Система автоматического реферирования новостных сообщений на основе машинного обучения

© Павел Браславский[*], Василий Густелев[**]

[*]Институт машиноведения УрО РАН
[**]Уральский государственный университет
[*]pb@imach.uran.ru, [**]gustelev@gmail.com

Аннотация

В статье описана макетная система автоматического реферирования новостных сообщений на основе машинного обучения. В качестве основного набора данных использует корпус из 1183 документов новостного ресурса «Газета.Ру», в которых выделены информативные предложения. Для построения классификатора используется библиотека LibSVM – реализация метода опорных векторов. Классификация производится на основе набора легко вычислимых признаков. Дополнительно проведена оценка на небольшом корпусе статей из газеты «Коммерсант», которые были размечены вручную. Оценка метода дала удовлетворительные результаты.

1 Введение

Методы автоматического формирования сжатого представления, или реферата (*summary*) текстовых документов разрабатываются с конца 1950-х годов [13]. Однако в современной ситуации быстрого роста объемов свободно доступной информации автоматические методы приобретают особую актуальность. Методы автоматического реферирования существенно различаются по типам объектов реферирования (универсальные методы; методы, ориентированные на обработку документов определенной темы и структуры), назначению реферата (так, различают *индикативный*, *информативный* и *критический* рефераты), по используемым данным и техникам. В последнее время особое внимание уделяется реферированию новостных документов, в частности – создание реферата кластера новостных сообщений, относящихся к одному событию, полученного в рамках задачи обнаружения новой темы и слежения за потоком сообщений (*topic detection and tracking, TDT*) [10, 15]. Однако задача реферирования отдельного документа остается по-прежнему актуальной и, возможно, более сложной, чем реферирование набора документов [14].

Большинство существующих систем автоматического реферирования ориентированы на извлечение фрагментов (обычно – предложений, отсюда общее обозначение подхода – *sentence extraction*) исходного текста, из которых и составляется реферат. Предложения извлекаются из оригинального документа на основе комбинации статистических и лингвистических признаков, признаков, базирующихся на словарях, информации о структуре документа и др. Набор извлеченных предложений может быть объектом дальнейшей обработки – например, сокращения (выделяются наиболее информативные фрагменты предложений) или сглаживания (для получения более удобочитаемого связного реферата).

В данной статье мы описываем экспериментальную систему автоматического реферирования новостных сообщений на основе машинного обучения. Основные элементы нашего подхода: выделение значимых предложений (без последующей обработки), использование машинного обучения и «готового» набора данных (т.е. данные не были специально подготовлены для решения задачи автоматического реферирования), использование расширенного набора признаков.

В нашей работе мы преследовали несколько целей:

- исследовать эффективность подхода с использованием машинного обучения по сравнению с более традиционным подходами (например, [2]);
- провести эксперименты с корпусом большого размера по сравнению с предыдущими работами;
- использовать дополнительные признаки, соответствующие синтаксическому уровню (подходы к использованию синтаксической информации в задачах автоматического реферирования излагались в [6]);
- разработать экспериментальную систему реферирования русскоязычных новостей.

Обзор различных подходов к автоматическому реферированию можно найти в нашей работе [2], см. также детальный обзор систем автоматического

Труды 9[ой] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль-Залесский, Россия, 2007.

реферирования [5]. Хорошее представление о состоянии исследований и разработок в области автоматического реферирования дают материалы ежегодной конференции *Document Understanding Conference* [10], в рамках которой в течение двух лет (2001 и 2002) проводилась сравнительная оценка методов реферирования отдельного документа на стандартном наборе данных. Здесь мы хотели бы сделать краткий обзор двух работ, в которых задача автоматического реферирования решается методами машинного обучения.

Пионерской работой, в которой задача выделения предложений была сформулирована как задача автоматической классификации на основе машинного обучения, была статья [12]. В работе используется относительно небольшой набор признаков, который был опробован в более ранних работах: длина предложения, «сигнальные фразы», положение предложения в абзаце, наличие в предложении частотных (для этого документа) слов, присутствие слов, набранных в верхнем регистре. В качестве метода использовался байесовский классификатор. Корпус для экспериментов состоял из 188 научно-технических статей на английском языке, для которых профессиональными библиографами были составлены рефераты. Так как предложения из составленных вручную рефератов могли отличаться от предложений оригинального текста, проводилась дополнительная процедура нахождения соответствия между предложениями реферата и оригинала. Результат оценки методом перекрестной проверки (*cross validation*): при формировании рефератов, совпадающих по длине с ручными рефератами (в среднем три предложения), доля совпадений с сопоставимыми предложениями рефератов-образцов составила 42%.

Работа [11] отстоит от предыдущей работы на семь лет и хотя в целом близка ей по подходу, хорошо отражает прогресс в области методов машинного обучения и автоматического реферирования. В работе используется больший набор признаков, а также более совершенный метод классификации – метод опорных векторов (*support vector machine*). В качестве корпуса используются 180 газетных статей корпуса *Text Summarization Challenge* на японском языке. Для каждого документа в корпусе редакторами выделены наборы наиболее важных предложений, соответствующие 10%, 30% и 50% объема исходного текста (по количеству предложений). В работе используется достаточно широкий набор признаков: позиция предложения, длина предложения, вес предложения (сумма модифицированных весов TF*IDF всех слов предложения), а также веса предыдущего и последующего предложений, плотность ключевых слов документа, наличие в предложении имен собственных определенных типов, присутствие определенных союзов и других функциональных слов, учитываются части речи, «семантическая глубина» существительных (определяется по словарю), жанр документа (каждому документу в корпусе приписан один из четырех жанров), наличие специальных символов, прямая речь, утвердительные высказывания. Пятикратная перекрестная проверка дала такой результат: 46,2% совпадающих предложений при объеме реферата 10% от исходного документа и 51,6% – при объеме 30%.

2 Данные, инструменты и методы

2.1 Корпус «Газета.Ру»

В качестве корпуса мы использовали статьи новостного ресурса Газета.ру (http://www.gazeta.ru). Особенность этого ресурса состоит в том, что в каждой статье выделены «ключевые фрагменты» (рис. 1), которые можно рассматривать как квазиреферат документа. Мы автоматически скачали около 1200 документов (вариант «для печати») из разделов «Бизнес» и «Политика» этого новостного ресурса. Документы были переведены в формат *plain text*, произведена разбивка на предложения, помечены заголовок и предложения, относящиеся к квазиреферату. Корпус был разделен на обучающую (ОВ) и тестовую (ТВ) выборки. Характеристики корпуса приведены в табл. 1.

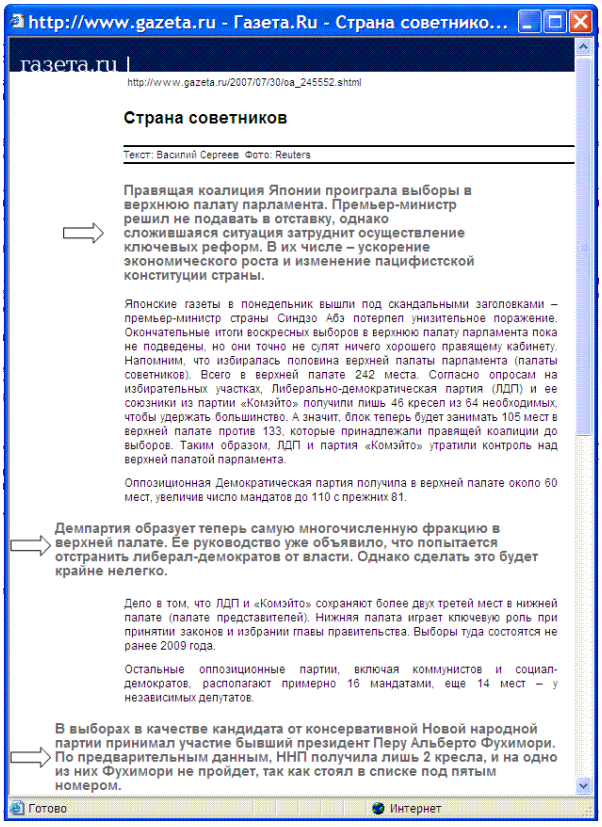

Рис. 1. Статья и ее квазиреферат

Таблица 1. Характеристики корпуса «Газета.Ру»

	Раздел	Документы	Предложения	Квазиреферат
ОВ	Бизнес	294	9534	1870 / 19,6%
	Политика	570	21153	4272 / 20,1%
	Всего	864	30687	6142 / 20,0%
ТВ	Бизнес	89	2843	540 / 18,8%
	Политика	230	8753	1692 / 19,3%
	Всего	319	11596	2232 / 19,2%
Всего		1183	42283	8374 / 19,8%

2.2 Метод классификации

Для построения классификатора мы используем пакет LibSVM [8] – реализацию метода опорных векторов (*support vector machine, SVM*).

Метод опорных векторов – это метод бинарной классификации с учителем, основы которого были разработаны Владимиром Вапником еще в 1970-х годах [3]. На основании обучающей выборки – множества объектов, заданных в виде векторов признаков с явным указанием принадлежности к одному из классов (+1, –1), строится гиперплоскость, разделяющая положительные и отрицательные примеры. Гиперплоскость строится таким образом, чтобы максимизировать ширину границы между положительной и отрицательной частями обучающей выборки. Таким образом, в построении гиперплоскости участвуют только *опорные вектора*, т.е. объекты на границе между частями обучающей выборки. В результате решения задачи оптимизации мы получаем классифицирующую функцию вида:

$$g(x) = \sum_{i=1}^{l} \lambda_i c_i K(x_i, x) + b$$, где

λ_i – коэффициент, определяемый в ходе оптимизации,

x_i – опорный вектор,

c_i – метка класса соответствующего опорного вектора,

K – ядерная функция (*kernel function*), которая используется для перехода к нелинейной разделяющей поверхности.

По рекомендации разработчиков LibSVM мы использовали радиальную базисную функцию:

$$K(x_i, x_j) = e^{-\gamma \left\| x_i^T - x_j \right\|^2}$$, где

γ – параметр функции.

Для построения классифицирующей функции мы должны определить два параметра: γ и C (последний определяет толерантность к ошибкам классификации). Для поиска наилучших значений (γ, C) перебираются пары параметров, после чего по каждой паре проводится перекрестная проверка. При первом проходе по сетке шаг перебора выбирается достаточно большим. Затем выделяется регион, в котором были получены наилучшие результаты, и процедура запускается в этом диапазоне с меньшим шагом. Подробнее о процедуре подбора параметров см. [8].

Количество положительных и отрицательных примеров в обучающей выборке (т.е. предложений из квазирефератов и остальной части документов) существенно различается (см. табл. 1), поэтому использование все обучающей выборки «как есть» ведет к получению классифицирующей функции, отдающей предпочтение классу «не из квазиреферата». Для балансировки положительных и отрицательных примеров мы формируем обучающую выборку из положительных примеров и равного числа отрицательных примеров, выбранных случайным образом.

Кроме того, мы проводили процедуру выбора наиболее информативных признаков в соответствии с рекомендациями разработчиков LibSVM [9]. Процедура основана на подсчете значения F для каждого признака:

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2},$$

где

n_+ – количество положительных примеров,

n_- – количество отрицательных примеров,

$\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$, $\bar{x}_i$ – средние значения i-го признака в положительной, отрицательной частях и всей выборке,

$x_{k,i}^{(+)}$, $x_{k,i}^{(-)}$ – значения i-го признака k-го положительного и отрицательного примеров соответственно.

Признаки упорядочиваются по значению F, после этого методом итеративной перекрестной проверки подбирается пороговое значение F. Признаки с меньшим значением отбрасываются.

Для того, чтобы строить реферат задачной длины (т.е. выбирать заданное количество предложений из текста, а не просто разделять все предложения текста на два класса), на этапе классификации мы используем не «классический» подход, когда объект классифицируется на основе знака функции *g(x)*, а используем значение функции *g(x)*, чтобы упорядочить предложения-кандидаты по расстоянию от разделяющей гиперплоскости.

Детали, касающиеся метода опорных векторов, можно найти в [3, 7, 8].

2.3 Признаки

Первоначально для вычисления синтаксических признаков мы планировали использовать библиотеки проекта «АОТ» [1]. К сожалению, из-за технических сложностей в процессе эксперимента мы отказались от использования признаков, основанных на полноценном синтаксическом разборе предложения, заменив их косвенно связанными легко вычислимыми признаками. Для лемматизации использовался модуль морфологического анализа *mystem* [4].

Признаки, которые мы первоначально планировали использовать для обучения и классификации, можно условно разделить на несколько групп. Признаки, которые не были

реализованы в данном эксперименте, помечены в списке ♠. Признаки, которые были отобраны после этапа сокращения набора признаков, помечены ♥.

Тематические признаки

♥Доля топ-10 частотных слов документа (за исключением стоп-слов) в предложении.

♥Сумма весов слов (TF*IDF), входящих в предложение.

♠ Присутствие в предложении имени и фамилии.

♥Наличие слов заголовка в предложении.

Структурные признаки предложения

♥Признак вопросительного предложения.

Признак предложения с прямой речью.

♥Длина предложения, отнесенная к длине самого длинного предложения в тексте.

♥Наличие в предложении подчинительного союза из списка.

♠Синтаксическая сложность предложения: комбинация длин и количества клауз.

Структурные признаки текста

Доля топ-10 частотных слов документа для предыдущего и последующего предложений.

Сумма весов слов (TF*IDF), для предыдущего и последующего предложений.

♥Предложение начинается с личного местоимения третьего лица или с указательного местоимения (простой сигнал анафоры).

♥Положение предложения в тексте.

Формальные признаки

♥Длина предложения в знаках.

♥Количество запятых в предложении.

♥Количество точек в предложении.

♥Количество чисел в предложении.

2.4 Методика оценки

Для первого типа оценки метода мы используем независимую тестовую выборку, ее характеристики приведены в табл. 1. Методика оценки аналогична описанным в [11, 12]. С помощью построенного классификатора мы выделяем из документа столько предложений, сколько было выделено в оригинале, и считаем долю совпавших предложений. В качестве «отправной точки» (*baseline*) мы используем соответствующее количество предложений с начала документа.

Однако использование статей одного издания за относительно небольшой период времени, с более или менее выдержанной редакционной политикой (более того – проводимой в жизнь ограниченным кругом авторов и редакторов) может привести к неявному переобучению (*overfitting*). Для того, чтобы выяснить, насколько метод, полученный на статьях одного издания, годится для обработки других документов того же жанра, мы сформировали дополнительный тестовый корпус из статей онлайновой версии газеты «Коммерсант» (http://www.kommersant.ru) – всего 30 документов, общий объем – 1046 предложений.

Каждый документ из корпуса «Коммерсант» был предъявлен двум экспертам, который выделили заданное количество важных с их точки зрения предложений (20% от общего количества предложений, округленное до целого значения). Каждый эксперт выделил по 231 предложению, в случае 83 предложений мнения экспертов совпали (35,93%). Таким образом, мы получили несколько «образцовых» рефератов: 1) набор выделенных рефератов, соответствующих каждому эксперту, 2) «рефераты согласия» (сюда вошли только те 83 предложения, которые были выделены обоими экспертами), а также 3) «рефераты разнообразия» (379 предложений, которые были отмечены хотя бы одним экспертом). Мы рассматривали все эти наборы как варианты «идеальных рефератов» и проводили по ним оценку.

3 Результаты и обсуждение

Пример результата работы системы приведен на рис. 2 (оригинальный документ находится по адресу http://www.gazeta.ru/2007/08/01/oa_245777.shtml, документ не входит в корпус «Газета.ру», который использовался для обучения и тестирования). Для данного реферата совпадение с квазирефератом оригинала – 60% (три предложения из пяти, помечены *).

Погосян ушел от рутины

- Михаил Погосян ушел с поста гендиректора "Опытно-конструкторского бюро Сухого".*
- Перестановка дает ему возможность сосредоточиться на стратегических вопросах.*
- "В том состоянии, в котором сейчас находится ОАК – постоянные заседания, совещания, невозможно сосредоточиться на других задачах", – говорит начальник аналитического отдела агентства "Авиапорт" Олег Пантелеев.*
- "Работы сейчас очень много и в ОАК, и в компании, - говорит Вадим Разумовский. - Мы сейчас диверсифицируемся с чисто военного на военно-гражданское производство".
- "Менеджер такого высокого уровня должен быть сосредоточен на решении стратегических вопросов", – уверен начальник аналитического отдела агентства "Авиапорт".

Рис. 2. Пример реферата

Таблица 2. Оценка работы метода

	наша система	baseline
Политика	52,83%	52,74%
Бизнес	50,28%	51,57%
Вся ОВ	52,12%	51,97%
Эксперт 1	31,16%	35,90%
Эксперт 2	51,94%	64,53%
Согласие	54,21%	74,70%
Разнообразие	36,41%	64,91%

Оценка результатов по описанным выше методикам приведена в табл. 2. Как видно из таблицы, на тестовой выборке корпуса «Газета.Ру» наша система примерно в половине случаев выбирала те же предложения, которые были в оригинальном квазиреферате. Этот результат очень близок отправной точке (*baseline*). Мы считаем это хорошим результатом. Во-первых, эти значения несколько выше, чем в экспериментах с аналогичной методикой оценки [11, 12]. Во-вторых, в рамках задачи реферирования отдельного документа DUC аналогичная отправная точка (первые 100 слов документа) не была превзойдена ни одной системой (при несколько иной методике оценки) [14]. Признается, что в случае реферирования новостных сообщений начало документа – это очень хорошая стратегия в силу особенностей организации новостного текста (например, для научных статей выводы и заключение играют не менее важную роль, чем введение). Тем не менее, этот факт не делает задачу автоматического реферирования бессмысленной – люди справляются с задачей реферирования значительно лучше «отправной точки» (особенно это верно по отношению не к отдельным людям, а к коллективным результатам группы людей). Известно, что при ручном извлечении рефератов эксперты демонстрируют низкое согласие (см., например, [6, 15]), что подтверждают и наши данные. Разрыв между «отправной точкой» и нашей системой на данных корпуса «Коммерсант» существенно больше и не в пользу нашей системы. Вероятно, полученная модель все же учитывает какие-то частности, присущие только этому изданию. К тому же стоит заметить, что выделенные предложения в Газете.ру – не полноценный реферат. Как нам показалось, иногда выделение предложений служит просто задаче упрощения *визуального* восприятия текста, делит длинный текст на фрагменты. При этом хорошим результатом можно считать то, что наша система не ухудшила свои показатели на рефератах второго эксперта и на «рефератах согласия».

Заключение

Мы реализовали макетную систему автоматического реферирования новостных документов. Оценка системы продемонстрировала приемлемое качество, сравнимое с качеством аналогичных систем, реализованных для других языков.

В дальнейшем мы планируем улучшить качество макета как формальными методами (подбор параметров модели, отбор признаков), так и содержательными – за счет расширение набора признаков.

Мы хотим поблагодарить экспертов, которые провели обработку статей корпуса «Коммерсант».

Литература

[1] Автоматическая обработка текста, http://www.aot.ru/

[2] Браславский П., Колычев И. eXtragon: экспериментальная система для автоматического реферирования веб-документов. *Труды РОМИП-2005*. СПб., 2005. С. 40-53.

[3] Вапник В.Н. Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979.

[4] Парсер *mystem*, http://company.yandex.ru/technology/products/mystem/mystem.xml

[5] Alonso L., Castellón I., Climent S., Fuentes M., Padró L., Rodríguez H. Approaches to Text Summarization: Questions and Answers. In *Revista Iberoamericana de Inteligencia Artificial*, No. 20, pp. 34-52, 2003.

[6] Braslavski P., Yampolska N. Syntax: Should It Be Counted in Automatic Summarisation Tasks? Presented at the *6th International Tbilisi Symposium on Language, Logic and Computation*. Batumi, Georgia, September 12-16, 2005. Available online: http://kansas.ru/pb/presentation/batumi2005.pdf

[7] Burges Ch. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. In *Data Min. Knowl. Discov*. Vol. 2(2), pp. 121-167, 1998.

[8] Chang Ch., Lin Ch. LIBSVM : A Library For Support Vector Machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[9] Chen Y.-W., Lin C.-J. Combining SVMs with various feature selection strategies. In *Feature extraction, foundations and applications* / I. Guyon, S. Gunn et al.(Eds.). Springer, 2006, pp. 315-324. Available online: http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf

[10] Document Understanding Conferences, http://duc.nist.gov

[11] Hirao T., Isozaki H., Maeda E., Matsumoto Y. Extracting Important Sentences with Support Vector Machines. In *Proc. of the 19th International Conference on Computational Linguistics*, vol. 1, pp. 1–7, 2002.

[12] Kupiec J., Pedersen J., Chen F. A Trainable Document Summarizer. In *Proc. of SIGIR'95*, pp. 68–73.

[13] Luhn H. The automatic creation of literature abstracts. In IBM Journal of Research and Development, Vol. 2(2), pp.159–165, 1958.

[14] Nenkova A. Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference. In *Proc. of AAAI 2005*, pp. 1436-1441.

[15] Radev D.R., Jing H., Stys M., Tam D. Centroid-based summarization of multiple documents. In *Information Processing and Management*, vol. 40, pp. 919–938, 2004.

News Summarization System Based On Machine Learning Approach

Pavel Braslavski, Vasilii Gustelev

The paper describes an experimental automatic summarization system for news stories based on machine learning approach. As a main dataset we use a corpus of 1183 news stories from Gazeta.ru, a popular Russian online news service. The news stories have highlighted sentences that are used as summary. For classifier building we use LibSVM – an implementation of support vector machine. We use a set of easily computable features for classification. Additionally, we performed evaluation on a smaller manually tagged Kommersant corpus. Evaluation shows acceptable quality of the results.