УДК 004.023

АВТОМАТИЧЕСКИЙ РАЗБОР И АННОТИРОВАНИЕ СТАТЕЙ

Ефремова М.И.

Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, e-mail: korixd@gmail.com

В данном исследовании рассматривается задача автоматического создания аннотаций, для решения которой был разработан метод, сочетающий извлекающий подход (он используется для нахождения основных доминант текста и ключевых слов) и генерирующий (для формирования итоговых предложений аннотации). В основе обоих подходов лежат шаблоны, составленные на основе морфологического анализа текстов и их семантической разметки, а также словаря клише. Формирование шаблонов происходит при помощи контекстно-свободных грамматик, семантической разметки и морфологического анализа исходного текста и словарей ключевых слов. Метод состоит из нескольких этапов: для начала из статьи выделяются ключевые слова, на их основе извлекаются N-граммы, из которых впоследствии составляются предложения и уже из предложений строится итоговая аннотация.

Ключевые слова: аннотирование, квазиреферирование, контекстно-свободные грамматики, морфологический анализ, научная статья

AUTOMATIC ARTICLE PARSING AND ANNOTATION

Efremova M.I.

Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, Saint-Petersburg, e-mail: korixd@gmail.com

Automatic article annotation is the problem to be considered in this study. The suggested approach is a combination of extracting method (to find the main ideas of the text and its keywords) and generating method (to create final sentences of annotation). Both used approaches are based on patterns, which are pieced out from morphological analysis, semantic formating and clichés vocabulary. Drafting patterns is done by context-free grammatics and dictionaries of keywords. Method consist of several stages: at first, key words of article are extracted, then, retrieved N-grams, of which later final sentences are constructed, and finally they form the result annotation.

Keywords: annotate, quasiabstract, context-free grammar, morphological analysis, research article

В настоящее время все большее применение находят системы автоматического реферирования информации, расположенной в Интернете. Они используются для того, чтобы получить краткое содержание, а также краткую характеристику первоисточника, тем самым уменьшив время на его изучение. Аннотация – это краткая характеристика научной статьи с точки зрения ее назначения, содержания, вида, формы и других особенностей.

По способу построения текста методы автоматического реферирования делятся на две группы: извлекающие (квазиреферирование, Sentence extraction) и генерирующие (генерация реферата с порождением нового текста, Abstraction) [1, 2]. При использовании извлекающих методов из исходного текста выделяются наиболее важные фрагменты (предложения, абзацы). При этом данные фрагменты не обрабатывают, а извлекают в таком порядке и виде, в каком они приведены в тексте. Генерирующие методы предполагают автоматическое определение содержания реферата с последующей генерацией нового текста, не представленного явно в тексте исходного документа [3, 4]. При использовании генерирующих методов текст реферата строится на правилах, предполагающих наличие лингвистической базы знаний.

В настоящее время автоматическое аннотирование и реферирование производится в таких системах, как:

– Microsoft Word (функция автоматического реферирования);
– Intelligent Text Miner (IBM);
– Oracle Context;
– NewsBlaster [6];
– Ultimate Research Assistant [7];
– Google News, Яндекс. Новости [8].

Одна из проблем в применении извлекающих методов реферирования – это требование сжатия. Объем аннотации или реферата должен составлять от 5 до 30% исходного текста. Кроме этого, проблемой является недостаток различных лингвистических ресурсов (толковые, лексические и частотные словари, грамматики, тезаурус) и сложность автоматического создания текстов на естественном языке. Кроме того, большинство алгоритмов, построенных на машинном обучении, требуют больших вычислительных ресурсов.

Цель исследования. Перед автором статьи была поставлена цель – разработать

метод автоматического создания аннотаций к статьям. В ходе разработки была создана система, на вход которой подаётся статья, набор ключевых слов, согласующийся с контекстом этой статьи, а также словарь клише со встроенными морфологическими правилами, а результатом работы является краткая аннотация.

Предложенный подход состоит из нескольких модулей, каждый из которых выполняет определённые функции. Необработанная статья поступает на вход модуля извлечения ключевых слов, который служит для определения основных слов и установления веса каждого слова. Полученные на этом этапе данные и сама статья подаются в модуль извлечения фактов, хранящий в себе морфологические правила, на которых строятся грамматики для извлечения основных фактов из статьи. На выходе этого модуля формируется набор фактов (а именно – лексически согласованных N-грамм, включающих в себя ключевые слова), а также рассчитываются относительные веса для каждого факта. Полученные на предыдущих этапах данные поступают в модуль создания аннотаций, который выбирает из перечня фактов наиболее подходящие (с большим весом) и соединяет их в аннотацию в соответствии с правилами, перечисленными в словаре клише.

Разработанный алгоритм основан на эвристическом методе, ПО написано на языке программирования java, для извлечения фактов из текста по составленным шаблонам используется томита-парсер.

Модуль извлечения ключевых слов

На вход модуля поступает файл с описанием ключевых слов (используется синтаксис грамматик Томита-парсера [8], в примере далее для упрощения восприятия показано текстовое описание грамматик). Примеры используемых правил извлечения и используемые для них эмпирически вычисленные веса:

- Слово, имеющее грамматические признаки: [именительный или винительный падеж, одушевлённое, единственное число] – 0,8.
- Слово, имеющее грамматические признаки: [родительный падеж, одушевлённое, единственное число] – 0,75.
- Глагол – 0,75.
- Редко встречающееся или забытое слово – 0,72.
- …..
- Географическое название – 0.

Из статьи извлекаются ключевые слова, параллельно определяется частота каждого слова, слова с частотой 1 убираются из дальнейшего анализа. На выходе каждому слову присваивается относительный вес, который рассчитывается по формуле

$$\frac{a_i}{Median(a)} \cdot v_i,$$

где a_i – это частота конкретного слова; v_i – словарный вес извлечённой сущности.

Здесь и далее в примерах работы метода в качестве входных данных будет использоваться эта самая статья.

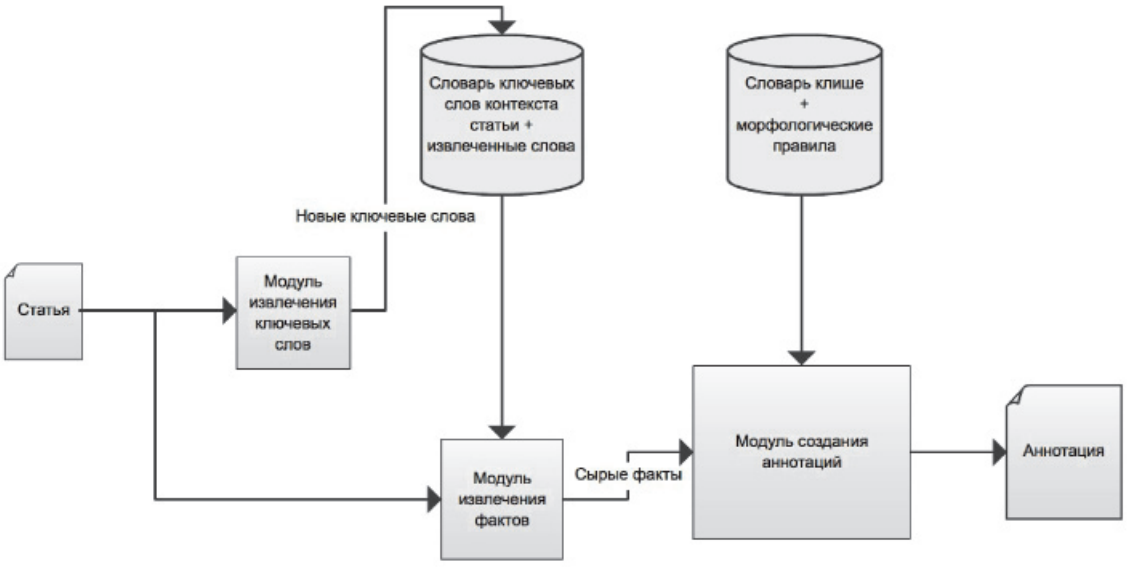

Схема работы

Пример получившегося после экстракции массива сущностей (не показаны сущности с весами < 0,32):

{вес = 1.5, извлечение = 1.1785, модуль = 1.071, пример = 0.8571, слово = 0.8571, статья = 0.75, аннотация = 0.75, грамматик = 0.72, клише = 0.6428, код = 0.6171, сущность = 0.6171, алгоритм = 0.5357, создание = 0.4285, вход = 0.4285, коэффициент = 0.4285, метод = 0.3214, набор = 0.3214, обработка = 0.3214, описание = 0.3214, правило = 0.3214, частота = 0.3214, номер = 0.3214}

Модуль извлечения фактов

На вход модуля поступает статья, ключевые слова контекста статьи (например, «результат», «задача», «проблема», «алгоритм», «исследование», «анализ», «факторы» и т.д.) и ключевые слова, полученные на предыдущем этапе с соответствующими им весами. В результате работы получаются несогласованные факты.

Модуль также содержит в себе набор грамматик на языке томита-парсера [9], которые описывают строение извлекаемых сущностей. Используются следующие обозначения (приведена расшифровка только тех обозначений, которые встречаются в примерах далее, полные обозначения можно найти в документации томита-парсера):

1. Части речи: Word – любое слово, Noun – существительное, Adj – прилагательное, Verb – глагол, APRO – местоименное прилагательное.

2. Gnc-agr – требование согласования нескольких слов по роду, падежу и числу. Gn-agr – требование согласования по роду и падежу.

3. kwtype = ‘ключевые_слова’ – ключевое слово, полученное на предыдущем этапе, или словарное слово из контекста статьи.

4. indic – изъявительное наклонение, abl – аблатив, исходный падеж, nom – номинатив, именительный падеж, partcp – причастие, praes – настоящее время.

5. + – одно и более повторение, * – 0 и более повторений.

Примеры правил извлечения:

1. Word < gnc-agr > * Noun < kwtype = ‘ключевые_слова’, gnc-agr > Word APRO Word*;

2. Word < gnc-agr > * Noun < kwtype = ‘ключевые_слова’, gnc-agr > Word < gnc-agr > *;

3. Verb < ndic > Word < gnc-agr > * Noun < kwtype = ‘ключевые_слова’, gnc-agr > Word < gnc-agr > *;

4. Word < gnc-agr[1] > * Noun < kwtype = ‘ключевые_слова’, gnc-agr > Word < gnc-agr > Word < abl > + ;

5. Verb < indic > Word < gnc-agr > * Noun < kwtype = ‘ключевые_слова’, gnc-agr > Word < gnc-agr > Word < abl > + ;

6. …..

Примеры извлечённых из текста фактов:

1) из проблем в применении извлекающих методов;

2) задача разработать метод автоматического создания аннотаций;

3) в данный исследовании предложен;

4) морфологический анализа текстов;

5) словарь клише;

6) помощь контекстно-свободная грамматик;

7) служить для извлечения основных слов;

8) с описанием ключевых слов;

9) использоваться синтаксис грамматик Томита-парсера;

10) каждый слову присваивается, относительный вес;

11) факт с большими весами;

12) последовательно меняя их морфологическую форму, подставлялся.

Для того, чтобы рассчитать удельный вес каждого полученного факта и «удачность» его использования в аннотации, кроме всего прочего следует учитывать номер абзаца, из которого этот факт был получен – в начале и в конце текста обычно располагается наиболее важная информация.

Для этих расчётов будет использована следующая формула:

$$c_i \cdot \alpha \cdot \frac{\sum_{0}^{N} l_i \cdot mystem_i}{N},$$

где c_i – словарный вес факта; l_i – вес для слова, посчитанный на предыдущем этапе; mystemi – уверенность утилиты mystem в том, что морфологические характеристики

сущности определены верно; α – это коэффициент абзаца, который равен

$$\alpha = \begin{cases} 0,7 \ \ if \ \ x \in (1;3), \\ 0,3 \ \ if \ \ x \in (3;N-2), \\ 0,8 \ \ if \ \ x \in (N-2;N). \end{cases}$$

Задача создания полной аннотации сводится к следующим подзадачам:

– создать грамматические правила для словаря клише;

– составить связный текст из упорядоченных предложений.

Примеры грамматических правил на основе словаря клише (вместе с этим для каждого правила проставляется коэффициент, который отвечает за номер предложения в составе аннотации, на месте которого может стоять фраза, построенная по данному правилу):

1. ‘Рассматривается метод’ <gnc-arg[1]> Word + Noun <GU = [acc], gnc-arg[1]> Word + ; ‘Приводится описание|пример’ <gnc-arg[1]> Noun <GU = [acc, nom], gnc-arg[1]> Word + ; ‘На основе метода’ <gnc-arg[1]> Noun <GU = [gen], gnc-arg[1]> Word + ;

2. ‘В даннои работе рассматривается’ <gnc-arg[1]> Word + Noun <GU = [acc], gnc-arg[1]> Word + ‘в применении к’ Word + ‘которыи’ Verb Word +

Из отобранных на предыдущем этапе фактов выбираются факты с большими весами и, после изменения их морфологической формы, подставляются на место грамматических правил в клише. Если форма найдена – то полученное предложение сохраняется как кандидат на присутствие в аннотации.

Пример получившейся аннотации:

Рассматривается задача автоматического создания аннотаций. В данном исследовании предложен подход, сочетающий извлекающий метод (для нахождения основных идей текста и ключевых слов) и генерирующий метод, который строится на шаблонах, составленных на основе морфологического анализа текстов и словаря клише. Составление шаблонов происходит при помощи контекстно-свободных грамматик и словарей ключевых слов. Для начала из статьи выделяются основные слова, на их основе извлекаются N-граммы и уже из них формируется итоговая аннотация.

Главная сложность реализации поставленной задачи заключалась в последнем этапе, потому что не все автоматически полученные предложения будут согласованы лексически, особенно на аннотациях больших размеров. Однако же, даже из автоматически сгенеренных предложений можно отобрать те, которые окажутся корректными, и составить аннотацию.

Заключение

Был разработан метод автоматического создания аннотаций, состоящий из нескольких этапов: для начала из текста выделяются ключевые слова, далее на их основе составляются ключевые N-граммы, далее с помощью словаря клише и описания предметной области статьи составляются предложения. Полученные на последнем этапе предложения компилируются в итоговую аннотацию.

Список литературы

1. Абрамова Н.Н., Абрамов В.Е. Автоматическое составление обзорных рефератов новостных сюжетов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL’2007: труды 9-й Всероссийской научной конференции. – Переславль-Залесский, Россия, 2007.

2. Лукашевич Н.В., Добров Б.В. Автоматическое аннотирование новостных кластеров на основе тематического представления // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог». Периодическое издание. – 2009. – Вып. 8 (15).

3. Тарасов С.Д. Исследование и оптимизация параметров алгоритма Manifold Ranking на основе метрики автоматической оценки качества обзорного реферирования ROUGE-RUS // Электронные библиотеки. Перспективные методы и технологии, электронные коллекции: труды XI Всероссийской научной конференции. – Петрозаводск, 2009. – С. 86–93.

4. Luhn, H.P. The Automatic Creation of Literature Abstracts // IBM Journal of Research and Development. – 1958. – V. 2, № 2. – P. 159–165.

5. Edmundson, H.P. New methods in automatic extracting // Newspaper of ACM tea, 16–2 (1969). – P. 264–285.

6. Noura Farra, Nadi Tomeh, Alla Rozovskaya, Nizar Habash. Generalized character-level spelling error correction. In Proceedings of the Conference of the Association for Computational Linguistics (ACL), Baltimore, Maryland, USA, 2014 [Computer file] URL: http://www.cs.columbia.edu/nlp/papers/2014/W14-2202.pdf (дата обращения: 20.03.2015).

7. Or Biran and Kathleen McKeown. Justification narratives for individual classifications. In Proceedings of the AutoML workshop at ICML 2014, Beijing, China, 2014. [Computer file] URL: http://www.cs.columbia.edu/nlp/papers/2014/justification_automl_2014.pdf (дата обращения: 22.03.2015).

References

1. Abramova N.N., Abramov V.E. Avtomaticheskoe sostavlenie obzornyx referatov novostnyx syuzhetov. // Trudy 9-oĭ Vserossiĭskoĭ nauchnoĭ konferencii «Elektronnye biblioteki: perspektivnye metody i texnologii, elektronnye kollekcii» RCDL2007. Pereslavl-Zalesskiĭ, Rossiya, 2007.

2. Lukashevich N.V., Dobrov B.V. Avtomaticheskoe annotirovanie novostnyx klasterov na osnove tematicheskogo predstavleniya // Kompyuternaya lingvistika i intellektualnye texnologii. Po materialam ezhegodnoĭ mezhdunarodnoĭ konferencii «Dialog». Periodicheskoe izdanie. 2009. Vypusk 8 (15).

3. Tarasov S.D. Issledovanie i optimizaciya parametrov algoritma Manifold Ranking na osnove metriki avtomaticheskoĭ ocenki kachestva obzornogo referirovaniya ROUGE-RUS // Trudy XI Vserossiĭskoĭ nauchnoĭ konferencii «Elektronnye biblioteki. Perspektivnye metody i texnologii, elektronnye kollekcii». Petrozavodsk, 2009. pp. 86–93.

4. Luhn, H.P. The Automatic Creation of Literature Abstracts / H.P. Luhn // IBM Journal of Research and Development. 1958. V. 2, no. 2. pp. 159–165.

5. Edmundson, H.P. New methods in automatic extracting / H.P. Edmundson // Newspaper of ACM tea, 16–2 (1969). P. 264-285.

6. Noura Farra, Nadi Tomeh, Alla Rozovskaya, Nizar Habash. Generalized character-level spelling error correction. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Baltimore, Maryland, USA, 2014 [Computer file] URL: http://www.cs.columbia.edu/nlp/papers/2014/W14-2202.pdf (дата обращения: 20.03.2015).

7. Or Biran and Kathleen McKeown. Justification narratives for individual classifications. In *Proceedings of the AutoML workshop at ICML 2014*, Beijing, China, 2014. [Computer file] URL: http://www.cs.columbia.edu/nlp/papers/2014/justification_automl_2014.pdf (дата обращения: 22.03.2015).

Рецензенты:

Бессмертный И.А., д.т.н., доцент, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, г. Санкт-Петербург;

Назаров И.А., д.т.н., помощник ректора, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина) (СПбГЭТУ «ЛЭТИ»), г. Санкт-Петербург.