Брунова Елена Георгиевна

ОСОБЕННОСТИ ПАРАМЕТРИЧЕСКОЙ ЛЕКСИКИ ПРИ КОНТЕНТ-АНАЛИЗЕ МНЕНИЙ

В статье определяются основные особенности параметрической лексики при контент-анализе мнений на материале отзывов клиентов о качестве банковского обслуживания. Предлагается усовершенствованная структура лексикона для контент-анализа мнений. Результаты исследования показывают, что параметрическая лексика выражает мнение имплицитно. Некоторая часть параметрической лексики может быть отнесена к одному из главных классов (положительному или отрицательному лексикону), причем такое отнесение является специфичным для данной предметной области. Большая часть параметрической лексики относится к вспомогательным классам (инкрементам или декрементам), и такое отнесение представляется универсальным.

Адрес статьи: www.gramota.net/materials/2/2014/12-1/9.html

Источник

Филологические науки. Вопросы теории и практики

Тамбов: Грамота, 2014. № 12 (42): в 3-х ч. Ч. I. С. 35-39. ISSN 1997-2911.

Адрес журнала: www.gramota.net/editions/2.html

Содержание данного номера журнала: www.gramota.net/materials/2/2014/12-1/

Список литературы

1. **Гроссман В. С.** Жизнь и судьба // Гроссман В. С. За правое дело. Жизнь и судьба: дилогия. М: Астрель, 2012. 1518 с.
2. **Гроссман В. С.** Треблинка // Черная книга / сост. под ред. В. Гроссмана, И. Эренбурга. К.: Обериг, 1991. С. 391-422.
3. **Гроссман В. С.** Убийство евреев в Бердичеве // Черная книга / сост. под ред. В. Гроссмана, И. Эренбурга. К.: Обериг, 1991. С. 32-43.
4. **История Минского гетто. По материалам: А. Мачиз, Гречаник, Л. Глейзер. П. М. Шапиро** // Черная книга / сост. под ред. В. Гроссмана, И. Эренбурга. К.: Обериг, 1991. С. 152-194.
5. **Палиевский П. В.** Документ в современной литературе // Палиевский П. В. Из выводов XX века. СПб.: Русский остров; Владимир Даль, 2004. С. 376-402.

DOCUMENTARY SOURCES OF THE "BLACK BOOK" IN THE NOVEL BY V. S. GROSSMAN "LIFE AND FATE"

Biryuchin Svyatoslav Vladimirovich
Moscow State Pedagogical University
biryuchin@yandex.ru

The article by means of comparative analysis of the texts examines the functioning of documentary sources of an anthology "Black Book" in the novel by V. S. Grossman "Life and Fate". The author pays special attention to the essays "The Murder of Jews in Berdichev" and "Treblinka" acquired most detailed and accurate artistic representation in the text of the novel.

Key words and phrases: documentary source; fact; author˝s fiction; "Black Book"; "Life and Fate"; V. S. Grossman.

УДК 81'322
Филологические науки

В статье определяются основные особенности параметрической лексики при контент-анализе мнений на материале отзывов клиентов о качестве банковского обслуживания. Предлагается усовершенствованная структура лексикона для контент-анализа мнений. Результаты исследования показывают, что параметрическая лексика выражает мнение имплицитно. Некоторая часть параметрической лексики может быть отнесена к одному из главных классов (положительному или отрицательному лексикону), причем такое отнесение является специфичным для данной предметной области. Большая часть параметрической лексики относится к вспомогательным классам (инкрементам или декрементам), и такое отнесение представляется универсальным.

Ключевые слова и фразы: обработка естественного языка; контент-анализ мнений; оценочный лексикон; предметная область; параметрическая лексика; инкремент; декремент.

Брунова Елена Георгиевна, д. филол. н., доцент
Тюменский государственный университет
egbrunova@mail.ru

ОСОБЕННОСТИ ПАРАМЕТРИЧЕСКОЙ ЛЕКСИКИ ПРИ КОНТЕНТ-АНАЛИЗЕ МНЕНИЙ©

1. Введение

Контент-анализ мнений (англ. *sentiment analysis*) является одной из бурно развивающихся методик автоматической обработки естественного языка. Первые работы были опубликованы в начале 2000-х гг. [11; 13; 14; 15], и с тех пор сделано достаточно много. Созданы оценочные лексиконы, разработаны алгоритмы [5; 8; 9; 10; 12]. Основные исследования в данной области проводились на материале английского языка, и казалось логичным применить их результаты для других естественных языков, перевести лексиконы и модифицировать средства синтаксического анализа. Однако попытки создания универсального оценочного лексикона, основного инструмента контент-анализа мнений, пока не увенчались успехом.

Оценочный лексикон (англ. *sentiment lexicon*) представляет собой множество слов, которые используются для выражения мнений и эмоций в документах контент-анализа мнений (отзывах и т.п.), он обычно состоит из двух классов – положительного и отрицательного лексикона [13]. Со временем стало очевидно, что такой лексикон является специфичным для конкретного языка и для конкретной предметной области.

Проблема специфичности по языку связана с особенностями морфологии и синтаксиса естественных языков, тогда как проблема специфичности по предметной области относится к сфере семантики. Некоторые слова из оценочных лексиконов оказываются специфичными для той или иной предметной области [6, p. 242], например, слово *долгий* относится к положительному лексикону при оценке времени работы аккумулятора (предметная область «смартфон»), однако при оценке затрат времени клиента (предметная область «качество банковского обслуживания») оно относится к отрицательному лексикону. В данной статье такие неоднозначные слова называются параметрической лексикой.

Параметрическая лексика – это слова, обозначающие объем некоторого параметра, специфичного для данной предметной области.

Целью данного исследования является определение основных особенностей параметрической лексики при контент-анализе мнений.

2. Материал и методика исследования

Материалом для исследования послужили отзывы клиентов о качестве банковского обслуживания на русском языке, взятые из Народного рейтинга банков на сайте [16]. Исследуемая предметная область – «качество банковского обслуживания». Для создания оценочного лексикона случайным образом были отобраны 20 текстов отзывов (10 положительных и 10 отрицательных). Из данного контента вручную был построен базовый лексикон (англ. *seed lexicon*) в объеме 100 слов. В дальнейшем данный лексикон был расширен до 700 слов с помощью синонимов, антонимов и технологии оценочной согласованности (англ. *sentiment consistency*) [9]. Технология оценочной согласованности, впервые представленная в [7], использует набор базовых оценочных имен прилагательных и набор ограничителей (*и, но, или-или, ни-ни*) для выявления оценочной лексики и определения ее полярности. Например, в предложении *Этот айфон красивый и легкий*, если заранее известно, что *красивый* относится к положительному лексикону, то подразумевается, что и *легкий* относится к положительному лексикону. И наоборот, в предложении *Этот айфон красивый, но дорогой*, если заранее известно, что *красивый* относится к положительному лексикону, то подразумевается, что *дорогой* относится к отрицательному лексикону.

Наш оценочный лексикон включает два главных класса: положительный и отрицательный лексиконы, т.е. слова, выражающие соответственно положительные и отрицательные мнения. Кроме того, он включает три вспомогательных класса: инкременты, модификаторы и антимодификаторы полярности [1; 2].

Инкрементами называются слова, усиливающие полярность других слов в предложении, при этом полярность не изменяется на противоположную, например, в контекстах *Это очень надежный банк* и *Это очень плохие условия кредита* слово *очень* является инкрементом, усиливающим соответственно положительную и отрицательную оценки.

Модификаторами полярности называются слова, изменяющие полярность других слов в предложении на противоположную, например, в контексте *Сами работники банка не грубые и не злые* имеются слова из отрицательного лексикона *грубые* и *злые*, а слово *не* является модификатором полярности, изменяющим их полярность на положительную.

Антимодификаторами полярности называются слова, которые отменяют изменение полярности, несмотря на наличие модификаторов полярности в предложении. Сравним два контекста: 1) *Меня никогда не обманывали* 2) *Меня никогда так не обманывали*. Несмотря на почти полное совпадение слов, которые в них входят, данные контексты выражают противоположные оценки – соответственно положительную и отрицательную. Разница заключается в том, что в первом случае под словом *никогда* подразумевается *никогда в этом банке*, а во втором – *никогда, кроме этого банка*. Слово *так* является антимодификатором полярности, оно отменяет смену полярности во втором примере, и оценка предложения остается отрицательной, поскольку имеется слово из отрицательного лексикона *обманывали*.

Для проведения контент-анализа мнений использовался алгоритм REGEX [3]. Алгоритм содержит 11 правил формальной грамматики и соответствующие синтаксические модели, которые идентифицируют определенные элементы текста, упрощают предложение и представляют текст в виде формальной модели.

Алгоритм преобразования схемы разметки включает последовательное применение правил замены в соответствии с установленными приоритетами. На определенном шаге алгоритма подсчитывается количество слов из положительного и отрицательного лексиконов, после чего определяется черновая численная оценка полярности мнения каждого предложения. Затем применяется группа правил для корректировки черновой оценки. На выходе алгоритма REGEX производится подсчет полярности текста, нормализованный по количеству слов.

Предлагаемый алгоритм был апробирован в системе SENTIMENTO, реализованной в виде интернет-приложения на базе web-сервера *Apache* [Там же]. Система предусматривает возможность для пользователя подтвердить или опровергнуть ее заключение, с этой целью появляется запрос *Your conclusion* (Ваше заключение) и две кнопки: *Positive* (Положительная оценка) и *Negative* (Отрицательная оценка). После того, как пользователь нажимает одну из кнопок, система проверяет свое заключение на соответствие с заключением пользователя. В случае соответствия документ включается в базу данных. Кроме того, результаты такой проверки используются для расчета эффективности алгоритма.

Эксперименты по контент-анализу мнений, проведенные в системе SENTIMENTO, выявили ряд проблем, связанных с параметрической лексикой. Например, пользователь выставил предложению *Предлагают маленький процент по вкладу* отрицательную оценку, а система выставила 0, т.е. нейтральную оценку, поскольку она не обнаружила слов из отрицательного лексикона. С другой стороны, предложению *Очередь была совсем маленькая* пользователем дается положительная оценка, а система дает отрицательную оценку, поскольку обнаруживает слово из отрицательного лексикона *очередь*.

Таким образом, поведение параметрической лексики в отзывах клиентов отличается от поведения слов из положительного и отрицательного лексиконов, и пренебрежение этим фактом приводит к некорректным результатам контент-анализа мнений.

3. Результаты

Такие слова как *очень*, *совершенно*, *долго*, *медленно* и т.п. демонстрируют свою неоднозначную природу при контент-анализе мнений. Н. В. Лукашевич и И. И. Четверкин предлагают выделять параметрическую лексику в качестве операторов, влияющих на степень оценки [4], однако в их исследовании под операторами

понимаются только отрицательные частицы и лексические усилители прилагательных (*не, нет, полный, очень, самый* и т.п.), а не собственно прилагательные или наречия. Мы полагаем, что прилагательные, наречия и даже существительные (например, *максимум*), выражающие объем того или иного параметра предметной области, необходимо включать в оценочный лексикон.

Возрастание или убывание определенного параметра может вызывать положительные или отрицательные эмоции. Так, *высокий* применительно к скорости, надежности или устойчивости вызывает положительные эмоции, а по отношению к цене или затраченному времени – отрицательные. Именно параметр определяет специфичность такой лексики для предметной области.

Для определения основных особенностей параметрической лексики из корпуса объемом в 70 отзывов о качестве банковского обслуживания, отобранных случайным образом с сайта [16], были выделены контексты слов *большой, маленький, долгий, быстрый, максимум, минимум* и т.п. Изучение данных контекстов позволило определить параметры, специфичные для данной предметной области.

Рассмотрим параметры, имеющие существенное значение для качества банковского обслуживания.

Положительные отзывы

1. Возрастание параметра:

а) положительные эмоции клиента: *хочется отметить оперативность в работе и готовность оказать максимум помощи даже потенциальным клиентам*;

б) экономия средств клиента: *Карта с немалым лимитом*;

в) экономия времени клиента: *Наш кредит одобрили очень быстро*;

г) достаточность информации об услугах: *Много информации, листовки, плакаты с рекламой*.

2. Убывание параметра:

а) отрицательные эмоции клиента: *небольшой список замечаний*;

б) расходы средств клиента: *маленький процент по кредиту*;

в) расходы времени клиента: *Очередь была совсем маленькая*.

Отрицательные отзывы

1. Возрастание параметра:

а) отрицательные эмоции клиента: *хитрости для большого обмана*;

б) расходы средств клиента: *Я и так плачу немалый процент за пользование кредитом*;

в) расходы времени клиента: *Банк для тех, у кого много лишнего времени*.

2. Убывание параметра:

а) положительные эмоции клиента: *толку мало*;

б) экономия средств клиента: *Лимит по кредитной карте маленький*;

в) экономия времени клиента: *платежи проходят медленно*;

г) достаточность информации об услугах: *инфы мало*.

Извлеченные параметры представлены на схеме (см. Рис. 1).

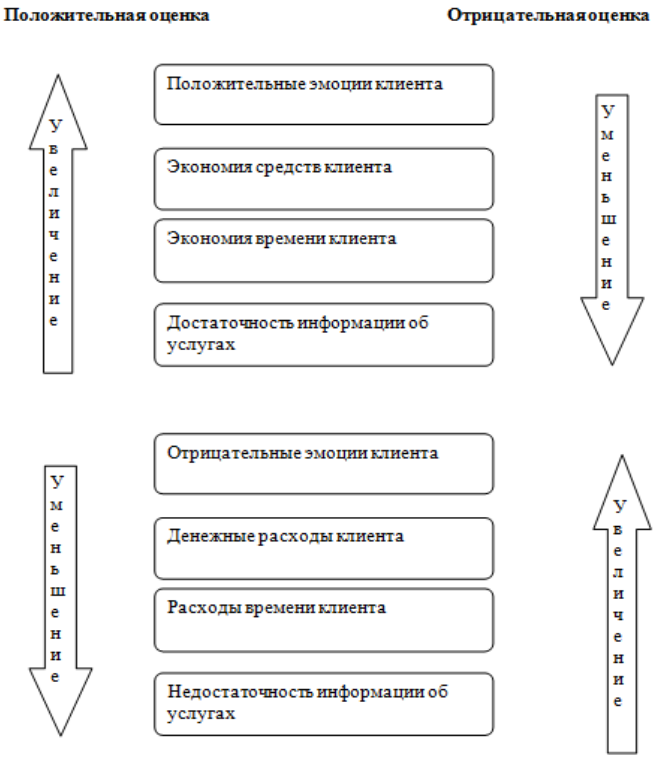

Рис. 1. *Параметры контент-анализа мнений, извлеченные из контекста отзывов о качестве банковского обслуживания*

Как видно из схемы на Рис. 1, если возрастание какого-либо параметра приводит к положительной оценке, то убывание того же самого параметра приводит к отрицательной оценке, и наоборот. Например, возрастание экономии средств клиента вызывает положительные эмоции, а ее убывание – отрицательные. Возрастание расходов клиента вызывает отрицательные эмоции, а убывание – положительные. Таким образом, поведение параметрической лексики не только специфично для данной предметной области, но является неоднозначным даже в пределах одной и той же предметной области. Это подтверждается встречаемостью такой лексики в одном (чаще всего – отрицательном) контексте, ср. *Много слов, но мало дела. Дают быстро, отдают долго. Большой минус и маленький плюс.*

Результаты исследования показывают, что пренебрежение параметрической лексикой приводит к некорректным выводам системы контент-анализа мнений, поэтому такая лексика должна входить в оценочный лексикон. Только небольшая часть параметрической лексики может быть отнесена к одному из главных классов (положительному или отрицательному лексикону), например, *быстро* мы отнесли к положительному лексикону, а *долго* и *медленно* – к отрицательному. Такое отнесение является специфичным, т.е. релевантным только для данной предметной области. Параметрические слова, выражающие возрастание параметра (*большой, много, максимум* и т.п.), следует отнести к классу инкрементов, поскольку они выражают усиление положительных или отрицательных эмоций автора отзыва. Слова, выражающие убывание параметра (*маленький, мало, минимум* и т.п.), следует отнести к новому классу, который мы назвали декрементами. Декременты – это слова, уменьшающие полярность оценочных слов в предложении, при этом полярность не изменяется на противоположную.

Таким образом, большинство параметрических слов относится к вспомогательным классам (инкрементам или декрементам), и такое отнесение представляется универсальным, т.е. релевантным для разных предметных областей.

Усовершенствованная структура оценочного лексикона выглядит следующим образом: два главных класса (положительный и отрицательный лексиконы) и четыре вспомогательных (инкременты, декременты, модификаторы и антимодификаторы полярности).

4. Заключение

В результате исследования определены основные особенности параметрической лексики при контент-анализе мнений, пересмотрена структура оценочного лексикона, добавлен новый класс – декременты полярности.

Поведение большинства параметрических слов в отзывах клиентов отличается от поведения слов из положительного и отрицательного лексиконов. Параметрическая лексика, как правило, выражает мнение имплицитно: она выражает не мнение *per se*, а интенсивность соответствующих эмоций. Сами параметры предметной области, как правило, не называются, однако именно они определяют специфическое поведение параметрических слов для данной предметной области.

Список литературы

1. **Брунова Е. Г.** Методика составления оценочного лексикона для контент-анализа мнений [Электронный ресурс] // Language and Science. 2012. № 1. URL: http://www.utmn.ru/docs/9317.pdf (дата обращения: 08.10.2014).
2. **Брунова Е. Г.** Составление лексикона для контент-анализа мнений // Теоретические и прикладные аспекты изучения речевой деятельности. Н. Новгород: НГЛУ им. Н. А. Добролюбова, 2013. Вып. 1 (8). С. 24-29.
3. **Брунова Е. Г., Бидуля Ю. В.** Алгоритм с элементами формальной грамматики для контент-анализа мнений // Вестник Тюменского государственного университета. Серия «Физико-математические науки. Информатика». 2014. № 7. С. 242-250.
4. **Лукашевич Н. В., Четверкин И. И.** Извлечение и использование оценочных слов в задаче классификации отзывов на три класса // Вычислительные методы и программирование. 2011. Т. 12. С. 73-81.
5. **Gamon M. et al.** Pulse: Mining Customer Opinions from Free Text // Proc. of the 6th International Symposium on Intelligent Data Analysis (IDA). 2005. P. 121-132.
6. **Ganapathibhotla M., Liu B.** Mining Opinions in Comparative Sentences // Proc. of the 22nd International Conference on Computational Linguistics. Manchester, 2008. P. 241-248.
7. **Hatzivassiloglou V., McKeown K.** Predicting the Semantic Orientation of Adjectives // Proc. of the 35th Annual Meeting of ACL. Madrid, 1997. P. 174-181.
8. **Hu M., Liu B.** Mining and Summarizing Customer Reviews // Proc. of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004. P. 168-177.
9. **Liu B.** Sentiment Analysis and Subjectivity [Электронный ресурс] // Handbook of Natural Language Processing: Second Edition. 2010. URL: http://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf (дата обращения: 08.10.2014).
10. **Manning C., Raghavan P., Schütze H.** Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008. 544 p.
11. **Nasukawa T., Yi J.** Sentiment Analysis: Capturing Favorability Using Natural Language Processing // Proc. of the 2nd International Conference on Knowledge Capture. Florida, 2003. P. 70-77.
12. **Pang B., Lee L.** Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval. 2008. Vol. 2. № 1-2. P. 1-135.
13. **Pang B., Lee L., Vaithyanathan S.** Thumbs up? Sentiment Classification Using Machine Learning Techniques [Электронный ресурс] // Proc. of EMNLP. 2002. URL: http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf (дата обращения: 08.10.2014).
14. **Turney P.** Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews // Proc. of the 40th Annual Meeting on Association for Computational Linguistics. 2002. P. 417-424.
15. **Wiebe J., Wilson T., Bell M.** Identifying Collocations for Recognizing Opinions // Proc. of ACL/EACL 01 Workshop on Collocation. 2001. P. 24-31.
16. **www.banki.ru** (дата обращения: 08.10.2014).

FEATURES OF PARAMETRIC VOCABULARY IN CONTENT ANALYSIS OF OPINIONS

Brunova Elena Georgievna, Doctor in Philology, Associate Professor
Tyumen State University
egbrunova@mail.ru

The article determines the main features of parametric vocabulary in a content analysis of opinions by the material of clients‟ reviews about the quality of banking service. An improved structure of the lexicon for the content analysis of opinions is suggested. The research results show that parametric vocabulary expresses the opinion implicitly. Some of the parametric vocabulary may be assigned to one of the main classes (positive or negative lexicon), and this classification is specific to the given subject sphere. Most of the parametric vocabulary refers to the auxiliary classes (increments or decrements), and this reference seems to be universal.

Key words and phrases: natural language processing; content analysis of opinions; evaluative vocabulary; subject sphere; parametric vocabulary; increment; decrement.

УДК 81‟347.78.034
Филологические науки

В статье рассматривается одна из парадоксальных особенностей субтитрированного перевода с японского языка на французский – необходимость сокращать оригинальный текст под влиянием технических требований и одновременно эксплицитно выражать значение ряда языковых единиц в силу значительных структурных отличий языка оригинала и языка перевода. На примере субтитров к фильмам Х. Миядзаки показывается, какие именно трансформации позволяют максимально точно передать содержание, сохраняя коммуникативные интенции говорящих и эмоциональный фон высказываний.

Ключевые слова и фразы: переводоведение; японоведение; переводческие трансформации; субтитрирование; экспликация; сокращение; сопоставительное языкознание; киноискусство.

Бубнова Анна Сергеевна, к. филол. н.
Нижегородский государственный лингвистический университет им. Н. А. Добролюбова
frenjap@yandex.ru

СУБТИТРИРОВАННЫЙ ПЕРЕВОД КАК СИНТЕЗ ЭКСПЛИКАЦИИ И СОКРАЩЕНИЙ (НА ПРИМЕРЕ ФРАНЦУЗСКИХ ВЕРСИЙ ФИЛЬМОВ Х. МИЯДЗАКИ)©

Преимущества субтитров по сравнению с дубляжом очевидны: они не влияют на качество оригинального звука, актѐры говорят «своими» голосами. Всегда можно расслышать малейший скрип двери или приглушѐнный шѐпот таинственного убийцы. Субтитры имеют особое значение для тех, кто изучает иностранные языки, поскольку позволяют обогатить словарный запас вне зависимости от того, насколько понятна оригинальная речь. Субтитры окажут неоценимую помощь всем, кто интересуется сопоставительным языкознанием: звуковая дорожка на одном языке и текст на другом позволят сделать выводы о структурных, лексических и культурно обусловленных различиях двух языков.

Работа по созданию субтитров – процесс чрезвычайно интересный, творческий, но при этом невероятно тяжелый. По словам С. Лакса, «Le grand art de transposer un dialogue parlé en sous-titrage visuel consiste à exprimer le **maximum d'idées** dans la compression avec le **maximum de naturel** dans l‟artifice» [4, p. 6]. / «Искусство перевода устного диалога в субтитры заключается в том, чтобы создать искусственный диалог, максимально сжатый, максимально информативный, максимально приближенный к естественной речи». (*Перевод автора – А. Б.*)

Технически, лаконичность субтитров обусловлена двумя основными пространственно-временными ограничениями: субтитры не могут занимать более двух строк, при этом каждая строка не может содержать более 33 знаков [1, p. 11]. Также необходимо учитывать, что если за одну секунду человек, не напрягаясь, читает 11 знаков, то максимальное время появления субтитров на экране – 6 секунд [Ibidem]. К этому следует добавить, что продолжительность появления субтитров на экране должна совпадать с продолжительностью речи персонажей. Т.е. если персонаж говорит в течение трех секунд, размер субтитров не может превышать 33 знаков.

Не следует забывать и о действии так называемого «визуального шока» [4]. Данный термин подразумевает, что каждый раз при внезапном появлении и исчезновении субтитра глаза зрителя подвергаются достаточно сильному воздействию. Исследователи утверждают, что, поскольку зритель и так достаточно напрягается, одновременно читая субтитры и глядя на изображение, не следует «травмировать» его глаза еще больше излишне частым мельканием субтитров [Ibidem, p. 11]. Из чего следует вывод, что субтитров должно быть как можно меньше. Что же получается в итоге? Субтитры должны максимально передавать содержание фильма, но при этом не появляться слишком часто, совпадать по времени с продолжительностью речи персонажей и содержать строго ограниченное количество знаков.

Сложность при субтитрировании японских фильмов заключается в том, что в силу некоторых характерных особенностей японского языка время звучания реплик весьма непродолжительно. Это создает сложности