УДК 004.912

ИЗВЛЕЧЕНИЕ И РАНЖИРОВАНИЕ КЛЮЧЕВЫХ ФРАЗ В ЗАДАЧЕ АННОТИРОВАНИЯ

С.В. Попова, И.А. Ходырев

Для решения задачи аннотирования проводится сравнительный анализ двух подходов ранжирования ключевых фраз. Первый основан на оценке веса извлекаемых фраз с помощью TextRank, второй основан на использовании *tf-idf* оценки. Исследование проведено на базе коллекции INSPEC dataset. Представлены описание экспериментов и сравнительные результаты. Экспериментально показано, что подход, основанный на использовании *tf-idf*, дает лучший результат.

Ключевые слова: аннотирование, извлечение и ранжирование ключевых фраз, оценка качества аннотаций.

Введение

Тенденция к распространению электронных форматов представления научной информации стимулирует активное развитие научного сектора Интернета. Выражено это появлением огромного числа электронных публикаций и каталогов цитирования, доступных через сеть интернет, что, в свою очередь, способствует развитию и научных электронных библиотек. Очевидно, что комфортная работа пользователя с таким большим объемом информации невозможна без быстрого автоматического поиска нужных материалов. Для решения этой задачи необходимы данные о смысловом содержании документа, представленного в виде короткой аннотации. В работе под аннотацией понимается список ключевых слов/словосочетаний (фраз), характеризующих электронный документ. Наборы ключевых фраз или слов могут быть также использованы в задачах кластеризации и классификации, в задаче автоматического построения/пополнения онтологий, в задаче определения основных трендов, в задаче поиска новой информации и т.д. Под аннотированием в работе будем иметь в виду автоматическое извлечение из текста ключевых слов/словосочетаний (фраз).

Для решения задачи аннотирования выделяют два подхода. Первый использует обучающую выборку, второй – нет.

В первом подходе задача сводится к разработке классификатора, определяющего для поступившего на вход текста, какие из его частей являются ключевыми фразами, а какие нет [1, 2]. В работе [2] предложен генетический алгоритм и параметризованная система по извлечению ключевых фраз Extractor. Генетический алгоритм позволяет определить оптимальные значения параметров. В [1] использован наивный байесовский классификатор. В [3] выполнена интеграция лингвистических данных в машинное обучение, показано преимущество использования информации о частях речи.

В рамках второго подхода наиболее популярным является метод, основанный на представлении текста в виде графа, предложенный в работе [4]. Вершины графа – целостные части текста (отдельные слова, *n*-граммы, предложения). Веса дуг графа характеризуют тип связи между вершинами по выбранному принципу (например, встречаться вместе в окне размера *n*, т.е. на расстоянии не более *n* слов друг от друга). В [4] в качестве вершин графа рассматриваются отдельные слова текста; вес дуги, соединяющей две вершины-слова, показывает, сколько раз эти два слова встретились в тексте в окне *n*. Для оценки веса каждой вершины-слова в [4] используется величина, основанная на модификации формулы PageRank [5]:

$$S(v_i) = (1-d) + d \sum_{V_j \in In(V_i)} \frac{1}{|Out(v_j)|} S(v_j),$$

где $In(v_i)$ – дуги, входящие в вершину v_i; $Out(v_j)$ – дуги, исходящие из вершины v_j. Представленная выше формула была изменена [4] с учетом того, что каждая дуга имеет вес w:

$$S(v_i) = (1-d) + d \sum_{v_j \in In(V_i)} \frac{w_{ij}}{\sum_{v_k \in Out(v_j)} w_{jk}}. \qquad (1)$$

Формула (1) получила название TextRank. Данная формула в [4] используется для расчета весов вершин-слов. Вершины-слова ранжируются по значению их веса, после чего отбирается только часть вершин с наибольшим весом. В [4] отобранные таким образом слова склеивались в фразы. Два слова объединяются в одну фразу, если в оригинальном тексте они непосредственно следуют друг за другом. В работе [8], помимо слов, рассмотрены многословные части текста – тогда вершинами графа текста являются группы слов. В качестве таких групп слов, например, взяты *n*-граммы (последовательность из *n* слов, следующих в тексте друг за другом), существительные фразы [8]. Для полученных многословных вершин рассчитывался вес на основе (1), лучшие вершины отбирались как ключевые фразы. После этого две последовательности слов склеивались в одну фразу, если они непосредственно следовали друг за другом в оригинальном тексте. В противном случае группа слов, образующих вершину, рассматривалась как самостоятельная фраза. Использование многословных частей текста при описанном выше подходе не дает лучших результатов в сравнении с использованием слов [8].

Оценка веса части текста с помощью TextRank в задаче аннотирования получила дальнейшее развитие. В работе [6] при построении графа учитывается содержание *k* ближайших документов. В [7] учитывается информация о семантической близости между построенными вершинами, дополнительно используются WordNet и Wikipedia. Полученные результаты являются одними из лучших в предметной области.

В данной работе задача извлечения ключевых фраз разделена на два этапа: 1) построение фраз-претендентов; 2) оценка веса каждой фразы-претендента и выбор лучших *k* из них как ключевых фраз. Также как и в работе [4], вес слов рассчитывается с помощью (1). Но, в отличие от [4], здесь не отбираются слова с наивысшей величиной TextRank, которые затем склеиваются в фразы. Вместо этого все слова после предобработки склеиваются в фразы, после чего оценивается вес каждой фразы, и лучшие фразы отбираются в качестве ключевых. Вес фразы считается равным весу лучшего слова (лучшее слово – слово в фразе с набольшим весом).

В работе поставлена задача проверки адекватности оценки веса фразы на основе информации о весах входящих в фразу слов, оцененных с помощью (1). Для сравнительного анализа в работе использован второй способ оценки весов слов и фраз соответственно [8]. Для этого использована популярная оценка веса слова *tf-idf* [11], которая рассчитывается для каждого конкретного слова в каждом документе как произведение частоты слова в данном документе *tf* на инвертированную частоту документов *idf*,

$$idf = \log \frac{|N|}{df},$$

где N – множество документов, df – число документов, в которых хотя бы раз встретилось слово. Далее будем обозначать вес слова v_i в документе как (*tf-idf*)(v_i). С помощью *tf-idf* оценивается вес каждого слова в документе. Вес фразы, построенной для текста, считается равным весу слова, входящего в фрау с максимальным значением *tf-idf*.

Для работы выбрана коллекция INSPEC dataset – одна из самых популярных в исследованиях по аннотированию текстов ключевыми фразами [3, 4, 7, 8, 10]. Коллекция содержит англоязычные аннотации к научным публикациям (abstracts, далее в тексте использован термин «абстракт», под которым понимается аннотация к научной публикации. Заметим, что термин «абстракт» использован во избежание путаницы, так как термин «аннотация» в работе используется для обозначения набора ключевых фраз документа). Коллекция состоит из трех подколлекций: training dataset (1000 документов), evaluation dataset (500 документов), testing dataset (500 документов). Каждый текст имеет «золотой стандарт», состоящий из фраз, отобранных для документа экспертом. Золотой стандарт включает в себя два подмножества аннотаций: contr set и unconter set. Аналогично [3, 4, 7, 8, 10], в работе использовано множество uncontr set. Подробное описание коллекции приведено в [3].

Оценка качества

Изначально для оценки качества извлеченных ключевых фраз использовалась оценка, основанная на F-score [3, 11], интегрирующая информацию о точности и полноте извлеченных фраз (Precision и Recall [11]). Однако данный способ оценки не накладывает четких ограничений на число извлеченных фраз. В работе [8] предложен подход, основанный на использовании R-Precision вместо F-score. R-Precision – значение Precision при условии, что число извлеченных ключевых фраз в точности совпадает с числом фраз в золотом стандарте. Для расчета R-Precision требуется информация о числе правильно извлеченных фраз. Пусть G_t – множество автоматически извлеченных фраз из текста t, C_t – множество ключевых фраз золотого стандарта для текста t (в золотой стандарт для каждого текста входят идеальные аннотации, вручную построенные экспертом). Возникает вопрос, какие из фраз G_t принадлежат $G_t \cap C_t$. В большинстве работ по извлечению ключевых фраз используется точное совпадение. Тогда фраза k («продвинутый автоматический перевод») и фраза золотого стандарта g («автоматический перевод»)

распознаются как различные. Аналогично ситуация возникает, например, и для фраз «хороший автомобиль» и «хорошая машина». В [8] сравниваются способы определения правильности извлеченной фразы:

- точное совпадение двух фраз (exact);
- совпадение двух фраз при наличии только морфологического различия (morph);
- совпадение двух фраз, если фраза k содержит в себе фразу g (include);
- совпадение двух фраз, если фраза g содержит в себе фразу k (part-of).

Показано, что использование part-of является неудачным. Для оценки совпадения двух фраз в [8] использовались exact и include+morph. В работе [9] расширено число способов оценки извлеченной ключевой фразы: BLEU, METEOR, NIST, ROUGE. Определено отношение, лучше отражающее подход экспертов к определению корректности фразы:

$$rp = \frac{|\, слова\_фразы \in G_t \cap слова\_фразы \in C_t \,|}{\max\{|\, слова\_фразы \in G_t \,|, |\, слова\_фразы \in C_t \,|\}}.$$

В данной работе для оценки качества результатов аннотирования используются метод и мера, предложенные в работах [8, 9], что позволяет в дальнейшем использовать результаты, полученные в данной работе, для сравнительной оценки качества новых алгоритмов, также опирающихся на [8, 9].

Для оценки качества аннотирования используется R-Precision [8]. Для оценки качества каждой конкретной извлеченной фразы рассматриваются три способа.

1. Случай exact. Здесь
   $|\, G_t \cap C_t \,| = |\, \{k \in G_t : \exists g \in C_t \wedge exact(k,g) = 1\} \,|$;
   где $exact(k,g) = 1$, если фразы k и g совпадают, $exact(k,g) = 0$ – в противном случае.
2. Случай include. Здесь
   $|\, G_t \cap C_t \,| = |\, \{k \in G_t : \exists g \in C_t \wedge include(k,g) = 1\}$;
   где $include(k,g) = 1$, если фраза k содержит в себе фразу g, иначе $include(k,g) = 0$.
3. Случай гр [9]. Здесь
   $$|\, G_t \cap C_t \,| = \frac{1}{|\, G_t \,|} \sum_{k \in G_t} \max_{g \in C_t} \left(\frac{|\, k \cap g \,|}{\max(|\, k \,|, |\, g \,|)} \right).$$

Описание эксперимента

Для формирования исходных данных выполнена предобработка текстов. Каждый текст t представлен набором слов, из которого изъяты стоп-слова и все другие слова, кроме существительных и прилагательных [4]. Для определения части речи слов использован Stanford POS tagger tool [12]. Для каждого текста t построены фразы-претенденты: несколько слов из множества b_t объединялись в одну фразу, если в t они непосредственно следовали друг за другом. В построенной фразе-претенденте не могло быть больше четырех слов [8]. В работе поставлена следующая задача: проанализировать способы ранжирования полученных фраз-претендентов, для отбора лучших n из них как ключевых фраз, где $n = |\, C_t \,|$.

Примем следующие обозначения: $\mu(k)$ – вес фразы k, где $k = \{s_1, s_2, ...s_n\}$; s_i или s_i^k – слова фразы k; $TR(s_i^k)$ – вес слова s_i^k, рассчитанный с использованием формулы (1). Для ранжирования фраз на основе TextRank и *tf-idf* вес фразы рассчитывался по формулам

$$\mu(k) = \max_{s_i \in k} TR(s_i), \quad \mu(k) = \max_{s_i \in k} \left(tf - ifd\right)(s_i).$$

Для расчета $TR(s_i^k)$ требуется определить способ оценки веса дуг w_{ij}, соединяющий вершины-слова v_i и v_j. Рассмотрены варианты:

1. $w_{ij} = occur(v_i, v_j)$, где $occur(v_i, v_j)$ – число раз, когда слова v_i и v_j встретились вместе в тексте t в окне $n = 2$ (в работе [4] было показано, что такой размер окна является наилучшим);
2. $w_{ij} = mi(v_i, v_j)$, где $mi(v_i, v_j) = \log \frac{p(v_i, v_j) \cdot N}{p(v_i) \cdot p(v_j)}$ есть взаимная информация между v_i и v_j, $p(v_i, v_j)$ – число раз, когда слова v_i и v_j встретились вместе в текстах коллекции в окне 2, $p(v_i)$, $p(v_j)$ – число вхождений слов v_i и v_j в текстах коллекции, N – общее число словоформ в коллекции;
3. $w_{ij} = \left(tf - idf\right)(v_i) \cdot \left(tf - idf\right)(v_j)$.

Также поставлен дополнительный эксперимент: для ранжирования на основе TextRank строился граф, вершинами которого являлись все слова коллекции после предобработки, а $w_{ij} = p(v_i, v_j)$. Отличие от рассмотренного выше способа расчета весов вершин на основе формулы (1) – в том, что используется один граф для всей коллекции, а не отдельные графы для каждого текста.

Фразы ранжировались по значению весов. Лучшие n фраз с наибольшими значениями весов отбирались как ключевые фразы. На завершающей стадии эксперимента проводилась постобработка. Если для одного текста в качестве претендентов были получены две фразы, такие, что одна является частью другой, то оставлялась одна фраза наибольшего размера.

Результаты экспериментов и обсуждение

Результаты экспериментов представлены для трех подколлекций INSPEC dataset: test dataset, evaluate dataset, trial dataset (табл. 1–3). Приняты следующие обозначения:

1. ранжирование фраз с помощью TextRank:
   TR for text – граф строился для каждого отдельного текста, $w_{ij} = occur(v_i, v_j)$;
   TR for text*mi – граф строился для каждого отдельного текста, $w_{ij} = mi(v_i, v_j)$;
   TR for text**tf-idf – граф строился для каждого отдельного текста, $w_{ij} = (tf - idf)(v_i) \cdot (tf - idf)(v_j)$;
   TR collection – граф строился для всей коллекции, $w_{ij} = p(v_i, v_j)$;
2. результаты ранжирования с помощью *tf-idf*.

INSPEC dataset	TR for text	TR for text*mi	TR for text**(*tf-idf*)	TR collection	*tf-idf*
test dataset	0,22	0,21	0,23	0,19	0,28
evaluate dataset	0,21	0,20	0,22	0,17	0,26
trial dataset	0,21	0,19	0,22	0,16	0,26

Таблица 1. R-Precision exact

INSPEC dataset	TR for text	TR for text*	TR for text**	TR collection	*tf-idf*
test dataset	0,29	0,27	0,31	0,25	**0,37**
evaluate dataset	0,27	0,26	0,29	0,23	0.34
trial dataset	0,27	0,26	0,29	0,23	0.34

Таблица 2. R-Precision include

INSPEC dataset	TR for text	TR for text*	TR for text**	TR collection	*tf-idf*
test dataset	0,42	0,41	0,43	0,37	0,48
evaluate dataset	0,40	0,39	0,42	0,36	0.46
trial dataset	0,40	0,40	0,42	0,35	0.46

Таблица 3. R-Precision rp

Результаты показали, что ранжирование фраз, основанное на TextRank, уступает ранжированию на основе *tf-idf*. TextRank использует данные, собранные из одного текста. Такая ограниченность плохо сказывается на качестве аннотирования. *Tf-idf* интегрирует информацию о важности слов в отдельном тексте и информацию о распространенности слов в коллекции, позволяя отсеять общеупотребительные для коллекции слова, не потеряв важных для текста терминов. Интересно, что *tf-idf* хорошо работает для коллекций абстрактов, несмотря на то, что документы коллекции является короткими текстами, а используемая коллекция – сравнительно небольшая. Формула *tf-idf* требует сбора статистики по встречаемости слов, для чего обычно требуется большой объем текстовых данных. Вывод: важные слова часто повторяются в конкретной аннотации к научной статье и редки в других аннотациях (по другим темам). Такая закономерность позволяет собрать нужную статистику на небольшом наборе данных и обосновывает использование *tf-idf*.

Использование $w_{ij} = mi(v_i, v_j)$ для TextRank не улучшает качества ранжирования по сравнению с классическим вариантом $w_{ij} = occur(v_i, v_j)$. Некоторое улучшение достигается за счет использования для оценки дуг $w_{ij} = (tf - idf)(v_i) \cdot (tf - idf)(v_j)$. Использование взаимной информации между словами: $mi(v_i, v_j)$, не улучшает качества ранжирования в сравнении с *tf-idf*, несмотря на то, что данные собираются по всей коллекции. При использовании mi оценивается не значимость отдельных слов, а значимость связей между словами. Эта мера имеет тенденцию к завышению значимости связей с/между редкими (случайными) словами. В работе при расчете mi не ставилось ограничение, которое позволило бы рассматривать связи только со словами, встретившимися в коллекции больше, чем l раз (в противном

случае вес связи был бы равен нулю). Полученный результат служит индикатором того, что в аннотациях к публикациям ключевые слова не бывают случайными (одиночными относительно коллекции).

Результаты, полученные в дополнительном эксперименте, могут считаться неудовлетворительными. Вероятно, это связано с тем, что теряется информация об особенностях отдельных текстов.

Еще одно интересное наблюдение касается способа оценки каждой автоматически извлеченной фразы. Рассматривалось три способа: exact, inclide, rp. Существует высокая корреляция между результатами, полученными для каждого из этих способов. Коэффициент корреляции Спирмена рассчитан для пар exact-include, include-rp, exact-rp и составил 0,98–0,99, для расчета коэффициентов использовано 15 наблюдений (Табл. 1 для exact, Табл. 2 для include, Табл. 3 для rp), корреляция значима на уровне 0,001. Была проверена корреляция между результатами, полученными в [8], где использовались такие способы оценки, как exact и include+morph. Получено значение 0,91, корреляция значима на уровне 0,001. Планируется провести дополнительные исследования, отвечающие на вопрос: «насколько высока вероятность того, что рассматриваемые способы могут быть взаимозаменяемыми», что позволит использовать единственный способ оценки качества.

Заключение

В работе представлены результаты серии экспериментов по извлечению ключевых слов на базе коллекции INSPEC datset с использованием Stanford POS tagger tool. Показано, что для данной коллекции *tf-i*df предложенный авторами способ ранжирования дает лучший результат, чем TextRank. Использование *tf-idf* требует сбора статистики по встречаемости слов, для чего обычно требуется большой объем данных. Экспериментально полученный результат показывает пригодность *tf-idf* для обработки небольших коллекций абстрактов. Это означает, что важные для текста слова часто повторяются в конкретном абстракте и редки в других абстрактах, что можно считать обоснованием использования *tf-idf* для небольших коллекций абстрактов. По результатам экспериментов сделан вывод, что в абстрактах ключевые слова имеют вхождение в коллекцию большее, чем 1–2 раза. Показано, что используемые в работе три способа оценки качества аннотирования могут оказаться взаимозаменяемыми.

Работа выполнена в рамках ФЦП «Научные и научно-педагогические кадры инновационной России» поисковые научно-исследовательские работы по лоту шифр «2011-1.2.1-302-031», государственный контракт № 16.740.11.0751.

Литература

1. Frank E., Paynter G.W., Witten I.H., Gutwin C., Nevill-Manning C.G. Domain-specific keyphrase extraction // Proc. of IJCAI. – 1999. – P. 688–673.
2. Turney P. Learning to Extract Keyphrases from Text. – NRC/ERB-1057. – 1999. – February 17. – 43 p.
3. Hulth A. Improved automatic keyword extraction given more linguistic knowledge // Proc. of the Conference on Empirical Methods in Natural Language Processing. – 2003. – P. 216–223.
4. Mihalcea R., Tarau P. TextRank: Bringing order into texts // Proc. of the Conference on Empirical Methods in Natural Language Processing. – 2004. – P. 404–411.
5. Brin S. and Page L. The anatomy of a large-scale hypertextual web search engine // Computer Networks and ISDN Systems. – 1998. – V. 30. – № 1–7. – P. 107–117.
6. Wan Xiaojun and Jianguo Xiao Single document keyphrase extraction using neighborhood knowledge // Proceedings of the 23rd AAAI Conferenceon Artificial Intelligence. – 2008. – P. 855–860.
7. Tsatsaronis G., Varlamis I., Norvag K. SemanticRank: Ranking Keywords and Sentences Using Semantic Graphs // Proc. of the 23rd International Conference on Computational Linguistics. – 2010. – P. 1074–1082.
8. Zesch T., Gurevych I. Approximate Matching for Evaluating Keyphrase Extraction // International Conference RANLP 2009. – Borovets, Bulgaria, 2009. – P. 484–489.
9. Su Nam Kim, Baldwin T., Min-Yen Kan. Evaluating N-gram based Evaluation Metrics for Automatic Keyphrase Extraction // Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). – Beijing, 2010. – P. 572–580.
10. Kazi Saidul Hasan, Vincent Ng. Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art // Proceedings of Coling 2010. – Poster Volume, Beijing, 2010. – P. 365–373.
11. Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval // Cambridge University Press. – 2009. – 581 p.
12. Интернет страница инструмента Standford POS tagging tool [Электронный ресурс]. – Режим доступа: http://nlp.stanford.edu/software/tagger.shtml, свободный. Яз. рус. (дата обращения 09.11.2012).

Попова Светлана Владимировна – Санкт-Петербургский государственный университет, Санкт-Петербургский государственный политехнический университет, ст. преподаватель, svp@list.ru

Ходырев Иван Александрович – ООО «Висмарт», зам. генерального директора по науке, kivan.mih@gmail.com