Ягунова Е.В., Макарова О.Е., Антонова А.Ю., Соловьев А.Н. Разные методы компрессии в исследовании понимания новостного текста // Понимание в коммуникации: Человек в информационном пространстве: сб. научных трудов. В 3 тт. – Ярославль – Москва: Изд-во ЯГПУ 2012. Т.2. С. 414-422

Ягунова Е.В., Макарова О.Е., Антонова А.Ю., Соловьев А.Н. (СПбГУ, С.-Петербург; I-Teco, Москва
iagounova.elena@gmail.com, makarova.olga.e@gmail.com, a.antonova@i-teco.ru, a.solovyev@i-teco.ru

РАЗНЫЕ МЕТОДЫ КОМПРЕССИИ В ИССЛЕДОВАНИИ ПОНИМАНИЯ НОВОСТНОГО ТЕКСТА

Elena Yagunova, Olga Makarova, Anna Antonova, Alexej Solov'ev
(St.-Petersburg State University, I-Teco, Moscow)

VARIOUS SUMMARIZATION METHODS AND THE INVESTIGATION OF NEWS TEXTS UNDERSTANDING

Text understanding, perception, news texts, summarization, compression, text structure, keywords, sentiment analysis, information

Исследование понимания новостного текста (соотношения роли информационной и эмоциональной составляющей) строится на базе разных методов компрессии текста (от разных вариантов квазиреферирования до наборов ключевых слов).

The investigation on the news text understanding (with its informational and emotional aspects) is based on the different text compression approaches (varying from automatic summarization to key words extraction).

0. Введение

В данной работе мы рассматриваем некоторые аспекты понимания текста на примере текстов новостных лент как имеющих наиболее простую структуру. Что оказывается в фокусе внимания читающего новостную ленту? Можем ли мы попробовать оценить вклад в понимание тех процедур, которые основываются преимущественно на информационной составляющей и преимущественно на эмоциональной составляющей?

Далее мы будем сопоставлять результаты экспериментов по различным вариантам компрессии текста. Что такое компрессия? Компрессия – это то, что принято соотносить с процессом/результатом понимания текста, когда в результате понимания остается сжатый определенным образом текст (т.е. осталось только то, что оказалось в фокусе внимания). Способ сжатия

определяется задачами компрессии (субъективной и объективной, см. (Леонтьева 2006)). Таким образом, компрессия текста зависит от методов компрессии и целей адресата, вариативность компрессии неизбежна. Традиционным результатом компрессии является реферат. В данной работе мы рассматриваем автоматические рефераты, полученные методами квазиреферирования, основанные на информационной составляющей (количество информации и TF*iDF) или эмоциональной составляющей предложения (сентимент-анализ уровня (под)предложения)[1]. Именно эти два направления, условно противопоставляющие собственно информационную и эмоциональную составляющие, наиболее интересны для нас.

Еще одним вариантом компрессии текста является сворачивание его до набора ключевых слов (КС), выделяемых автоматически или вручную в ходе эксперимента с информантами. В рамках данной работы этот вид компрессии использовался для оценивания качества реферата, поэтому мы решили ограничиться только экспериментом с информантами.

1. Материал. Методика

В качестве материала для исследования мы взяли новостные тексты сравнительно простой структуры, относящиеся к типу (жанру) новостная лента. Один и тот же материал использовался для проведения двух различных независимых экспериментов.

Из четырех новостных текстов (см. п. 2) вручную выделялись КС. Для этого использовалась традиционная методика проведения эксперимента с информантами со стандартной инструкцией, предложенной А.С. Штерн (Мурзин, Штерн 1991): «*Прочитайте текст. Подумайте над его содержанием. Выпишите 10-15 слов, наиболее важных с точки зрения его смысла»*. В эксперименте принял участие 31 информант.

Ранее (Solov’ev et al 2012) был проведен эксперимент по «интегральной» оценке рефератов (19 информантов без специальных экспертных навыков

[1] См. статью «Using sentiment-analysis for text information extraction» в (Solov'ev A. N..et al. 2012). В нашей работе использована часть методов квазиреферирования, которые в максимальной степени отвечают нашим задачам.

анализа рефератов). Информантам предлагалось прочитать исходный текст и оценить по трехбалльной шкале (без дополнительных пояснений):

- **точность автореферата:** «1» – хорошая точность (кратко содержит значимую информацию); «2» – удовлетворительная точность (содержит только часть информации); «3» – плохая точность (не отражает исходную информацию);
- **избыточность** автореферата: «1» – нет избыточности, «2» – есть небольшая избыточность и «3» – есть большая избыточность.

Итак, что лежит в основе понимания текста - скорее информационная или скорее эмоциональная составляющая? Какие рефераты лучше отражают процедуры понимания текста (являются «хорошими» рефератами)?

Рассмотрим две гипотезы оценки рефератов с помощью набора КС:

1) точность реферата по оценке эксперта коррелирует с иерархией КС,

2) в «хороших» рефератах представлены те темы, которые соответствуют КС с максимальными весами.

Дополнительным направлением является классификация текстов СМИ, в основу которой положено согласованность всех рассматриваемых параметров[2].

2. Материал и результаты. Обсуждение результатов

Работа с КС как сверткой текста, объективирующей значимость для информантов того или иного слова (и, соответственно, той или иной составляющей текста), описана в ряде публикаций (напр., (Ягунова 2008; Ягунова 2012)). В эксперименте с информантами анализировались т.н. «пробельные слова» (от пробела до пробела или знака препинания), поэтому возможны были разрывы сложных номинаций, когда компоненты сложных номинаций имели разные значения меры (выделялись разным количеством информантов).

[2] В силу ограничений на объем в эту статью мы не включаем данные по классификации текстов (хотя основной принцип такой классификации просматривается уже в ходе анализа тех четырех текстов, что рассматриваются как примеры). В рамках исследования классификации будет рассматриваться оба типа КС: как КС, выделяемые информантами, так и КС, выделяемые на основании меры TF*iDF (с разными контрастивными коллекциями, см., напр., (Salton, Buckley 1988)).

Таблица 1. Наборы КС для текстов 1-4

Исходный текст	**Набор КС (по уменьшению веса «ключевости»)**
(1) В Москве сорван план хакеров по краже денег из 96 банков	Хищения (27), хакеры (26), ключи (23), паролями (19), обеспечения (17), банках (16), предотвратили (16), программного (16), хранения (15), порядка (14), денег (13), несвоевременным обновлением (12), вредоносного (11), завладела (10), электронные (10)
(2) Паника в Бангкоке: жители бегут из города	Наводнение (31), Бангкок (26), Паттайю (19), затопила (15), паника (15), прекратить (15), закончилась (13), поездки (13), вода (питьевая) (12), уехать (12), выходные (11), пик (11), еда (10), оверквотинг (9)
(3) В Петербурге сотрудник ФСБ застрелил прохожего	Расстрелял (29), прохожего (28), ФСБ (28), скончался (23), сотрудник (23), депутат (18), в Петербурге (14), задержаны (12), Саратове (12), травматического (11), инцидент (10), пострадавший (10)
(4) Умер известный композитор-песенник Георгий Мовсесян	Скончался (28), инфаркт (27), Мовсесян (27), композитор (24), артист (23), Георгий (19), народный (17), автором (13), песням (11), 7 ноября (10), искусств (8), в Москве (8), Робертом Рождественским (8)

Приведем пример разметки исходных текстов статей, в которых отражены результаты квазиреферирования и выделения КС, т.к. без конкретного исходного материала сложно оценить полученные результаты. Условные обозначения: ***п/ж курсив*** – темы, отраженные в реферате и словах; ~~зачеркнутый~~ – темы, не отраженные в рефератах; *подчеркнутый курсив* – темы, неважные для информантов, но отраженные в рефератах.

Текст 2. Паника в Бангкоке: жители бегут из города

МИД России призвал отечественных путешественников ***прекратить поездки в Бангкок из-за сильного наводнения***. Тем временем столицу Таиланда ~~охватила паника~~ - тысячи туристов и жителей города пытаются одновременно покинуть его и атакуют железнодорожные и автовокзалы, а также единственный доступный аэропорт, сообщает Chinadaily.

Главная река Бангкока - Чао Прайя затопила центральные части города и территорию Королевского дворца, нетронутым остается пока лишь деловой центр города. Впрочем, большинство учреждений столицы вынуждено было закрыться раньше - сотрудники не могли добраться в офисы. ***Сейчас в Бангкоке объявлены длинные выходные, чтобы жители могли уехать из города***. *На ближайшие выходные, по данным экологов, в Бангкоке придется пик наводнения.* Впрочем, вернуться в столицу и после выходных не удастся - город еще месяц будет стоять под водой.

Часть жителей не собирается покидать Бангкок во время наводнения, они опасаются мародеров. Но большинство спешит ***поскорее уехать***, ~~в магазинах давно закончилась еда и бутылированная питьевая вода~~. *Тем временем, частные учреждения и государственные компании тайской столицы перебираются в курортные города****, в том числе и в Паттайю.***

"Мы откроем в Паттайе временный центр и переведем туда значительную часть сотрудников, если ситуация с наводнением резко ухудшится", - заявил президент тайской фондовой биржи. В Бангкоке проблемы с работой биржи заключаются в том, что уже сейчас 10% сотрудников не могут добраться до офиса. *А к выходным экологи обещали, что уровень воды поднимется и столицу затопит окончательно.*

По информации авиакомпании Thai Airways, главный авиаперевозчик Таиланда собирается также переносить свой офис в Паттайю и пытается забронировать 300 номеров в отелях города для своих сотрудников сразу на месяц вперед.

Многие жители Бангкока уже уехали в Паттайю**, город находится всего в 140 километрах от столицы, и улицы его наводнили автомобили с бангкокскими номерами. **Для российских туристов такое нашествие может обернуться "оверквотингом" в гостиницах.

Текст 3. В Петербурге сотрудник ФСБ застрелил прохожего

В Петербурге сотрудник ФСБ расстрелял из травматического пистолета случайного прохожего. Скандальный инцидент произошел на Среднеохтинском проспекте в минувшую пятницу, однако известно об этом стало только сейчас, сообщает РБК-Петербург.

По предварительной информации, двое сотрудников Федеральной службы безопасности поссорились с двумя прохожими. *Между ними завязалась словесная перепалка, и тогда один из силовиков выхватил из кобуры травматический пистолет "Оса" и открыл огонь.*

Пострадавший мужчина был госпитализирован и спустя некоторое время скончался в больнице. Сотрудники ФСБ вскоре были задержаны. ФСБ в настоящий момент от комментариев отказывается, однако, по неофициальным данным, по факту инцидента проводится всесторонняя проверка, отмечает издание. *В Следственном комитете по Петербургу РБК сообщили, что в их ведомстве это дело не проходит.*

~~Похожий инцидент произошел в августе 2011г. в Саратове,~~ только тогда уже сотрудники ФСб выступили потерпевшими по делу. ***Депутат Саратовской областной думы от "Единой России" Леонид Писной открыл стрельбу в центре города***. *Как позднее пояснил сам избранник народа, он открыл огонь "в целях безопасности", однако в правоохранительных органах уверены, что он стрелял по сотрудникам полиции, проводившим операцию.*

Выяснилось, что вечером 16 августа улицу Вольскую, по которой следовала депутатская "Волга", перекрыли три автомобиля ВАЗ-2109. *В правоохранительных органах заявили, что в это время ГУ МВД РФ по Саратовской области совместно с сотрудниками регионального управления ФСБ проводило оперативно-розыскные мероприятия. Около 21:45 силовики перекрыли ул.Вольскую.* КамАЗ следовал в составе колонны, а дорога была закрыта, чтобы избежать ДТП при развороте грузовика.

По словам полицейских, Л.Писной вышел из "Волги" и в грубой форме потребовал убрать автомобили с дороги. К нему подошел одетый в гражданскую форму сотрудник МВД, предъявил удостоверение и объяснил, что через две минуты проезд будет свободен. Однако депутат достал пистолет, выстрелил в лобовое стекло "девятки" со стороны водителя, сел в свою машину и уехал.

Таблица 2. Распределение КС в компрессированных текстах[3]

(1) В Москве сорван план хакеров по краже денег из 96 банков
В Москве сотрудники управления экономической безопасности столичного ГУВД **предотвратили** многомиллионные **хищения денег** со счетов 457 компаний в 96 **банках.** Об этом сообщили в пресс-службе управления. Преступники похитили **электронные ключи** 457 компаний. Банковские **ключи**, логины и **пароли** 457 клиентов были похищены **хакерами** с помощью **вредоносного программного обеспечения.**
В Москве сотрудники управления экономической безопасности столичного ГУВД **предотвратили** многомиллионные **хищения денег** со счетов 457 компаний в 96 **банках.** Сотрудники милиции узнали о том, что преступная группа, в которую входят профессиональные **хакеры, завладела паролями** к доступу управления счетами клиентов в 96 российских и зарубежных **банках.** Преступники похитили **электронные ключи** 457 компаний.
В Москве сотрудники управления экономической безопасности столичного ГУВД **предотвратили** многомиллионные **хищения денег** со счетов 457 компаний в 96 **банках.** Преступники похитили **электронные ключи** 457 компаний. Банковские **ключи**, логины и **пароли** 457 клиентов были похищены **хакерами** с помощью **вредоносного программного обеспечения.**
(2) Паника в Бангкоке: жители бегут из города
МИД России призвал отечественных путешественников **прекратить поездки в Бангкок** из-за сильного **наводнения.** Сейчас в **Бангкоке** объявлены длинные **выходные**, чтобы жители могли **уехать** из города. На ближайшие выходные, по данным экологов, в **Бангкоке** придется **пик наводнения.** А к выходным экологи обещали, что уровень **воды** поднимется и столицу **затопит** окончательно. Для российских туристов такое нашествие может обернуться "**оверквотингом**" в гостиницах.
МИД России призвал отечественных путешественников **прекратить поездки в Бангкок** из-за сильного **наводнения.** Впрочем, большинство учреждений столицы вынуждено было закрыться раньше - сотрудники не могли добраться в офисы. "Мы откроем в **Паттайе** временный центр и переведем туда значительную часть сотрудников, если ситуация с **наводнением** резко ухудшится ",- заявил президент тайской фондовой биржи. Для российских туристов такое нашествие может обернуться " **оверквотингом** " в гостиницах.
Главная река **Бангкока** Чао Прайя **затопила** центральные части города и территорию Королевского дворца, нетронутым остается пока лишь деловой центр города. Сейчас в **Бангкоке** объявлены длинные выходные, чтобы жители могли уехать из города. На ближайшие выходные, по данным экологов, в **Бангкоке** придется **пик наводнения.** Часть жителей не собирается покидать **Бангкок** во время **наводнения**, они опасаются мародеров. Тем временем, частные учреждения и государственные компании тайской столицы перебираются в курортные города, в том числе и в **Паттайю.**
(3) В Петербурге сотрудник ФСБ застрелил прохожего
В Петербурге сотрудник ФСБ расстрелял из **травматического** пистолета случайного **прохожего.**

[3] Порядок следования методов квазиреферирования (квазирефератов) как в табл. 1.

В Следственном комитете по Петербургу РБК сообщили, что в их ведомстве это дело не проходит. **Депутат** Саратовской областной думы от "Единой России" Леонид Писной открыл стрельбу в центре города. Около 21:45 силовики перекрыли ул.Вольскую. КамАЗ следовал в составе колонны, а дорога была закрыта, чтобы избежать ДТП при развороте грузовика. По словам полицейских, Л.Писной вышел из "Волги" и в грубой форме потребовал убрать автомобили с дороги.
В Петербурге сотрудник ФСБ расстрелял из **травматического** пистолета случайного **прохожего.** **Пострадавший** мужчина был госпитализирован и спустя некоторое время **скончался** в больнице.
В Петербурге сотрудник ФСБ расстрелял из **травматического** пистолета случайного **прохожего.** Между ними завязалась словесная перепалка, и тогда один из силовиков выхватил из кобуры травматический пистолет Оса и открыл огонь. Сотрудники ФСБ вскоре были **задержаны.** **Депутат** Саратовской областной думы от Единой России Леонид Писной открыл стрельбу в центре города. Как позднее пояснил сам избранник народа, он открыл огонь в целях безопасности, однако в правоохранительных органах уверены, что он стрелял по сотрудникам полиции, проводившим операцию. В правоохранительных органах заявили, что в это время ГУ МВД РФ по Саратовской области совместно с сотрудниками регионального управления ФСБ проводило оперативно-разыскные мероприятия.
(4) Умер известный композитор-песенник Георгий Мовсесян
С 1969г. - **артист** Москонцерта, солист и концертмейстер инструментальной мастерской. Работал над **песнями** в соавторстве со знаменитыми поэтами **Робертом Рождественским** и Михаилом Таничем. Создал цикл песен на стихи **Р.Рождественского** "Поговорим".
В Москве вечером **7 ноября** в возрасте 66 лет скоропостижно **скончался** известный **композитор, народный артист** России **Георгий Мовсесян.** Как сообщили РБК в Союзе композиторов России, причиной смерти стал **инфаркт,** врачи скорой не успели помочь **артисту.** В 1995г. Г. **Мовсесяну** было присвоено звание заслуженного деятеля **искусств**, в 2001г.- звание народного **артиста** России.
В Москве вечером **7 ноября** в возрасте 66 лет скоропостижно **скончался** известный **композитор, народный артист** России **Георгий Мовсесян.** Создал цикл песен на стихи **Р.Рождественского** «Поговорим». РБК приносит свои соболезнования родным и близким композитора.

В Таблице 3 приведен пример оценки разных типов компрессии, в графах «точность» и «избыточность» приводится значение медианы (в скобках – среднего) результатов экспериментов по «интегральной» оценке рефератов, в графе Q – доля количества КС в реферате по отношению к объему набора КС, R – доля КС в реферате (с повторами) по отношению к общему количеству слов в реферате, X – количество предложений без КС (в скобках

их объем в пробельных словах). Лучшие результаты по каждому тексту выделены серым фоном.

Таблица 3. Пример оценки компрессии (квазирефератов)

	точность	избыточность	Q	R	X
(1) В Москве сорван план хакеров по краже денег из 96 банков					
Количество информации	1,5 (1,5)	2 (1,8)	**0,67**	0,24	1(6)
Тональность	1,5 (1,6)	2 (1,9)	**0,73**	0,24	0
TF-iDF	1 (1,6)	2 (1,8)	**0,73**	0,30	0
(2) Паника в Бангкоке: жители бегут из города					
Количество информации	2 (1,8)	2 (1,7)	0,57	0,14	0
Тональность	2 (2,1)	3 (2,3)	0,50	0,11	1(15)
TF-iDF	2 (1,8)	2 (1,8)	0,50	0,01	0
(3) В Петербурге сотрудник ФСБ застрелил прохожего					
Количество информации	2 (2,3)	2 (2,3)	0,58	0,09	3(51)
Тональность	2 (1,9)	2 (1,1)	**0,67**	0,35	0
TF-iDF	2 (1,7)	2 (2,3)	**0,67**	0,08	3(67)
(4) Умер известный композитор-песенник Георгий Мовсесян					
Количество информации	3 (2,9)	2,5 (2,2)	0,23	0,13	0
Тональность	1 (1,2)	2 (1,8)	**0,77**	0,25	0
TF-iDF	2 (1,6)	2 (1,9)	**0,69**	0,27	1(8)

Итак, все 4 текста включают очевидные эмоциональные составляющие. Два текста дают лучшую оценку по тональности. Два других – по «информационным мерам»: (количество информации или TF*iDF). Лучше всего опора на тональность работает для текста 3, т.е. именно для него наиболее важна эмоциональная составляющая; хуже всего – для текста 2.

3. Выводы. Заключение

Очевидно, что в основе понимания текста лежит коммуницируемый смысл, который представлен в виде переплетения информационной, и эмоциональной составляющих. Сложное взаимодействие этих составляющих зависит от типа текста. Рефераты текста можно рассматривать как формальным образом полученные варианты вторичных текстов, а автомат – как субъект понимания особого типа (ср. (Новиков 2001; Леонтьева 2006)). Мы предлагаем строить изучение понимания новостного текста следующим образом:

- сопоставление разных методов компрессии – и разных способов квазиреферирования, и выделения КС (желательно разными способами);

- выделения информационной и эмоциональной составляющих на основе результатов компрессии (разных рефератов и наборов КС);
- построения предполагаемой типологии текстов, объясняющей «включение» определенной стратегии анализа (понимания) текста, где стратегия понимания – это соотношение значимости эмоциональных и/или информационных составляющих.

Одним из результатов такого изучения будет формализованная оценка «адекватности» методов компрессии типу текста. Точность реферата по оценке эксперта – сложный и неоднозначный признак, он коррелирует с иерархией КС обычно или для «своего» типа текста (согласно получаемой нами типологической схеме), или для представительной выборки текстов. Более того, по-видимому, эта экспертная оценка может служить косвенным показателем стратегии понимания разных групп экспертов. В обобщенном виде (ср. табл. 1-3), она, по-видимому, представляет неоднозначное взаимодействие разных стратегий оценивания и понимания.

Литература

1. Леонтьева Н.Н. Автоматическое понимание текста: системы, модели, ресурсы: учебное пособие – М.: Издательский центр «Академия», 2006.

2. Мурзин Л. Н., Штерн А. С. Текст и его восприятие.– Свердловск : Изд-во Урал. ун-та, 1991.

3. Ягунова Е.В. Набор опорных слов как вид свёртки текста (в сопоставлении с набором ключевых слов) // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14).– М.: РГГУ, 2008

4. Ягунова Е.В. Как использовать ключевые слова в исследовании произведений Н.В. Гоголя? // "X выездная школа-семинар "Проблемы порождения и восприятия речи": Материалы.- Череповец: ГОУ ВПО "Череповецкий государственный университет", 2011. с. 39-52

5. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 1988. – № 24(5). – P. 513-523.

6. Solov'ev A.N., Antonova A.Ju., Pazel’skaja A.G. Using sentiment-analysis for text information extraction // Компьютерная лингвистика и

интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2012 г.). Вып. 11 (18): В 2 т. Т. 1: Основная программа конференции. — М.: Изд-во РГГУ, 2012.