Система распознавания тематики устных сообщений

Дорофеев Д.В., *студент ФТЦ НАН Украины*

Чучупал В.Я., *кандидат физико-математических наук, ведущий научный сотрудник ВЦ РАН*

• *распознавание тематики устных сообщений* • *речевая база данных* • *распознавание речи*.

Распознавание слитной речи в телефонном канале связи — достаточно сложная задача. Существующие в настоящее время методы допускают слишком много ошибок, из-за чего практическое их применение невозможно. В данной работе сделана попытка использовать систему автоматического распознавания ключевых слов для классификации телефонных переговоров по заданным темам. Применялся статистический алгоритм классификации, как наиболее устойчивый к шумам. Показано, что даже при очень низком качестве распознавания слов можно добиться приемлемого качества тематической классификации.

• *topic classification* • *topic indexing and summarizing of spoken documents* • *keyword spotting* • *speech recognition*.

The paper describes the experimental work on topic recognition in Russian telephone conversational speech. The probabilistic topic recognition algorithm based on TF-IDF measures has been applied to output of the HMM based keyword spotting system. It has been shown that despite the high keyword error detection rate the topic recognition performance may be quite reliable for practical use.

Введение

Проблема автоматического распознавания и тематической классификации речевых сообщений возникла относительно недавно. Она связана с необходимостью автоматизации процедур анализа массивов текстовой, аудио- и видеоинформации, которые поступают по различным каналам связи и на разных языках. Автоматическая обработка часто является единственным возможным вариантом анализа больших объёмов речевой информации.

Большинство современных научных и прикладных работ по тематической классификации речевых сообщений посвящено обработке текстов. Это вызвано как удельным весом текстов в Интернете, так и недостаточной эффективностью современных технологий распознавания устной речи.

Появление серии работ по тематической классификации устных сообщений в конце прошлого века было обусловлено финансированием со стороны силовых структур США в рамках проектов Национального института стандартов США (NIST) [1], а позднее — программы «Global Autonomous Language Exploitation» (GALE) агентства передовых исследовательских оборонных проектов (DARPA).

Распознавание тематики, как и другие задачи распознавания, можно рассматривать в двух вариантах:
1) идентификация темы разговора как единого целого;
2) обнаружение моментов смены темы.

Поскольку задача тематической классификации устной речи существенно пересекается с соответствующей задачей для текстов, алгоритмы тематической классификации сообщений используют те же подходы, что и алгоритмы классификации текстов.

Отличия классификации устных сообщений заключаются, в частности, в том, что:
- используемый для классификации текст сообщения получен с помощью систем распознавания речи и содержит ошибки распознавания (причём для телефонных систем процент ошибочно распознанных слов может превышать 80%);
- в тексте устного сообщения отсутствуют пунктуационные и орфографические выделения;
- использование методов грамматического анализа литературной речи неэффективно в связи с существенной аграмматичностью разговорной речи.

Анализ литературных источников показывает, что практически все опубликованные алгоритмы используют признаковое представление устного документа, которое основано на понятиях т.н. частот терминов и обратных частот документов или их модификациях[2].

Частота слова определяется как:

$$TF - \frac{\#w_i}{\sum_{j=1}^{N} \#w_j}, \quad (1)$$

где N — число слов в документе, число раз, сколько слово с индексом i встретилось в документе.

Обратная частота документа определяется как:

$$IDF - \frac{\#d : w_i \in d}{\#d}, \quad (2)$$

где — общее число документов в корпусе число документов, содержащих слово w_i .

Оба определения введены специально для автоматической тематической классификации текстов. Чаще всего в качестве признака используют скалярное значение, объединяя два признака в один: «частота слова — обратная частота документа» параметр *TF-IDF*. Если задана текстовая выборка для данной тематики, то после оценки *TF-IDF* для каждого слова из словаря любой документ (необязательно из этой выборки) может быть представлен как вектор параметров.

Таким образом, любой документ можно представить в виде вектора фиксированной размерности, равной размеру словаря. Это означает, что задача распознавания или классификации тематики сводится к постановке классической задачи распознавания образов, для которой уже существует ряд эффективных методов решения, например, на основе байесовских классификаторов, нейронных сетей, машины опорных векторов и т.п.

Кроме рассмотренного выше подхода к классификации тематики на основе использования признаков *TF-IDF* и классических методов распознавания образов в литературе предлагаются и другие подходы к построению алгоритмов, основанные, в значительной сте-

пени, на семантико-синтаксическом и морфологическом анализе текстов разговора. Прямое заимствование подходов, принятых в обработке текстов, может оказаться неэффективным, так как текстовый подход учитывает признаки, связанные с орфографией и пунктуацией, что пока не реализовано в системах распознавания устной речи.

В работе [3] предложен метод определения тематики, основанный на использовании морфологического словаря. Для слов, важных для определения данной тематики с точки зрения частоты их встречаемости, с помощью морфологического словаря строятся расширения — осуществляется поиск и выбор слов, близких по морфологическим и семантическим свойствам к данному слову. В результате каждое такое слово становится корнем целого дерева производных слов, близких по значению к исходному. Идентификация темы разговора производится на базе таких деревьев-слов. Достоинством алгоритма является то, что он интуитивно соответствует представлениям о том, как осуществляется идентификация тематики человеком. Например, подход, основанный на использовании *TF-IDF*, не учитывает всех возможных словоформ, т.е. слова, которые отличаются от данного слова родом, числом или падежом могут не иметь достаточно высокого значения *TF-IDF*. При построении дерева учитываются все словоформы. Это позволяет сократить размерность задачи и не «размазывает» значение *TF-IDF* по словоформам.

Однако в описанном подходе есть и значительные недостатки. Основной недостаток — существенная зависимость от морфологического словаря. С учётом обычного отставания текстов словарей от современных реалий языка, использование морфологического словаря в такой системе означает обработку, скорее, грамматически правильных разговоров с лексикой 5–10-летней давности. Отметим также практическую сложность реализации алгоритма, основанного во многом на эмпирических правилах. Поэтому отсутствие количественных сведений об эффективности описанного подхода, кроме косвенных, представленных авторами [3], объяснимо.

Существующая система автоматического распознавания ключевых слов допускает значительное количество ошибок при распознавании слитной речи. Особенно, если использовать акустическую модель, неадаптированную для данного канала связи (см. раздел 4.1). Поэтому мы попробовали подойти к данной задаче с другой стороны — не трогать систему распознавания, а определять по результатам её работы тему сообщения. Подобные исследования, проводимые для русского языка, авторам не известны.

В данной работе предложен алгоритм для определения тематики устного разговора. Преобразование документов в вектор признаков осуществляется схожим с *TF-IDF* методом, описанным выше. В качестве алгоритма классификации используется решающее правило, которое выбирает тему с максимальным отношением правдоподобия, т.е. наиболее вероятную. Для сокращения размерности задачи используется метод отбора признаков — наиболее характерных слов для каждой из тематик. В статистическом подходе, даже в условиях, когда только некоторые слова правильно распознаются автоматической системой, вероятность встретить слова, характерные для данной темы, высока. Это обеспечивает значительную устойчивость к шуму.

В разделе 3 описан вероятностный подход к распознаванию тематики сообщения. Предложены различные эвристики для улучшения качества классификации, такие, как: сглаживание при восстановлении плотности вероятности по обучающей выборке, отбор наиболее информативных признаков или взвешивание признаков методом градиентного спуска.

В разделе 4 описаны результаты проведённого вычислительного эксперимента. Приведены графики с оптимизацией параметров алгоритма. И обозначены сложности, с которыми столкнулись экспериментаторы.

Определения и обозначения

Во избежание путаницы приведём здесь основные понятия, которые авторы часто используют в данной работе, как синонимы.

Исходный объект, с которым мы работаем, — устное сообщение или звуковой фрагмент, после обработки системой распознавания превращается в цепочку слов, которую мы называем документом. Это определение заимствовано из области обработки текстов. Эта цепочка с помощью словаря преобразуется в векторов слов W = $(\#w_1,...,w_N)$, где число вхождений слова w_i в документ. Часто все понятия в этом абзаце использу-ются для обозначения W, поскольку он является конечным результатом преобразования.

Тема или тематика, иногда класс обозначается буквой t, а всё множество тематик — T.

Корпус данных и обучающая выборка тоже часто отождествляются. Небольшой размер корпуса не позволяет разделить его на обучающую и тестовые части. Для контроля используется критерий LOO (Leave One Out), что позволяет использовать всю выборку, кроме одного объекта для обучения алгоритма.

Алгоритм тематической классификации устных сообщений

Особенности задачи

В нашем случае задача классификации выглядит следующим образом:
X — множество объектов, t — множество классов.
Есть признаковое описание объектов $X_i = (f_{i1},...,f_{iN})$, f — некоторые признаки.
Имеется конечная обучающая выборка:

$$\left\langle \begin{array}{c|c} f_{11},\ldots,f_{1N} & t_1 \\ \vdots & \vdots \\ f_{m1},\ldots,f_{mN} & t_m \end{array} \right\rangle . \qquad (3)$$

Требуется построить алгоритм $\alpha: X \to t$, который для любого объекта определит его принадлежность к классу. Это задача с замкнутыми непересекающимися классами.

Для её решения известно много готовых методов, таких, как: нейронные сети, SVM (машина опорных векторов), метрические алгоритмы классификации, статистические алгоритмы.

Рассмотрим нашу задачу более подробно, чтобы выбрать подходящий метод решения. Число классов более двух — значит, бинарные классификаторы непригодны. Размерность задачи или число признаков равно размеру словаря (порядка 30000). Число объектов — около 500 (см. раздел 4.1). Как видим, нам очень сильно мешает «проклятие размерности». Это исключает возможность использовать метрические алгоритмы: все объекты в таком пространстве будут находиться одинаково далеко друг от друга [4]. К тому же у нас мало данных, на которых можно обучать модель, и в них присутствует значительный процент шума. На этом шаге отпадают слишком сложные, с большим числом параметров и неустойчивые к шумам методы.

Мы остановились на статистическом подходе, поскольку он больше других удовлетворяет приведённым выше требованиям. Алгоритм схож с тем, который использовали авторы [5].

3.2. Решающее правило

Пусть дан звуковой фрагмет, преобразованый в векторов слов W = $(\#w_1,...,w_N)$ с помощью системы автоматического распознавания речи, где — число вхождений слова w_i в звуковой фрагмент, N — размер словаря. Для определения наиболее вероятной тематики используем критерии отношения правдоподобия.

Отношением правдоподобия называется:

$$T(W,t)=\frac{P(W|t)}{P(W|\bar{t})}, \quad (4)$$

где $P(W|t)$ — вероятность того, что вектор слов *W* порождается тематикой *t*, а $P(W|\bar{t})$ — вектор порождается какой-то другой тематикой.

Решающее правило, которое используется в этой работе, делает выбор в пользу той тематики *t* (из набора тематик *T*), для которой отношение правдоподобия максимально:

$$\alpha(W)=\arg\max_{t\in T}\frac{P(W|t)}{P(W|\bar{t})}. \quad (5)$$

Подобный подход применим и для открытого множества тематик. Отказ от классификации происходит, если отношение правдоподобия меньше заранее заданного порога .

$$T(W,t)<c_0 \Rightarrow t\notin T. \quad (6)$$

При подсчёте вероятности $P(W|t)$ используем наивный байесовский подход, полагаем статистическую независимость отдельных слов из *W<$>*, что в данном случае недалеко от действительности, поскольку при малом количестве правильно распознанных слов логические связи между ними теряются. В соответствии с этим предположением вероятность порождения *W* тематикой *t* представляется как:

$$P(W|t)\approx\prod_{i=1}^{N}P(w_i|t). \quad (7)$$

При использовании конечного словаря *V* выражение перепишется следующим образом:

$$P(W|t)\approx\prod_{\forall w\in V}P(w_i|t)^{\#w}. \quad (8)$$

Напомним, что — одна из компонент вектора *W*, индекс которой соответствует индексу слова *w* в словаре *V*. В данной интерпретации возможны нецелочисленные значения параметра, в качестве которых можно использовать вероятность появления *w* , полученную системой распознавания речи.

Вероятность $P(W|\bar{t})$ определяется следующим образом:

$$P(W|\bar{t})\approx\frac{1}{N_T-1}\sum_{\forall t_i\neq t}P(W|t_i), \quad (9)$$

где N_T — общее количество тематик. Выражение (9) получено в предположении, что все тематики имеют одинаковые априорные вероятности $P(t_i)$.

На практике удобно использовать логарифмическую шкалу. Это позволяет возведение в степень в формуле (8) заменить простым умножением. В этом случае оценка принадлежности вектора *W<$>* к классу *t* представима в виде функции правдоподобия:

$$F(t|W)\approx\sum_{\forall w\in V}\#w\cdot\log\frac{P(w|t)}{P(w|\bar{t})}. \quad (10)$$

Решающее правило или алгоритм классификации выглядит следующим образом:

$$\alpha(W) = \arg\max_{t \in T} \sum_{\forall w \in V} \#w \cdot \log \frac{P(w|t)}{P(w|\overline{t})}. \quad (11)$$

Обозначим логарифм отношения правдоподобия в виде матрицы *H[N x m]* , где *N* — размер словаря, а *m* — число тематик.

$$H = \log \frac{P(w|t)}{P(w|\overline{t})}. \quad (12)$$

Тогда решающее правило (11) будет выглядеть как скалярное произведение:

$$\alpha(W) = \arg\max_{t \in T} (W \cdot H(t)), \quad (13)$$

где *H*(*t*) — столбец соответствующий тематике *t*. Заметим, что *W* в нашем случае пропорционально *TF*, а *H* — аналог множителя *IDF*.

Оценка параметров распределений

Для подсчёта функции правдоподобия необходимо иметь распределение условных вероятностей для каждого из слов. Пусть *P*(*w*) означает априорную вероятность возникновения слова *w* независимо от темы и оценивается как максимум апостериорной вероятности с применением δ-сглаживания:

$$P(w) = \frac{N_w + \delta}{N_W + \delta N_V}, \quad (14)$$

где N_w — число вхождений слова *w* в обучающую выборку, а N_w — общее число слов в выборке. Использование δ-сглаживания позволяет избежать нулевых вероятностей для некоторых слов, которые не встречаются в обучающей выборке.

Вероятность $P(W|t)$ оценивается аналогичным образом как максимум апостериорной вероятности с применением δ-сглаживания:

$$P(w|t) = \frac{N_{w|t} + \delta N_V P(w)}{N_{W|t} + \delta N_V}, \quad (15)$$

где N_V — число слов в словаре, $N_{(w|t)}$ — число слов *w*, принадлежащих классу *t*.

Отбор признаков

Часто бывает, что небольшое число ключевых слов хорошо характеризует определённую тематику и редко или вообще не встречается в других. В то время как другие слова (союзы, предлоги, междометия) не вносят никакого вклада в определения тематики. По этой причине необходимо использовать метод отбора информативных признаков. В данной работе предлагается выбирать N наиболее значимых слов для каждого класса, имеющих наибольшую апостериорную вероятность для данного класса $P(t|w)$, которая оценивается следующим образом:

$$P(t|w) = \frac{N_{(w|t)} + \delta}{N_w + \delta N_T}, \quad (16)$$

где N_T — общее количество классов, $N_{(w|t)}$ — число слов *w*, принадлежащих классу *t*.

Взвешивание признаков

Отбор признаков можно рассматривать как частный случай взвешивания признаков, где каждый признак берётся с весом 0 или 1. В общем случае мы можем брать веса с любыми значениями (в данной работе мы ограничимся неотрицательными).

Введём вес λ_w для каждого слова *w*, тогда выражение (10) примет следующий вид:

$$F(t|W) \approx \sum_{\forall w \in V} \lambda_w \#w \cdot \log \frac{P(w|t)}{P(w|\bar{t})}. \quad (17)$$

Наша цель — найти вектор весов, минимизирующий ошибку классификации.

Введём понятие отступа для объекта — вектора слов *W*<$>:

$$M(W) = F(t_I|W) - F(t_G|W), \quad (18)$$

где t_G — класс, к которому принадлежит *W*, а t_I — ошибочный класс, получивший наибольшую оценку. Если объект верно классифицирован, значение отступа будет отрицательным, если неправильно — положительным.

Значение отступа можно гладко отобразить в отрезок [0,1] с помощью сигмоидной функции потерь:

$$l(W) = \frac{1}{1 + e^{-\beta M(W)}}, \quad (19)$$

где β определяет наклон сигмоидной функции. Среднее значение функции потерь аппроксимирует ошибку классификации и становится точным при $\beta \to \infty$. Поэтому, минимизируя среднее значение функции потерь *l*(*W*), мы уменьшим ошибку классификации. Так как функция потерь сглажена монотонной сигмоидной функцией, то её можно продифференцировать по отдельным признакам и минимизировать с помощью методов градиентного спуска.

Частная производная функции потерь *l*(*W*) по весу признака λ_w будет равна:

$$\frac{\partial l(W)}{\partial \lambda_w} = \beta l(W)(1 - l(W))(f(t_I|W) - f(t_G|W))G_{w|W}. \quad (20)$$

Пересчёт весов происходит по следующей формуле:

$$\lambda'_w = \lambda_w - \varepsilon \frac{1}{N_D} \sum_{\forall W} \frac{\partial l(W)}{\partial \lambda_w}, \quad (21)$$

где N_D — число объектов в обучающей выборке, а ε — параметр, определяющий скорость обучения. В нашей задаче мы не допускаем отрицательных значений весов λ_w, но не ограничиваем их сверху.

Вычислительный эксперимент

Создание корпуса данных для распознавания тематики устных сообщений

Собранный корпус данных состоял из аудиозаписей интервью различных людей на различные темы с текстовыми расшифровками, написанными челове-

ком. В основном записи были выбраны из сети Интернет. Всего собранный корпус содержал 25 часов аудиоданных, разбитых на пять тем: криминал, политика, спорт, быт и финансы.

Особенность аудиоматериала заключается в том, что он исходно сжат в формате с кодированием по стандарту MP3 со скоростью 32 кбит/с. Записи разборчивые, но искажения в речевом сигнале заметные. Таким образом, для преобразования, например, в сигнал качества мобильной телефонной связи проводилось преобразование форматов и низкочастотная фильтрация типа:

$$\text{MP3} \rightarrow \text{WAV} \rightarrow \text{ФНЧ } 3.2\text{кГц} \rightarrow \text{GSM6.10} \rightarrow \text{WAV}.$$

Более того, часть аудиозаписей состояла из телефонных интервью, которые записывались в студии по громкой связи, а уже потом конвертировались в MP3.

Это впоследствии существенно влияло на результаты измерений, так как система распознавания ключевых слов использовала модели «натурального» GSM сигнала.

Другой особенностью аудиосигналов было присутствие значительного числа посторонних сигналов, которые часто сопровождают аудиоинтервью, например, фоновой музыки или коротких рекламных сообщений.

Технология проведения измерений эффективности распознавания тематики предусматривала, что длинные записи сегментировались на отрезки по 200–500 слов (2–5 мин), в пределах файла тематика считалась неизменной, что не всегда было верно, особенно для коротких отрезков речи, где ведущие и интервьюируемые могли отклоняться от темы. Поэтому качество распознавания даже на текстовых данных не могло быть стопроцентным.

Размер словаря корпуса данных — 28K слов.

Для преобразования речи в текст использовалась система автоматического распознавания ключевых слов в потоке слитной речи с акустическими моделями для телефонного канала связи.

Экспериментальная оценка эффективности алгоритма распознавания тематики устных сообщений

Эффективность распознавания тематики измерялась в зависимости от:

- длительности фрагмента записи (100–300–500–1000) слов 1–3-5–10 минут;
- числа ключевых слов на тематику;
- эффективности работы распознавания.

Влияние качества сигнала (в виде оценки СШО) оценивалось косвенным образом, как зависимость от пословной ошибки системы распознавания ключевых слов.

Эксперимент проводился на модельных данных, распознанных человеком, и так называемых реальных данных, распознанных программой. Модельные использовались для выявления зависимостей, а сильно зашумлённые реальные — для проверки устойчивости алгоритма.

В качестве критерия эффективности распознавания тематики использовалась величина ошибки распознавания:

$$P_{err} = \frac{N_{incorrect}}{N_{total}}, \qquad (22)$$

где $N_{incorrect}$ — число записей, тематика которых была неправильно определена, N_{total} — общее число проверявшихся записей. Эффективность оценивалась как индивидуально для каждой тематики, так и суммарно по всем тематикам.

При оценке эффективности в зависимости от длительности разговора длинные записи корпуса данных (в среднем более 15 минут) сегментировались на отрезки от 0.5 до 5 минут (что при темпе речи 100–120 слов в минуту соответствовало 100–600 словам), затем каждый отрезок обрабатывался как отдельная реализация.

В пределах записи интервью — файла тематика считалась неизменной, что не всегда было верно, особенно для коротких сегментов, где ведущие и интервьюируемые иногда существенно отклонялись от темы. Ошибки такого рода не корректировались.

Функционалом качества являлась суммарная ошибка классификации при исключении контрольного объекта из обучающей выборки (метод скользящего контроля или «Leave One Out»).

Эффективность распознавания тематики в зависимости от числа ключевых слов

Для реальных и модельных данных проводился отбор признаков, который позволял выбирать *N* самых значимых слов за каждый класс.

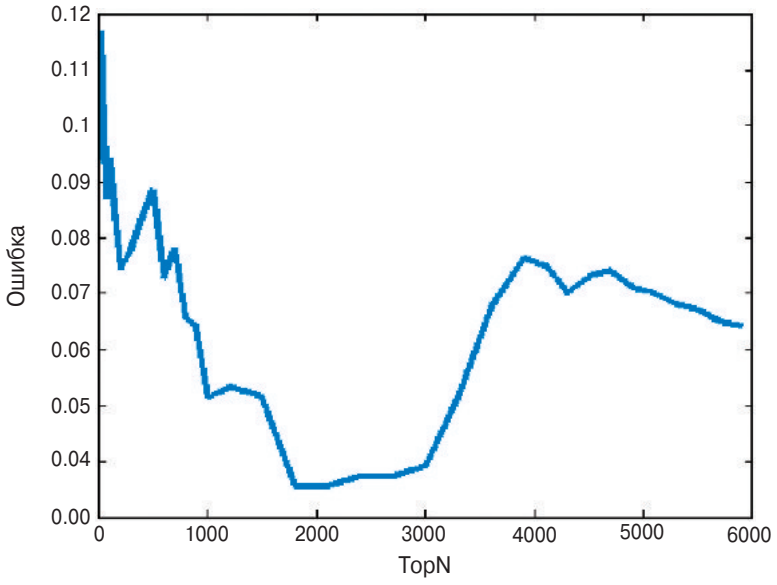

Рис. 1. Ошибка классификации в зависимости от числа ключевых слов (модельные данные)

На рис. 1 изображена зависимость общей ошибки классификации от количества отобранных признаков за каждый класс на модельных данных. Видно, что есть характерный набор слов, около 1800 для каждого класса, который выделяет его относительно других. При дальнейшем увеличении параметра выше 3000 качество распознавания ухудшается.

В лучшем случае удалось достичь ошибки 3,5%; дальнейшее её уменьшение не является главной целью, так как исходные данные не были идеально разделены на классы.

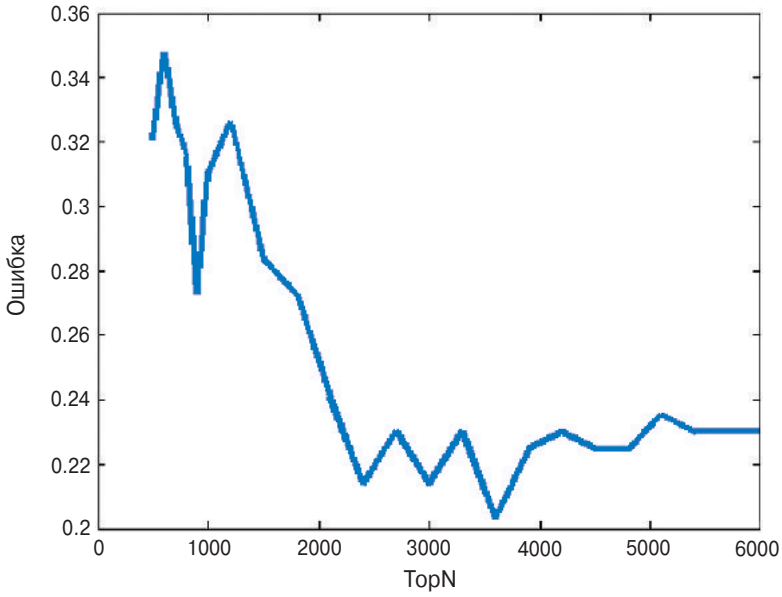

Рис. 2. Ошибка классификации в зависимости от числа ключевых слов (реальные данные)

Для реальных данных, распознанных программой, использование для классификации матрицы $\boldsymbol{H}$ (12), вычисленной по модельным данным не привело к хорошим результатам (ошибка около 70%). Обучение проводилось на реальных данных. График на рис. 2 отображает такую же зависимость ошибки от числа признаков для зашумлённых данных и имеет вначале похожую структуру. Однако качество распознавания практически не ухудшается при увеличении числа признаков. По данным графикам можно сказать, что размерность задачи нельзя существенно сократить. Она остаётся значительной $n \approx 10000$. Возможно, в дальнейших работах для этой цели мы будем использовать лемматизацию (приведение всех слов к нормальной форме) с помощью морфологического словаря.

В следующей таблице (Таблица 1) в качестве иллюстрации приведены наиболее информативные 20 слов для каждой из тематик. Таблица вычислена по модельным данным.

Таблица 1. Первые 20 наиболее характерных слов для каждой тематики (модельные данные)

Криминал	Политика	Спорт	Быт	Финансы
метро	союзного	спорта	<смех>	банков
терроризм	разного	спортсменов	***	банки
терроризмом	существу	спорте	угу	евро
терроризма	господина	катании	давай	кредиты
терактов	точности	сборной	блин	банка
теракты	демократия	катания	<вдох>	дозор
террористов	образцы	команды	абонент	финансовой
спецслужбы	индустрия	футбола	<шум>	кризиса
теракт	сюжет	тренер	ага	ставка
спецслужб	акционеров	тренеров	<кашель>	процентов
война	главой	тренеры	короче	банк
террористической	детского	спорт	***	прибыль

Таблица 1 (окончание)

террористы	замечательный	спортсмены	двадцать	банковской
взрыв	пальцев	фигурном	пятьдесят	индекс
взрывы	письмом	фигурного	чтоб	спрос
боевиков	собирают	тренера	знаешь	индексы
присяжных	соглашаются	школа	<выдох>	товары
безопасности	усыновления	большом	тобой	кредитов
взорвали	гарантии	спортивной	всё	экономики
взрывов	слушаю	спортсмен	щас	банкам

Очевидно, что вычисленные слова действительно отражают специфику соответствующих тематик. <…> — означает, что в записи присутствуют не слова, а специфические посторонние шумы, *** — ненормативная лексика.

В случае, когда алгоритм обучался на данных, распознанных программой с большим числом ошибок, получалась таблица, где характерные слова для класса не всегда соответствовали ожидаемым:

Таблица 2. Первые 12 наиболее характерных слов для каждой тематики (реальные данные)

Криминал	**Политика**	**Спорт**	**Быт**	**Финансы**
терроризма	договоры	выеду	личном	банки
террористов	сохранилась	команду	лучшему	кредиты
потому	охота	футболе	заходя	ударены
каждому	высока	спорта	кончились	численность
напряги	дошли	увеличивается	отнимали	фамилия
устройся	заходит	учились	охотно	банков
величия	казахского	футбольного	триллион	вставать
курам	концы	футбольную	банальности	кампании
трупов	мешалкой	шланги	взрослая	отраслевые
налажена	нагрузка	внедрена	догма	прибрать
ростову	надежд	командой	исхожу	расходов

Взвешивание признаков

Взвешивание признаков методом градиентного спуска по эффективности оказалось таким же, как и отбор признаков. Вначале он на 3,5% улучшил качество классификации, по сравнению с методом, использующим все признаки, но после 250 итераций ошибка резко возросла, хотя отступ $N(W)$ продолжал уменьшаться. Это обусловлено несбалансированной выборкой и сильным выделением отступа для какого-то одного класса в ущерб остальным. Такой эффект носит название «переобучения модели». На рис. 3 показана зависимость между качеством распознавания и числом итерации по пересчёту весов (21).

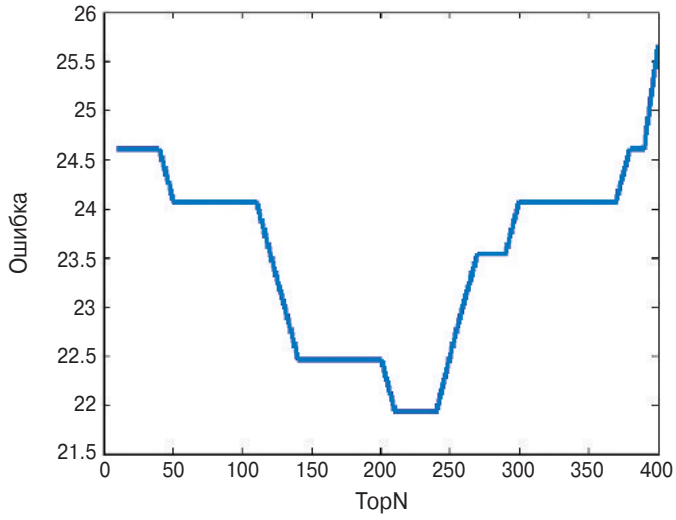

Рис. 3. Зависимость качества распознавания от числа итераций

Метод взвешивания признаков имеет больший потенциал, его невысокая эффективность в наших экспериментах обусловлена малым размером обучающей выборки. Поскольку число параметров модели возросло (бинарные веса заменили вещественными), то и корпус данных нужно значительно увеличить.

Эффективность распознавания тематики в зависимости от длительности записи

Так же мы проверили зависимость качества распознавания от длины фрагмента текста. Чем длиннее фрагмент, тем вероятнее, что встретится больше слов, характерных для данного класса и качество классификации улучшится.

На рис. 4 показана зависимость ошибки распознавания тематики от длины анализируемого речевого фрагмента. На графике видно, что в коротких отрывках по 50–150 слов очень сложно выделить тематику. Снижение качества распознавания в правой части графика, при длине фрагмента 1500 слов (10 минут) связано с тем, что выборка вырождается, так как в ней становится очень мало объектов (около 150).

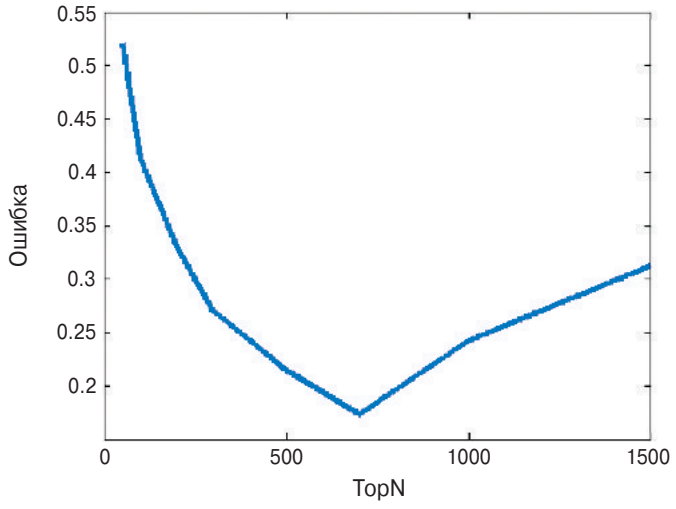

Рис. 4. Зависимость качества распознавания от длины (число слов) речевого фрагмента

Полученные результаты сведены также в таблице 3 с ошибками за каждый класс. Темп речи — примерно 150 слов в минуту.

Таблица 3. Вероятность правильного распознавания тематики в зависимости от длины записи

Длительность анализируемых записей	**Ошибка P_{err} в зависимости от тематики**					
	Криминал	Политика	Спорт	Быт	Финансовая	Всего
30 сек.	52.2	60.1	58.0	46.7	46.8	52.1
1 мин.	43.1	53.3	46.6	31.4	34.0	41.3
2 мин.	32.4	43.3	44.9	25.6	24.8	33.0
5 мин.	16.7	33.3	30.0	12.5	20.4	21.4

Обсуждения и выводы

В данной работе предложен статистический алгоритм для автоматической классификации тематики разговора на основе системы распознавания русской речи.

Описан метод отбора признаков, который выделяет наиболее характерные слова для каждой тематики. Этот метод позволяет примерно в три раза сократить размерность задачи (с 30000 до 10000).

Так же реализован метод взвешивания признаков. Он демонстрирует эффективность, сравнимую с методом отбора признаков. Но мы полагаем, что если увеличить размер обучающей выборки, его эффективность должна возрасти.

Основной сложностью является низкий показатель качества распознавания слитной русской речи в телефонном канале связи (опускается до 5–10%). Однако даже при таких зашумленных исходных данных проведённые вычислительные эксперименты показывают приемлемое качество классификации, ошибку удаётся снизить до 22% при разбиении на пять тематик.

Для дальнейших исследований в этой области необходимо улучшить качество распознавания речи, используя подходящие модели языка и подстроить акустическую модель под используемый канал связи. Увеличить размер корпуса данных для проведения экспериментов. Также необходимо выполнить лемматизацию, т.е. приведение всех слов в нормальную форму, используя морфологический словарь, как предлагают авторы [3].

Литература

1. *Wayne C.L*. Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation // Proc. 2nd Int. Conf. Language Resources and Evaluation, LRE, p.168, 2000.

2. Wu H.C, Luk R.W.P., Wong K.F., Kwok K.L. Interpreting TF-IDF term weights as making relevance decisions // ACM Transactions on Information Systems, 26 (3), Pp. 1–37, 2008.

3. *Schone P., Nelson D*. A Dictionary based method for determining topics in text and transcribed speech //Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'96, 1996, Pp.295–298.

4. Whittaker E.W.D., Woodland P.C. «Language modeling for Russian and English using word and classes» // Computer Speech and Language, Vol. 17, Pp. 87–104, 2003.

5. *Hazen T., Margolis A*. Discriminative feature weighting using MCE training for topic identification of spoken audio recordings // Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP, Pp. 4965-4968, 2008.

6. *Hazen T., Richardson F., Margolis A*. Topic identification from audio recordings using word and phone recognition lattices // Proc. Int. Conf on Automatic Spoken Recognition and Understanding, ASRU, Pp. 659-664, 2007

Сведения об авторе

Дорофеев Данила Викторович —

студент ФТЦ НАН Украины (магистратура). В 2010 г. закончил МФТИ (ГУ) факультет управления и прикладной математики (диплом бакалавра)

Чучупал Владимир Яковлевич —

кандидат физико-математических наук, ведущий научный сотрудник ВЦ РАН. Основная область интересов — распознавание и обработка речевых сигналов.