 – С. 85-93.

Вормсбехер В.Ф. *100 страниц в час* / Вормсбехер В.Ф., Кабин В.А. – Кемерово: Кемеровское книжное издательство, 1980. – 144 с.

Николаев И.С., Митренина О.В., Ландо Т.М. *Прикладная и компьютерная лингвистика* - М.: URSS, 2016. – 320 с.

[1] what you see is what you get
[2] brat.nlplab.org

Blumenstock J.E. *Size matters: word count as a measure of quality on Wikipedia* / J.E. Blumenstock // Proceedings of the 17th International Conference on World Wide Web. – 2008. – pp. 1095-1096.

Gay C.W. *Semi-automatic indexing of full text biomedical articles* / C.W Gay, M. Kayaalp, A.R. Aronson // AMIA Annu. Symp. Proc. – 2005. – pp. 271-275.

Hasan K. *Automatic Keyphrase Extraction: A Survey of the State of the Art* / K. Hasan, V. Ng // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. – 2011. – Vol 1. – pp. 1262-1273.

Hulth A. *Improved Automatic Keyword Extraction Given More Linguistic Knowledge* / A. Hulth // EMNLP'03 Proc. 2003 Conf. Empir. Methods Nat. Lang. Process. – 2003. – № 2000. – pp. 216-223.

Kim S. *Semeval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles* / S. Kim, O. Medelyan, M Kan, T. Baldwin // Proceedings of the 5th International Workshop on Semantic Evaluation. – 2010. – pp. 21-26.

Krapivin M., Autayeu A., Marchese M. *Large Dataset for Keyphrases Extraction.* URL: http://eprints.biblio.unitn.it/archive/00001671/01/disi09055-krapivin-autayeu-marchese.pdf (дата обращения 01.02.2017).

Marujo L. *Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing* / L. Marujo, A. Gershman, J.G. Carbonell, R.E Frederking, J.P. Neto // 8th International Conference on Language Resources and Evaluation (LREC 2012). – 2012. pp. – 399-403.

Medelyan O. *Human-competitive automatic topic indexing*. PhD thesis / O. Medelyan. – Hamilton, 2009 –244 p.

Mihalcea R. *TextRank: Bringing order into texts* / R. Mihalcea, P. Tarau // Proceedings of EMNLP 2004. – 2004. – Vol. 4. – pp. 404-411.

Nguyen T. *Keyphrase extraction in scientific publications* / T. Nguyen, M. Kan // Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers. – 2007. – pp. 317-326.

Palshikar G. *Keyword Extraction from a Single Document Using Centrality Measures* / G. Palshikar // Pattern Recognition and Machine Intelligence, LNCS. – 2007. – Vol. 4851. – Iss l. – pp. 503-510.

Rose S. *Automatic Keyword Extraction from Individual Documents* / S. Rose, D. Engel, N. Cramer, W. Cowley // Text Min. Appl. Theory. – 2010. – pp. 1-20.

Stenetorp P. *BRAT : a Web-based Tool for NLP-Assisted Text Annotation* / P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii // Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. – 2012. – pp 102-107.

Wan X. *Single Document Keyphrase Extraction Using Neighborhood Knowledge* / X. Wan, J. Xiao // Chicago. – 2008. – pp 855-860.

Yimam S.M. *Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno* / S.M. Yimam, R. Eckart de Castilho, I. Gurevych, C Biemann // Proceedings of ACL-2014. – 2014. – pp. 91-96.