УДК 004.93'1, 004.912, 81`32 **DOI** 10.15622/sp.55.9

Д.А. Кочаров, А.П. Меньшикова

ПРИМЕНЕНИЕ ЛИНГВИСТИЧЕСКИХ ПРИЗНАКОВ ДЛЯ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ИНТОНАЦИОННО ВЫДЕЛЕННЫХ СЛОВ В РУССКОЯЗЫЧНОМ ТЕКСТЕ

Кочаров Д.А., Меньшикова А.П. **Применение лингвистических признаков для автоматического определения интонационно выделенных слов в русскоязычном тексте.**

Аннотация. В данной статье предлагается метод автоматического предсказания интонационно выделенных слов, то есть наиболее важной информации в высказывании. Метод опирается на использование лексических, грамматических и синтаксических маркеров интонационного выделения, что делает возможным его применение в системах синтеза речи по тексту, где реализация интонационного выделения может повысить естественность звучания синтезированной речи.

В качестве методов классификации независимо друг от друга использовалось несколько различных моделей: наивная байесовская модель, модель максимальной энтропии и условные случайные поля. Сопоставление результатов, полученных в ходе нескольких экспериментов, показало, что использовавшиеся дискриминативные модели демонстрируют сбалансированные и примерно равные значения метрик качества, в то время как генеративная модель потенциально более пригодна для поиска интонационно выделенных слов в речевом сигнале.

Результаты, представленные в статье, сравнимы и в некоторых случаях превосходят аналогичные системы, разработанные для других языков.

Ключевые слова: интонационное выделение, просодия, лексический анализ, синтаксический анализ, байесовский классификатор, метод максимальной энтропии, условные случайные поля, русский язык.

1. Введение. Интонационное выделение — это перцептивно значимое для носителей языка подчеркивание части устного высказывания, осуществляемое с помощью просодических средств: движения основного тона, интенсивности, длительности, тембра. С помощью интонации человек выделяет в высказывании наиболее важные слова. Интонационное выделение является одним из основных средств оформления информационной структуры и содержания высказываний наряду с лексическим составом и синтаксисом.

Интонационное выделение отдельных слов или словосочетаний повышает естественность звучания синтезированной речи и делает ее менее монотонной и более легко воспринимаемой, помогает слушателю понять, какая информация является в сообщении наиболее значимой. Последнее играет большую роль в таких приложениях, как диалоговые, информационно-справочные и навигационные системы, а повышение естественности звучания и простоты восприятия синтезированной речи важно при создании аудиокниг. Как было показано в ходе перцептивных исследований, описанных в [1], синтезированные высказывания, в которых присутствует

интонационное выделение, являются для большинства слушателей предпочтительными или более приятными на слух, чем нейтральные варианты, особенно в тех случаях, когда озвучиваются отрывки из художественной или детской литературы.

Существует достаточно много работ, посвященных автоматическому определению интонационного выделения [2-10]; большая их часть опирается не только на текстовые, но и на паралингвистические данные и анализ речевого сигнала. Такие системы достигают высоких результатов, однако могут быть применены при разработке систем автоматического распознавания и понимания речи, но не синтеза речи по тексту. В этом случае для предсказания того, какие фрагменты предложения могут быть акустически выделены, необходимо обращаться непосредственно к тексту и той информации, которую можно из него извлечь. Примерами исследований, где описываются такие подходы, могут служить работы [2-4], где для предсказания применялись такие критерии, как коммуникативный тип предложения, различные меры оценки частотности слов, мера TF-IDF (статистическая мера, используемая для оценки важности слова для конкретного документа), части речи слов и их гиперонимы (понятия, обозначающие семантические множества или группы, к которым относятся рассматриваемые слова), которые извлекаются с помощью семантических ресурсов (например WordNet). Однако подобные признаки зависят от конкретного языка, и их использование для русского языка требует адаптации.

В данной статье представлены результаты экспериментов по автоматическому определению интонационно выделенных слов в тексте с применением различных синтаксических, грамматических и лексических признаков. В статье сравниваются результаты, полученные при помощи трех разных классификаторов: наивного байесовского классификатора, метода максимальной энтропии и условных случайных полей. Полученные результаты сравнимы и в некоторых случаях превосходят аналогичные системы, разработанные для английского и японского языков.

2. Классификационные признаки. Благодаря тому, что явление просодической выделенности в русском языке всесторонне исследовано отечественными лингвистами, можно выделить наиболее распространенные маркеры рассматриваемого явления и выбирать признаки с опорой на них [11-15]. Такими маркерами, например, являются:

– акцентные частицы, которые зачастую сигнализируют о возможном интонационном выделении слов, рядом с которыми они находятся;

– инверсия (отличный от нейтрального порядок слов в предложении);

– противопоставление;

– определенные виды лексики, такие как оценочная и экспрессивная лексика, интенсивы адвербиального типа («очень», «весьма»), кванторные слова («все», «некоторые»), («иногда», «обычно») и прочие.

Для классификации использовались признаки, формализующие лингвистические маркеры интонационного выделения. Признаки можно условно разделить на три категории: синтаксические, грамматические и лексические.

1. Синтаксические признаки. Основным синтаксическим показателем интонационного выделения является порядок слов в предложении, отличный от нейтрального. Формализация данного признака основывается на следующем предположении: поскольку конструкции с нейтральным порядком слов встречаются чаще, их вероятность будет значительно выше вероятности конструкций с инвертированным порядком слов. Для каждого слова с конкретным типом синтаксической зависимости вероятность находится справа или слева от хозяина, то есть слово, от которого оно зависит, будет вычисляться следующим образом:

$$P_{right} = \frac{\sum_{i=1}^{k} \min(\max(\mathrm{I}_{w_i} - \mathrm{I}_{h_i}, 0), 1)}{k},$$

$$P_{left} = \frac{\sum_{i=1}^{k} \min(\max(\mathrm{I}_{h_i} - \mathrm{I}_{w_i}, 0), 1)}{k},$$

где P_{right} — вероятность того, что слово с данным типом зависимости находится справа от хозяина; P_{left} — вероятность того, что слово с данным типом зависимости находится слева от хозяина; k — общее количество слов с данным типом зависимости; I_{w_i} — номер позиции текущего слова в предложении; I_{h_i} — номер позиции слова, от которого зависит текущее слово. Для главного слова в предложении данный вероятностный признак всегда признавался равным единице. Для всех остальных слов предложения в качестве вероятностного

признака указывалось либо значение P_{right}, если слово находилось справа от слова-хозяина, либо P_{left}, если слово находилось слева от слова-хозяина. Пример использования данного признака приведен на рисунке 1.

Рис. 1. Использование вероятностного синтаксического признака

Так как рассчитанная вероятность того, что слово-определение находится слева от слова-хозяина, равна 0,91, для прилагательного «хороший» в данном предложении значение синтаксического признака будет равно 0,91. Вероятность того, что слово-определение находится справа от слова-хозяина, равная 0,09, будет указана для прилагательного «красный». Вероятность конструкции с нейтральным порядком слов (определение предшествует главному слову), как и предполагалось, выше, чем вероятность конструкции с инверсией (определение следует за главным словом), при этом меньшая вероятность соответствует тому слову, на котором в данном предложении делается наибольший акцент («красный» выделено сильнее, чем «хороший»).

Синтаксический разбор текстов осуществлялся с помощью парсера ParseySaurus, основанного на системе SyntaxNet, использующей нейронную сеть реализации библиотеки TensorFlow для глубокого обучения [16, 17]. В исследовании применялась синтаксическая модель русского языка из пакета SyntaxNet, которая обучена на синтаксически размеченном подкорпусе Russian-SynTagRus [18], входящем в корпус Universal Dependencies [19].

2. Грамматические признаки. К грамматическим признакам относятся части речи текущего, предыдущего и последующего слов. Они необходимы для поиска наречий, числительных, имен прилагательных, несущих на себе эмфатическое ударение, которое зачастую не маркировано ни инверсией, ни принадлежностью слова к одному из особых классов, притягивающих интонационное выделение. Часть речи определялась в процессе работы синтаксического анализатора ParseySaurus, описанного выше.

3. Лексические признаки. В качестве классификационных признаков использовалось само слово, предыдущее и последующее.

Каждое из слов было приведено к нормальной форме с помощью морфологического анализатора Pymorphy2 [20].

Также использовались признаки, определявшие принадлежность текущего слова какому-либо из следующих классов:

- акцентные частицы («даже», «же», «только»),
- интенсивы («очень», «невероятно»),
- итеративы («вновь», «всегда»),
- слова, связанные с отрицанием («ни», «нет», «никогда»),
- кванторные слова («каждый», «любой»),
- наречия меры и степени («вполне»),
- вопросительные слова («зачем»),
- наиболее частотная оценочная лексика («отвратительный»),
- наречия, связанные с обозначением времени («срочно», «немедленно», «теперь»).

Классы слов были составлены предварительно авторами исследования на основе работ Т. М. Николаевой [11].

Для каждого класса была высчитана точность, с которой он позволяет определить интонационную выделенность:

$$\Pr = \frac{N_{prom}}{N_{prom} + N_{nonprom}},$$

где N_{prom} — количество выделенных слов в корпусе, входящих в данный класс; $N_{nonprom}$ — количество невыделенных слов в корпусе, входящих в данный класс. Значения точности для описанных классов приведены в таблице 1.

Таблица 1. Значения точности для лексических классов, использованных в качестве признаков

Название класса	Точность
Отрицание	0,42
Кванторные слова	0,42
Акцентные частицы	0,45
Наречия меры и степени	0,57
Итеративы	0,57
Оценочная лексика	0,58
Интенсивы	0,60
Обозначение времени	0,61
Вопросительные слова	0,65

Полученная для каждого класса точность применялась в качестве маркера класса. Значение признака для слов, не принадлежащих к какому-либо из перечисленных классов, приравнивалось к нулю. Значения признака лежат в диапазоне от 0 до 1. В качестве признаков для каждого слова указывались маркеры лексических классов самого слова, предыдущего и последующего слов, а также синтаксического хозяина и зависимого слова. Если зависимых слов было несколько, то учитывался класс с наибольшим значением лексического признака.

3. Классификаторы. Тестовая и обучающая выборки представлялись в виде последовательности слов, для каждого из которых были указаны значения выбранных признаков. В результате классификации необходимо было отнести слово к одному из классов: (1) интонационно выделенное слово, (2) интонационно не выделенное слово. Было проведено несколько экспериментов с целью апробации и сравнения классификационных методов, предполагающих разные степени зависимостей слов и их признаков между собой:

– метода, предполагающего независимость объектов и независимость признаков объектов — наивного байесовского классификатора,

– метода, предполагающего независимость объектов и зависимость признаков объектов друг от друга — метода максимальной энтропии,

– метода, предполагающего зависимость объектов и признаков объектов друг от друга — условных случайных полей.

1. Наивный байесовский классификатор (НБК). Наивный байесовский классификатор — это алгоритм классификации, основанный на теореме Байеса и подразумевающий независимость признаков [21]. Данный классификатор рассматривает слова в высказывании как независимые друг от друга и не учитывает их последовательность. Класс c слова w вычисляется по формуле:

$$c = \arg\max_{c \in C} P(\mathrm{c}) \prod_i P(\mathrm{o}_i \mid \mathrm{c}),$$

где c — это класс слова w среди множества всех классов C, а o_i — это признаки слова w.

Наивный байесовский классификатор был реализован при помощи библиотеки Natural Language Toolkit (NLTK) для Python [22].

2. Метод максимальной энтропии (ММЭ). Метод максимальной энтропии — это дискриминативный метод классификации, который в

отличие от Наивного байесовского классификатора не предполагает независимости признаков [23]. Это свойство очень важно для обработки лингвистических данных. Например, лексический класс слова в некоторой мере связан его частью речи (см. «наречия, связанные с обозначением времени»).

Данный классификатор из всех возможных классов C слова w выбирает класс с наибольшей энтропией, определяемой как:

$$P(\mathrm{c} \mid w) = \frac{1}{Z(w)} \exp(\sum_i \lambda_i f_i(w, \mathrm{c})),$$

где $f_i(w, \mathrm{c})$ — классификационный признак, λ_i — вес классификационного признака, $Z(w)$ — коэффициент нормализации, вычисляемый следующим образом:

$$Z(w) = \sum_c \exp(\sum_i \lambda_i f_i(w, \mathrm{c})).$$

Метод максимальной энтропии применялся в качестве классификатора, который учитывает возможную зависимость признаков слова друг с другом, но при этом считает слова в высказывании набором независимых объектов.

Метод максимальной энтропии был реализован при помощи библиотеки Natural Language Toolkit (NLTK) для Python [22].

3. Метод условных случайных полей (УСП). Метод условных случайных полей — это дискриминативный метод классификации, получивший широкое применение в задачах по разметке и сегментации текстов [24]. Данный метод схож с методом максимальной энтропии, однако в отличие от него позволяет учитывать окружающий контекст.

Вероятность $P(c \mid w)$ того, что слово w принадлежит к классу c, вычисляется следующим образом:

$$\mathrm{P}(\mathrm{c} \mid w) = \frac{1}{Z(w)} \exp(\sum_i \lambda_i f_i(\mathrm{c}_i, \mathrm{c}_{i-1}, w_i)),$$

где c — класс-гипотеза, w — классифицируемое слово, — классификационный признак, λ_i — вес классификационного

признака, $Z(w)$ — коэффициент нормализации, вычисляемый следующим образом:

$$Z(w) = \sum_{c} \exp(\sum_{i} \lambda_i f_i(c_i, c_{i-1}, w_i)) .$$

Преимуществом метода считается отсутствие требования к независимости наблюдаемых объектов, что позволяет учитывать контекст классифицируемого объекта. Это особенно важно при поиске в тексте интонационного выделения слов, зачастую обусловленного не столько характеристиками самого выделенного слова, сколько характеристиками его окружения.

Для построения классификатора, опирающегося на метод УСП, использовалось программное обеспечение CRF++ [25].

4. Экспериментальный материал. В качестве экспериментального материала были использованы данные речевого корпуса CORPRES [26], созданного на кафедре фонетики и методики преподавания иностранных языков Санкт-Петербургского государственного университета. Корпус содержит около 30 часов чтения текстов профессиональными дикторами и фонетическую, орфографическую и интонационную разметку, включающую сведения об интонационно выделенных словах. Для данного исследования из корпуса были взяты тексты двух художественных повестей, прочитанных восьмью дикторами, где для каждого слова было указано количество выделивших его дикторов. Объем двух текстов составляет 34,000 словоупотреблений, из которых 3854 были интонационно выделены хотя бы одним диктором. Корпус содержит 3800 предложений, 2233 из них содержали хотя бы одно интонационно выделенное слово. В таблице 2 указано, какое количество слов было выделено во время прочтения одним, двумя, тремя и так далее дикторами. Третий столбец таблицы («Доля от всех выделенных слов») показывает, какую часть от общей массы интонационно выделенных слов корпуса составляют слова, выделенные одним, двумя и так далее дикторами.

Таблица 2. Количество интонационно выделенных слов в корпусе

Количество дикторов	Количество слов	Доля от всех выделенных слов, %
1	2222	57
2	805	21
3	420	11
4	206	5
5	102	3
6	62	2
7	32	0,9
8	5	0,1
ИТОГО	3584	100

Из таблицы видно, что в текстах преобладают выделения, реализованные только одним диктором, и практически отсутствуют случаи, когда слово выделили сразу все восемь дикторов. Это является следствием того, что исследуемое явление характеризуется большой вариативностью и зависимостью от интерпретации текста конкретным диктором.

5. Обсуждение полученных результатов. Использование такой метрики, как правильность (accuracy) при оценке качества предложенного метода нецелесообразно вследствие сильного дисбаланса между классами интонационно выделенных и нейтральных слов: первых в несколько раз меньше, чем вторых. Поэтому для оценки качества применялись следующие метрики: полнота (recall), точность (precision) и F1-мера:

$$\mathrm{Pr} = \frac{TP}{TP + FP},$$

$$\mathrm{Re} = \frac{TP}{TP + FN},$$

$$F_1 = 2 * \frac{\mathrm{Pr} * \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}},$$

где TP — истинно-положительное решение, TN — истинно-отрицательное решение, FP — ложно-положительное решение, FN — ложно-отрицательное решение.

В ходе исследования было проведено три эксперимента: определение слов, интонационно выделяемых большинством говорящих; определение выделенных слов в материале, включающем предложения без выделенных слов; определение слов, интонационно выделяемых любым количеством говорящих.

Эксперимент 1: определение слов, интонационно выделяемых большинством говорящих.

В первом эксперименте выборки для обучения и тестирования составлялись из предложений корпуса, каждое из которых содержало хотя бы одно слово, выделенное как минимум половиной дикторов, то есть четырьмя (384 предложения). Выборка для обучения содержала 75% данного подкорпуса (288 предложений), тестовая выборка — оставшиеся 96 предложений. Из-за малого объема доступного материала для апробации классификационных методов использовалась кроссвалидация, при которой весь материал был

разделен на 4 части. В таблице 3 приведены результаты эксперимента, которые были получены с применением наилучших наборов признаков, определенных эмпирически, которые в свою очередь приведены в таблице 4. В качестве базового уровня эффективности в таблице приводятся результаты классификации при помощи модели «мешок слов» (то есть с использованием только лексем в качестве признаков), реализованной на основе метода максимальной энтропии (полученное с помощью данного метода значение F_1-меры для базового уровня было наивысшим).

Таблица 3. Результаты первого эксперимента

	Базовый уровень	Наивный байесовский классификатор	Метод максимальной энтропии	Условные случайные поля
Полнота	0,22	0,65	0,38	0,39
Точность	0,53	0,43	0,52	0,59
F_1-мера	0,31	0,52	0,44	0,47

Из таблицы 3 видно, что для всех трех методов классификации значение средневзвешенной метрики F_1 выше, чем для базового уровня, что демонстрирует необходимость и полезность применения других признаков помимо лексем. Также можно наблюдать, что для более сложных дискриминативных моделей значения F_1-меры ниже, чем для генеративной, что, скорее всего, связано с малым количеством материала: в связи с тем, что при любом наполнении обучающей выборки интонационно выделенные слова являются крайне редко встречающимся классом, его вероятность во всех используемых моделях была весьма мала; это привело к получению низких значений полноты при классификации с помощью методов, не делающих предположения о независимости признаков (методы максимальной энтропии и условных случайных полей). В целом значения метрик для двух этих методов довольно близки, что можно объяснить родственностью методов. Однако при данной постановке эксперимента условные случайные поля позволяют получить небольшой прирост по точности и полноте классификации по сравнению с методом максимальной энтропии (0,07 и 0,01 соответственно).

С другой стороны, в результатах, полученных с помощью наивного байесовского классификатора, как и ожидалось, наблюдается больший уклон в сторону полноты. Благодаря этому качеству генеративной модели ее выгодно использовать в задачах несколько иного плана, например, при поиске интонационных выделений в речевом сигнале: наивный байесовский метод можно эффективно использовать для определения как можно большего числа

потенциальных выделенных единиц в целях их последующего «отсеивания» с помощью признаков другого типа (акустических).

Относительно низкие значения метрик можно объяснить крайней вариативностью интонационного выделения, зависимостью от конкретного диктора и текста; также значение точности несколько занижено вследствие того, что во время тестирования интонационно выделенными считались только слова, на которых акцент был сделан как минимум четырьмя дикторами. В случае, если выделенная одним-тремя дикторами единица была классифицирована как выделенная, это считалось ошибкой первого рода (ложное срабатывание).

Также определенную проблему представляет отсутствие признаков, полученных с помощью более глубокого семантического анализа, что является следствием ограниченности семантических ресурсов для русского языка. Большое количество экспрессивной и оценочной лексики не попало в составленные вручную списки лексических классов, что привело к большому количеству ошибок второго рода (ложноотрицательное срабатывание) и, соответственно, снижению полноты. На качество признаков, связанных с синтаксическими отношениями, сильно повлияло качество синтаксического парсера.

В таблице 4 указаны классификационные признаки в порядке убывания эффективности для каждого метода. В каждом столбце выше черты располагаются признаки, попавшие в итоговый набор, дающий наилучшие результаты (первые три типа признаков для наивного байесовского классификатора, первые шесть — для методов максимальной энтропии и условных случайных полей).

Таблица 4. Эффективность классификационных признаков

Наивный байесовский классификатор	Метод максимальной энтропии	Условные случайные поля
Части речи	Слово	Слово
Слово	Соседние слова	Соседние слова
Лекс. класс слова	Части речи	Части речи
Лекс. классы соседних слов	Лекс. классы соседних слов	Лекс. класс слова
Соседние слова	Лекс. класс слова	Лекс. классы соседних слов
Синтаксические признаки	Синтаксические признаки	Синтаксические признаки
Лекс. классы главного и зависимых слов	Лекс. классы главного и зависимых слов	Лекс. классы главного и зависимых слов

Наиболее результативными оказались лексические и грамматические признаки. Эффективность самого классифицируемого слова и соседних с ним как признаков можно объяснить наличием определенного небольшого количества лексических единиц, которые часто выделяются конкретными дикторами (например частица «тоже», встретившаяся в выборках 21 раз, была выделена 20 раз). Лексические классы обобщают группы таких единиц; вероятно, отдельные слова как признаки станут менее эффективными по сравнению с лексическими классами при тестировании на более объемном и разнородном корпусе, включающем в себя предложения из текстов большего количества жанров. Части речи позволяют находить оценочные прилагательные и наречия, не попавшие в список лексических классов. Лексические классы главного и зависимых слов показали низкую результативность, скорее всего, вследствие того, что в большинстве случаев их функцию выполняют признаки классов соседних слов; действительно, слова, связанные синтаксическими отношениями, оказывают наибольшее влияние на выделенность друг друга в том случае, если они находятся по соседству (например, частицы, непосредственно следующие за главными словами или предшествующие им, акцентируют их). Достаточно часто встречаются ситуации, когда главное и зависимое слова находятся в разных концах предложения (например, главное слово придаточного предложения считается зависимым от главного слова главного предложения); тогда с точки зрения интонационной выделенности они никак не взаимосвязаны.

В целом все три метода показывают наилучшие результаты при определении интонационной выделенности на словах из определенных описанных выше лексических групп, а также на словах, связанных синтаксическими отношениями с акцентными частицами. Наибольшее количество ошибок второго рода связано с пропуском интонационно выделенных противопоставлений и оценочных слов. Другим источником ошибок является смешение классификационными методами (в частности, УСП) акцентирующих и акцентируемых слов (первые притягивают выделенность на соседние слова, вторые сами являются объектами выделения), что особенно часто случается в предложениях с указательными частицами и неопределенными местоимениями (например, в предложении «Это я говорил с вами», где дикторами было выделено слово «я», при классификации как выделенная была определена частица «это»).

Эксперимент 2: включение в тестовый материал предложений без интонационно выделенных слов.

Во втором эксперименте тестовый материал был организован таким образом, чтобы он был максимально приближен к реальному

тексту по количеству в нем интонационно выделенных единиц. Тестовая выборка содержала 25% подкорпуса, использовавшегося в первом эксперименте (94 предложения), и 94 предложения без интонационно выделенных слов (всего 188 предложений). На 75% подкорпуса, использовавшегося в первом эксперименте (290 предложений), проводилось обучение модели. Метрики качества для данного эксперимента указаны в таблице 5. Как в таблице 3, в качестве базового уровня приводятся результаты классификации при помощи модели «мешок слов», реализованной на основе метода максимальной энтропии.

Таблица 5. Результаты второго эксперимента

	Базовый уровень	Наивный байесовский классификатор	Метод максимальной энтропии	Условные случайные поля
Полнота	0,24	0,66	0,42	0,39
Точность	0,32	0,27	0,41	0,40
F_1-мера	0,27	0,39	0,41	0,39

Как можно видеть из таблицы, такая постановка эксперимента сильнее всего сказывается на результатах наивного байесовского классификатора, снижая значение F_1-меры на 0,13. В итоге наиболее сбалансированные метрики и наибольшее значение F_1-меры наблюдается при использовании метода максимальной энтропии.

Эксперимент 3: определение слов, интонационно выделяемых любым из говорящих.

Целью третьего эксперимента было установить точность автоматического определения слов, которые были выделены читателями хотя бы единожды. При этом было сохранено высокое качество обучающего материала, то есть в обучающих выборках по-прежнему использовались предложения, содержащие слова, выделенные хотя бы четырьмя дикторами (384 предложения). Выборка для тестирования состояла из предложений, содержащих как минимум одно слово, выделенное одним, двумя или тремя дикторами (1767 предложений). В таблице 6 приведены данные о количестве интонационно выделенных слов в тестовой выборке в зависимости от количества дикторов, реализовавших выделение.

Таблица 6. Количество интонационно выделенных слов в тестовой выборке

Количество дикторов, выделивших слово	Количество слов	Доля от общ. кол-ва слов в выборке, %
1	1748	12,8
2	647	4,7
3	347	2,5

Для всех слов из тестовой выборки были получены вероятностные оценки каждого из двух возможных классов (оценка максимального правдоподобия для классификатора максимальной энтропии; граничная вероятность для условных случайных полей; вероятностная оценка, восстановленная из логарифмической оценки для байесовского классификатора). На основе этих оценок в каждом предложении было выбрано одно слово, для которого вероятность принадлежности к классу выделенных слов была наибольшей.

В таблице 7 указаны результаты третьего эксперимента. Предложением с правильно определенным интонационно выделенным словом считалось предложение, в котором слово с наибольшей вероятностью принадлежности к классу выделенных слов действительно было интонационно выделено хотя бы одним диктором. В качестве базового уровня эффективности была взята вероятность правильного определения выделенного слова случайным образом (относительное количество интонационно выделенных слов в предложении). (Значение базового уровня рассчитывалось отдельно для каждого предложения; в таблице указано среднее для всех предложений значение. В связи с тем, что распределение предложений по длинам ненормально и мультимодально, указанные базовые уровни не совпадают со значениями, получаемыми как отношение количества выделенных слов в предложениях и средней длины предложений — базовые уровни во всех случаях выше).

В последнем столбце таблицы приводятся результаты, полученные при помощи объединения вероятностных оценок всех трех классификаторов: в качестве интонационно выделенного слова выбиралось слово с наибольшим средним значением вероятностной оценки.

Таблица 7. Зависимость точности классификации в зависимости от длины предложения и количества интонационно выделенных слов

Кол-во выделенных слов в предложении	Кол-во предл.	Длина предл.	Базовый уровень	Доля предложений с правильно определенным выделенным словом			Объед. оценка
				НБК	ММЭ	УСП	
1	1094	8 ± 5	0,22	0,50	0,50	*0,52*	0,52
2	470	11 ± 6	0,30	0,59	0,57	*0,6*	0,6
3	133	14 ± 6	0,33	*0,72*	0,66	0,6	0,70
4	52	18 ± 6	0,32	0,65	*0,67*	0,65	0,65
5	17	18 ± 5	0,36	*0,88*	0,76	0,74	0,82
Средняя точность			0,30	0,67	0,63	0,62	0,66

Во-первых, стоит отметить тот факт, что с увеличением длины предложения растет вариативность его интерпретации говорящим, что подтверждается корреляцией между длиной предложения и количеством слов, интонационно выделенных хотя бы одним из прочитавших его. При этом процент таких слов в предложении практически не изменяется и колеблется около 33 % (см. столбец «Базовый уровень»), в то время как точность классификации растет с числом интонационно выделенных слов в предложении и в среднем на 34% точнее.

Объединение вероятностных оценок позволяет получить сбалансированные результаты: в четырех случаях из пяти полученное значение точности больше, чем среднее арифметическое значений точности отдельных классификаторов. Другими словами, при объединении оценок, как правило, удается достигать точности, которая либо равна лучшей из возможных при классификации одним методом, либо приближена к ней. Это выгодно, поскольку не наблюдается полного превосходства какого-либо одного метода над другими (для предложений с одним или двумя выделенными словами лучшие результаты показывают условные случайные поля, для предложений с тремя или пятью — наивный байесовский классификатор, для предложений с четырьмя — метод максимальной энтропии).

В таблице 8 приводится данные о корреляции между количеством выделивших слово дикторов и эффективностью его автоматического определения. Доля, указанная во втором столбце, считалась как отношение количества слов, выделенных N дикторами и классифицированных как выделенные, к общему количеству слов, выделенных N дикторами. Ожидаемо, для каждого классификационного метода доля тем выше, чем больше N. Это можно считать положительным результатом вследствие предположения о том, что слово считается тем более подходящим для интонационного выделения, чем большее количество дикторов выделили его при прочтении (акцент, реализованный одним диктором, может быть ошибкой или следствием специфической дикторской интерпретации текста, в отличие от акцента, реализованного тремя людьми).

Таблица 8. Корреляция между количеством выделивших слово дикторов и эффективностью его автоматического определения

Количество дикторов	Доля правильно классифицированных слов		
	Наивный байесовский классификатор	Метод максимальной энтропии	Условные случайные поля
1	0,30	0,30	0,31
2	0,41	0,39	0,40
3	0,55	0,51	0,55

В таблице 9 сравниваются результаты, представленные в данной статье, и результаты других исследователей, работавших над схожей задачей, то есть над определением выделенных слов предложения с помощью текстовых признаков. Условия и конкретные цели экспериментов, проводившихся в этих исследованиях, значительно отличаются друг от друга, поэтому для обеспечения наиболее корректного сравнения результаты данного исследования приводятся трижды, при этом указываются значения метрик, полученных в ходе экспериментов, наиболее близких по условиям к тем, что были представлены в других работах.

Таблица 9. Сравнение результатов с существующими системами

Исследование	Язык	Объем корпуса	Правильность (Accuracy)	F_1-мера	Точность (Precision)	Полнота
Nakajima et al. [3]	япон.	5737 слов	–	*67%*	70,6%	64,2%
Эксперимент 1	рус.	5745 слов	–	52%	43%	65%
Brenier et al. [2]	англ.	2906 слов		31%	42,4%	25,6%
Эксперимент 2	рус.	5745 слов		*41%*	41%	42%
Hovy et al. [4]	англ.	1000 предл.	44%	–		
Эксперимент 3	рус.	2151 предл.	*67%*	–		

В исследовании [2] обучение и тестирование модели проводилось на полном объеме доступного корпуса, 20% слов в котором были интонационно выделены; при этом выборки включали предложения без интонационно выделенных слов. В связи с этим результаты [2] сравниваются с результатами второго эксперимента, проведенного в данной статье.

Организация корпуса, применявшегося в [3], была такова, что около 70% предложений содержали более двух выделенных единиц, и около 5% не содержали ни одного, вследствие чего метод, описываемый в [3], был ориентирован на предсказание нескольких выделенных единиц в предложении; поэтому в таблице 8 приводится сравнение метрик качества для этого метода и метрик, приведенных в данной статье в описании первого эксперимента, где отсутствовали предложения без интонационно выделенных слов.

В [4] предсказание интонационно выделенных слов (т.е. наиболее важных в предложении и пригодных для интонационного

выделения) производилось способом, схожим с тем, что применялся в данном исследовании в третьем эксперименте: из всех слов предложения в качестве выделенного выбиралось одно, имеющее наибольшую вероятность оказаться в классе маркированных слов; вероятность оценивалась с помощью классификационного метода. Показателем точности считался процент предложений, где интонационно выделенное слово было определено верно.

В двух случаях из трех полученные в данном исследовании результаты выше аналогов. Вероятно, разница в полученных результатах связана не только с различиями в подходах, наборах признаков и классификационных методах, но и с особенностями исследуемых языков и объемами корпусов, на которых проводились исследования.

6. Заключение. Как показало проведенное исследование, несмотря на крайнюю вариативность и зависимость интонационной выделенности от конкретного диктора и его интерпретации художественного текста, автоматическое предсказание выделенных слов возможно осуществить с использованием только данных, извлекаемых из текста (грамматических, лексических и синтаксических признаков), с достаточной степенью эффективности. Результаты сопоставимы с результатами, полученными другими исследователями при решении той же задачи для других языков, и в двух случаях из трех превосходят их.

Сравнительный анализ методов классификации позволяет сделать вывод о том, что оптимальный метод следует выбирать, исходя из задач исследования. В частности, наивный байесовский классификатор, демонстрирующий наибольшие значения полноты, может быть использован при наличии второго этапа классификации, на котором будут отбрасываться лишние единицы, отнесенные к классу выделенных слов (например, это можно осуществить с помощью акустических признаков при определении интонационно выделенных слов в речи). Методы условных случайных полей и максимальной энтропии позволяют получить более сбалансированные и примерно равные значения метрик, что позволяет говорить о взаимозаменяемости методов при решении данной задачи.

Литература

1. *Strom V. et al.* Modelling Prominence and Emphasis Improves Unit-Selection Synthesis // Proceedings of Interspeech 2008. 2008.
2. *Brenier J., Cer D., Jurafsky D.* The detection of emphatic words using acoustic and lexical features // Ninth European Conference on Speech Communication and Technology (ICSLP'2005). 2005.

3. *Nakajima H., Mizuno H., Sakauchi S.* Emphasized Accent Phrase Prediction from Text for Advertisement Text-To-Speech Synthesis // 28th Pacific Asia Conference on Language, Information and Computation (PACLIC'2014). 2014. pp. 170–177.
4. *Hovy D. et al.* Analysis and Modeling of "Focus" in Context // Proceedings of Interspeech 2013. 2013. pp. 402–406.
5. *Mishra T., Sridhar V.K.R., Conkie A.* Word Prominence Detection using Robust yet Simple Prosodic Features // Proceedings of Interspeech 2012. 2012.
6. *Cernak M., Honnet P.E.* An empirical model of emphatic word detection // Proceedings of Interspeech 2015. 2015. pp. 573–577.
7. *Tamburini F.* Automatic detection of prosodic prominence by means of acoustic analyses // Lingue e linguaggio. 2015. vol. 14. no. 1. pp. 131–148.
8. *Johnson D.O., Kang O.* Automatic prominent syllable detection with machine learning classifiers // International Journal of Speech Technology. 2015. vol. 18. no. 4. pp. 583–592.
9. *Suni A., Aalto D., Vainio M.* Hierarchical representation of prosody for statistical speech synthesis // Computer Speech and Language Journal. 2017. vol. 45. pp. 123–136.
10. *Heckmann M.* Audio-visual word prominence detection from clean and noisy speech // Computer Speech & Language. 2018. vol. 48. pp. 15–30.
11. *Николаева Т.М.* Семантика акцентного выделения // М.: Наука. 1982. 106 с.
12. *Кодзасов С.В.* Законы фразовой акцентуации // Просодический строй русской речи. М.: Институт русского языка РАН. 1996. С. 181–206.
13. *Ковтунова И.И.* Современный русский язык. Порядок слов и актуальное членение предложения // М.: Едиториал УРСС. 2002. 240 с.
14. *Слюсарь Н.А.* На стыке теорий. Грамматика и информационная структура в русском и других языках // Либроком. 2009. 416 с.
15. *Luchkina T., Ionin T.* The effect of prosody on availability of inverse scope in Russian // Formal Approaches to Slavic Linguistics. 2015. vol. 23. pp. 418-437.
16. *Andor D. et al.* Globally normalized transition-based neural networks // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016. vol. 1. pp. 2442–2452.
17. *Alberti C. et al.* SyntaxNet Models for the CoNLL 2017 Shared Task // arXiv preprint arXiv: 1703.04929. 2017.
18. *Nivre J., Boguslavsky I., Iomdin L.* Parsing the SynTagRus Treebank of Russian // Proceedings of the 22nd International Conference on Computational Linguistics (CoLING'2008). 2008. vol. 2. pp. 641–648.
19. *Nivre J. et al*. Universal Dependencies v1: A multilingual treebank collection // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'2016). 2016.
20. *Korobov M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages // International Conference on Analysis of Images, Social Networks and Texts (AIST'2015). 2015. pp. 320–332.
21. *McCallum A., Nigam K.* A comparison of event models for Naive Bayes text classification // Proceedings of AAAI/ICML Workshop on Learning for Text Categorization. 1998. pp. 41–48.
22. *Bird S., Klein E., Loper E.* Natural Language Processing with Python // O'Reilly Media, Inc. 2009. 504 p.
23. *Berger A., Pietra V., Pietra S.* A Maximum Entropy Approach to Natural Language Processing // Computational Linguistics. 1996. vol. 22. no. 1. pp. 39–71.
24. C*ollobert R. et al.* Natural language processing (almost) from scratch // Journal of Machine Learning Research. 2011. vol. 12. no. Aug. pp. 2493–2537.

25. *Kudo T.* CRF++: Yet Another CRF Toolkit. 2013. URL: https://taku910.github.io/crfpp (дата обращения: 15.09.2017).
26. *Skrelin P. et al.* CORPRES - Corpus of Russian Professionally Read Speech // 13th International Conference Text, Speech and Dialogue (TSD'2010). 2010. pp. 392–399.

Кочаров Даниил Александрович — к-т филол. наук, доцент кафедры фонетики и методики преподавания иностранных языков, Санкт-Петербургский государственный университет (СПбГУ). Область научных интересов: автоматическая обработка речи и текста, математическая лингвистика, речевые технологии, фонетика, фонология. Число научных публикаций — 42. kocharov@phonetics.pu.ru; Университетская наб., 11, Санкт-Петербург, 199034; р.т.: +78123289565.

Меньшикова Алла Павловна — лаборант-исследователь филологического факультета, Санкт-Петербургский государственный университет (СПбГУ), студент кафедры математической лингвистики, Санкт-Петербургский государственный университет (СПбГУ). Область научных интересов: математическая лингвистика, речевые технологии. Число научных публикаций — 1. menshikova.alla2016@yandex.ru; Университетская наб., 11, Санкт-Петербург, 199034; р.т.: +79217638387.

Поддержка исследований. Работа выполнена в рамках проекта «Фонетические аспекты синтеза речевого сигнала высокого уровня естественности», финансируемого из средств Санкт-Петербургского государственного университета (№31.37.353.2015).

DOI 10.15622/sp.55.9

D.A. KOCHAROV, A.P. MENSHIKOVA

DETECTION OF PROMINENT WORDS IN RUSSIAN TEXTS USING LINGUSTIC FEATURES

D.A. Kocharov, A.P. Menshikova. **Detection of Prominent Words in Russian Texts Using Linguistic Features.**

Abstract. The article presents a method of detecting prosodically prominent words, i.e. words that carry most of the information in the utterance. The method relies on lexical, grammatical and syntactic markers of prominence, and can be used in Text-to-Speech synthesis systems to make synthesized speech sound more natural.

Three different classification methods were used: Naive Bayes, Maximum Entropy and Conditional Random Fields models. The results of the experiments show that discriminative models provide more balanced values of the performance metrics, while the generative model is potentially more useful for detecting prominent words in speech signal.

The results of the study are comparable with the performances of similar systems developed for other languages, and in some cases surpass them.

Keywords: prosodic prominence, emphasis, prosody, lexical analysis, syntax analysis, Naive Bayes classifier, Maximum Entropy classifier, Conditional Random Fields, Russian language.

Kocharov Daniil Alexandrovich — Ph.D., associate professor of phonetics and methods of teaching foreign languages department, Saint Petersburg State University (SPbSU). Research interests: automatic speech and text processing, computational linguistics, speech technologies, phonetics, phonology. The number of publications — 42. kocharov@phonetics.pu.ru; 11, Universitetskaya Emb., Saint-Petersburg, 199034; office phone: +78123289565.

Menshikova Alla Pavlovna — research assistant of the philology faculty, Saint Petersburg State University (SPbSU), student of mathematical linguistics department, Saint Petersburg State University (SPbSU). Research interests: computational linguistics, speech technologies. The number of publications — 1. menshikova.alla2016@yandex.ru; 11, Universitetskaya emb., Saint Petersburg, 199034, Russia; office phone: +79217638387.

Acknowledgements. The research was supported by SPbSU (project # 31.37.353.2015).

References

1. Strom V. et al. Modelling Prominence and Emphasis Improves Unit-Selection Synthesis. Proceedings of Interspeech 2008. 2008.
2. Brenier J., Cer D., Jurafsky D. The detection of emphatic words using acoustic and lexical features. Ninth European Conference on Speech Communication and Technology (ICSLP'2005). 2005.
3. Nakajima H., Mizuno H., Sakauchi S. Emphasized Accent Phrase Prediction from Text for Advertisement Text-To-Speech Synthesis. 28th Pacific Asia Conference on Language, Information and Computation (PACLIC'2014). 2014. pp. 170–177.
4. Hovy D. et al. Analysis and Modeling of "Focus" in Context. Proceedings of Interspeech 2013. 2013. pp. 402–406.
5. Mishra T., Sridhar V.K.R., Conkie A. Word Prominence Detection using Robust yet Simple Prosodic Features. Proceedings of Interspeech 2012. 2012.

6. Cernak M., Honnet P.E. An empirical model of emphatic word detection. Proceedings of Interspeech 2015. 2015. pp. 573–577.
7. Tamburini F. Automatic detection of prosodic prominence by means of acoustic analyses. *Lingue e linguaggio*. 2015. vol. 14. no. 1. pp. 131–148.
8. Johnson D.O., Kang O. Automatic prominent syllable detection with machine learning classifiers. *International Journal of Speech Technology*. 2015. vol. 18. no. 4. pp. 583–592.
9. Suni A., Aalto D., Vainio M. Hierarchical representation of prosody for statistical speech synthesis. *Computer Speech and Language Journal*. 2017. vol. 45. pp. 123–136.
10. Heckmann M. Audio-visual word prominence detection from clean and noisy speech. *Computer Speech & Language*. 2018. vol. 48. pp. 15–30.
11. Nikolaeva T.M. *Semantika akcentnogo vydelenija* [The semantics of the accentual prominence]. M.:Nauka. 1982. 106 p. (In Russ.).
12. Kodzasov S.V. [The laws of phrasal accentuation]. *Prosodicheskij stroj russkoj rechi* [Prosodic structure of Russian speech]. M.: Institut russkogo jazyka RAN. 1996. pp. 181–206. (In Russ.).
13. Kovtunova I.I. *Sovremennyj russkij jazyk. Porjadok slov i aktual'noe chlenenie predlozhenija* [Modern Russian: word order and the communicative structure of a sentence]. M.: Editorial URSS. 2002. 240 p. (In Russ.).
14. Sljusar' N.A. *At the junction of theories. The grammar and the informational structure in Russian and other languages* [Na styke teorij. Grammatika i informacionnaja struktura v russkom i drugih jazykah]. Librokom. 2009. 416 p. (In Russ.).
15. Luchkina T., Ionin T. The effect of prosody on availability of inverse scope in Russian. *Formal Approaches to Slavic Linguistics*. 2015. vol. 23. pp. 418–437.
16. Andor D. et al. Globally normalized transition-based neural networks. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016. vol. 1. pp. 2442–2452.
17. Alberti C. et al. SyntaxNet Models for the CoNLL 2017 Shared Task. arXiv preprint arXiv: 1703.04929. 2017.
18. Nivre J., Boguslavsky I., Iomdin L. Parsing the SynTagRus Treebank of Russian. Proceedings of the 22nd International Conference on Computational Linguistics (CoLING'2008). 2008. vol. 2. pp. 641–648.
19. Nivre J. et al. Universal Dependencies v1: A multilingual treebank collection. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'2016). 2016.
20. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages. International Conference on Analysis of Images, Social Networks and Texts (AIST'2015). 2015. pp. 320–332.
21. McCallum A., Nigam K. A comparison of event models for Naive Bayes text classification. Proceedings of AAAI/ICML Workshop on Learning for Text Categorization. 1998. pp. 41–48.
22. Bird S., Klein E., Loper E. Natural Language Processing with Python. O'Reilly Media, Inc. 2009. 504 p.
23. Berger A., Pietra V., Pietra S. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*. 1996. vol. 22. no. 1. pp. 39–71.
24. Collobert R. et al. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*. 2011. vol. 12. no. Aug. pp. 2493–2537.
25. Kudo T. CRF++: Yet Another CRF Toolkit. 2013. Available at: https://taku910.github.io/crfpp (accessed: 15.09.2017).
26. Skrelin P. et al. CORPRES - Corpus of Russian Professionally Read Speech. 13th International Conference Text, Speech and Dialogue (TSD'2010). 2010. pp. 392–399.