Информационные технологии и системы

УДК 004.934

Н.Н. Сажок

КЛАСТЕРИЗАЦИЯ СЛОВ ПРИ ПОСТРОЕНИИ ЛИНГВИСТИЧЕСКОЙ МОДЕЛИ ДЛЯ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧЕВОГО СИГНАЛА

Описано формирование лингвистической модели для распознавания слитной речи на основе объединения слов в классы. Широко применяемый для этого метод кластеризации с учетом рекурсии частот слов обеспечивает приемлемую скорость при работе со славянскими языками из-за огромного обилия словоформ. Анализ построения критерия точности аппроксимации дал возможность ввести рекурсию в итерации кластеризации на уровне компонентов критерия, что привело к существенному уменьшению сложности вычислений. Впервые автоматически сформированные для лингвистической модели распознавания украинской речи классы слов обобщают грамматические, семантические и фонетические признаки.

Введение. Характеристики систем распознавания речевого сигнала все больше приближаются к требованиям пользователя, которые, в свою очередь, неуклонно возрастают. Основными характеристиками эргономичности при этом выделим наличие практически всех слов, произносимых пользователем, в словаре системы распознавания и минимизацию количества времени, необходимого для «привыкания» системы к новому диктору.

Раньше пользователя удовлетворяло требование проводить обучение системы путем произнесения всех слов словаря, которым оперирует система. В настоящее время пользователь ожидает от системы возможности преобразования речи в текст уже с момента начала его взаимодействия с системой.

Переход от слова к фонеме в качестве базового элемента позволило избежать необходимости внесения в обучающую выборку всех слов словаря [1]. Таким образом, требования к настраиванию на голос диктора теоретически сократились до произнесения диктором некоторого фонетически обогащенного текста, составляющего несколько тысяч слов.

Современные системы распознавания речи процедуру настраивания на голос диктора в значительной мере компенсируют предварительно обработанными громадными корпусами речи кооператива дикторов. Дополнительная возможность адаптировать систему на голос диктора, особенно с учетом гендерности, приводит к уменьшению ошибок распознавания до двух раз [2]. Этот показатель достигается уже при произнесении примерно сотни слов.

Таким образом, на пути увеличения словаря распознавания было преодолено физическое препятствие, касающееся участия пользователя в

ISSN 0452-9910. **Кибернетика и вычисл. техника. 2012. Вып. 170**

настраивании системы. Это, в свою очередь, актуализировало проблему увеличения словаря.

Опыт создания системы реального времени для распознавания украинской речи на 100 тысяч слов показал, что объемы лингвистической базы данных и знаний такой системы превышают один гигабайт [3]. Согласно исследованиям [4], чтобы охватить лексикон произвольной речи на 99 % и более, система распознавания должна обрабатывать словарь объемом не менее 200 тысяч слов для западных языков и более миллиона слов для славянских языков, что обусловлено таким свойством славянских языков, как обилие словоформ. Соответственно, увеличение лингвистической базы данных и знаний станет еще более критичным, особенно учитывая дальнейшие работы по улучшению качества аппроксимации модели в целом.

Объединяя слова в классы, т.е. рассматривая определенные слова как эквивалентные, мы сможем существенно сократить количество элементов лингвистической модели и их контекстов. Существующие алгоритмы и инструментарий показывают свою ограниченность при их применении к славянским языкам. Таким образом, разработка усовершенствованных методов кластеризации слов по текстовому корпусу является критической задачей для развития речевой информатики в Украине.

Лингвистическая составляющая генеративной модели распознавания речи основывается на оценивании вероятности

$$P(\boldsymbol{w}) = \prod_{k=1}^{K_k} P(w_k \mid w_{k-1}, ..., w_1), \tag{1}$$

где $\mathbf{w}_{1:K} = (w_1, w_2, ..., w_K)$ — гипотетическая последовательность слов ответа распознавания [5]. Исходя из вычислительных соображений количество предыдущих слов ограничивается до $N-1$, и, таким образом, формируется N-граммная лингвистическая модель (ЛМ):

$$P(\boldsymbol{w}) = \prod_{k=1}^{K} P(w_k \mid w_{k-1}, w_{k-2}, ..., w_{k-N+1}), \tag{2}$$

где N обозначает ширину контекста ЛМ и обычно используется в пределах от двух до четырех. Вероятности N-грамм оцениваются по текстовому корпусу путем статистических подсчетов. Например, если обозначить $C(w_{k-N+1}, ..., w_{k-1}, w_k)$ как частоту N-граммы $(w_{k-N+1}, ..., w_{k-1}, w_k)$, то

$$P(w_k \mid w_{k-1}, w_{k-2}, ..., w_{k-N+1}) \approx \frac{C(w_{k-N+1}, ..., w_{k-1}, w_k)}{C(w_{k-N+1}, ..., w_{k-1})}. \tag{3}$$

Наибольшей теоретической проблемой при построении ЛМ является оценивание вероятностей тех N-грамм, для которых не набирается достаточно статистики. Тогда такая оценка проводится на основании статистик (N–1)-грамм. Другой проблемой является наличие в текстовом корпусе внесловарных слов, т.е. слов, не вошедших в рабочий словарь системы распознавания. Приемлемое решение этой проблемы достигается

путем введения категории *неизвестного слова*, заменяющего в текстовом корпусе все внесловарные слова. По своей сути, категория неизвестного слова уже является неким кластером или классом слов. Объединяя другие слова в классы, т.е. рассматривая определенные слова как эквивалентные, мы сможем существенно сократить историю наблюдения элементов ЛМ и их контекстов. С использованием функции классификации $G(w) = g$, где g принадлежит множеству классов слов, для биграмного контекста ($N = 2$) вероятности в произведении (2) могут быть представлены в виде [6]

$$P_{\text{class}'}(w_k \mid w_{k-1}) = P(w_k \mid G(w_k), G(w_{k-1}), w_{k-1}) P(G(w_k) \mid G(w_{k-1}), w_{k-1}). \quad (4)$$

Предполагается, что $P(w_k \mid G(w_k), G(w_{k-1}), w_{k-1})$ не зависит от $G(w_{k-1})$ и w_{k-1}, а $P(G(w_k) \mid G(w_{k-1}), w_{k-1})$ не зависит от w_{k-1}. Это приводит к заключительной модели:

$$P_{\text{class}}(w_k \mid w_{k-1}) = P(w_k \mid G(w_k)) P(G(w_k) \mid G(w_{k-1})). \quad (5)$$

Аналогичным образом рассматриваются более широкие контексты ($N > 2$).

Модель объединения слова в классы методами кластеризации. Пусть $\boldsymbol{W}$ – последовательность всех реализаций слов $(w_1, w_2, w_3, ...)$ из обучающей выборки на основе некоторого текстового корпуса, $\boldsymbol{V}$ – словарь или множество всех слов из $\boldsymbol{W}$. Тогда, исходя из выражения (1), для биграмного контекста наилучшее объединение слов в классы приводит к максимуму вероятности:

$$P_{\text{class}}(\boldsymbol{W}) = \prod_{x,y \in \boldsymbol{V}} P_{\text{class}}(x \mid y)^{C(x,y)}, \quad (6)$$

где (x, y) означает пару слов, в которой слово x следует за словом y в последовательности $\boldsymbol{W}$, а функция $C(\cdot)$ вычисляет частоту наблюдения аргумента в обучающей выборке. Во избежание проблем с малыми величинами применяется логарифмирование:

$$\log P_{\text{class}}(\boldsymbol{W}) = \sum_{x,y \in \boldsymbol{V}} C(x,y) \log P_{\text{class}}(x \mid y). \quad (7)$$

Применяя к $P_{\text{class}}(x \mid y)$ выражение (5) и (3), проведя ряд последующих преобразований и отбросив компоненты, не зависимые от функции классификации [6], приходим к формулировке окончательного критерия:

$$F_{\boldsymbol{G}} = \sum_{g,h \in \boldsymbol{G}} C(g,h) \log C(g,h) - 2 \sum_{g \in \boldsymbol{G}} C(g) \log C(g). \quad (8)$$

где (g, h) означает следование класса g за классом h.

Идея поиска некоего наилучшего объединения слов в заданное число классов состоит в вычислении изменения критерия $F_{\boldsymbol{G}}$ при гипотетическом отнесении каждого из слов в альтернативные классы с последующим перемещением в класс с наибольшим критерием. Таким образом, классы как

бы обмениваются словами, т.е. осуществляют некий **словообмен**, до тех пор пока критерий не перестает улучшаться. Ниже приводится пошаговая схема алгоритма словообмена.

Приведенный алгоритм относится к семейству «жадных» и не гарантирует глобального экстремума. На улучшение результата могут влиять способ инициализации и динамика наращивания классов. Кроме того, возможно получить лучший критерий, исходя из указания учителя-эксперта.

Шаг А. **Инициализация**: для всех слов из словаря вводится первоначальная функция классификации, которая относит все слова к одному классу, $G(w) = 1$, или распределяет слова по классам, исходя из соображений, например, частотности

Шаг Б. **Итерация**: для заданного количества итераций или по достижению определенного условия останова

Шаг Б1. **Итерация**: для каждого слова w из словаря $\boldsymbol{V}$

Шаг Б1а. **Итерация**: для каждого класса g из множества $\boldsymbol{G}$

– **Перемещаем** слово w в класс g, запоминая изначальный класс

– **Вычисляем** изменение F_G для этого перемещения (8)

– **Возвращаем** слово w в изначальный класс

Шаг Б1б. **Перемещаем** слово w в класс с наибольшим F_G

Для проведения дальнейших исследований алгоритма необходимо удостовериться в достижимости соответствующих вычислений на доступных вычислительных ресурсах за приемлемое время. Поэтому рассмотрим практическую сторону реализации алгоритма.

Рекуррентность в алгоритме словообмена. Прямое применение (8) приводит к сложности вычислений $O(IVG^3)$, где I — количество итераций шага Б, V — объем словаря обучающей выборки, G — число классов. В работе [7] предложен способ уменьшения количества арифметических операций за счет использования рекуррентных формул для частот при вычислении изменения критерия F_G на шаге Б1а. Это позволяет выйти на уровень сложности вычислений $O(IVG^2)$.

Произведенные замеры показали, что итерация с учетом базовой рекурсии для частот на шаге Б для словаря в 20 000 словоформ и количества классов в пределах 1000 протекает менее суток. Но, как видим далее, даже с учетом развития вычислительной техники достигнутый темп вычислений является неприемлемым при моделировании славянских языков. Общеизвестно, что, например, для украинского языка, количество словоформ на одно базовое слово в среднем превышает 15, тогда как для английского языка этот показатель равен двум. Т.е. соотношение объемов словаря равной по объему обучающей выборки для украинского и английского языков справедливо принять равным шести. Как показано в следующем разделе, соотношение количества классов слов для обоих языков примерно такое же. Таким образом, сложность вычислений может возрасти в 6^3 раз, а значит крайне необходимо дальнейшее ускорение работы алгоритма.

Предположим, что на предыдущей итерации шага Б при выполнении шага Б1б функция классификации $G^-(\cdot)$ отнесла слово w к классу u, т.е. $G^-(w)=u$. Далее, на шаге Б1а, предстоит проверить гипотезу перемещения w в класс v, т.е. $G(w)=v$.

Пронумеруем все классы $g_i \in \boldsymbol{G}$ от 1 до G и обозначим как $C_{ij}=C(g_i,g_j)$, $1\le i,j\le G$, частоту, вычисляемую при проверке гипотезы перемещения на шаге Б1а, а для частот в условиях изначального класса введем обозначение C_{ij}^-.

Очевидно, что наибольший вклад в сложность вычислений критерия (8) вносит первая сумма, которую, с учетом введенных обозначений, представляем в виде

$$\sum_{i,j} C_{ij}\log C_{ij} = \sum_{\substack{i,j \\ \{i,j\}\cap\{u,v\}=\varnothing}} C_{ij}\log C_{ij} + \sum_{\substack{i=u,v \\ j}} C_{ij}\log C_{ij} + \sum_{\substack{j=u,v \\ i\ne u,v}} C_{ij}\log C_{ij}\,. \tag{9}$$

Таким образом, мы разбили изначальную сумму на три части: 1) не меняющуюся при перемещении слова w из класса u в класс v (пустые ячейки на рис. 1) и 2) меняющуюся за счет исходного класса и 3) меняющуюся за счет потенциально перспективного тестируемого класса. Первая сумма в правой части выражения (9) представима в рекуррентном виде относительно первоначальной гипотезы:

$$\begin{aligned}\sum_{\substack{i,j \\ \{i,j\}\cap\{u,v\}=\varnothing}} C_{ij}\log C_{ij} &= \sum_{\substack{i,j \\ \{i,j\}\cap\{u,v\}=\varnothing}} C_{ij}^-\log C_{ij}^- = \\ &= \sum_{i,j} C_{ij}^-\log C_{ij}^- - \left(\sum_{\substack{i=u,v \\ j}} C_{ij}^-\log C_{ij}^- + \sum_{\substack{j=u,v \\ i\ne u,v}} C_{ij}^-\log C_{ij}^- \right).\end{aligned} \tag{10}$$

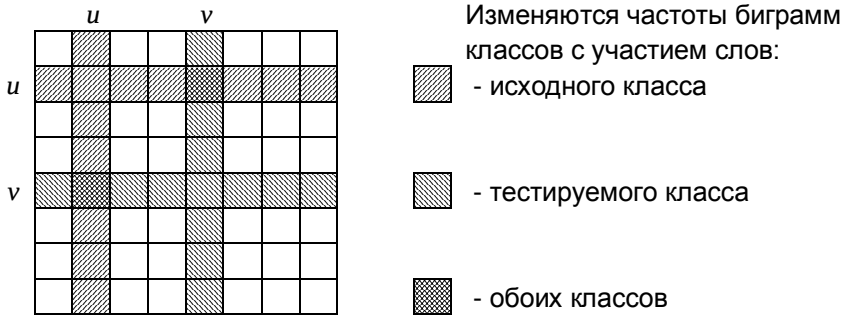

Рис. 1. Влияние перемещения некоторого слова из класса u в класс v на изменение частот биграмм классов C_{ij}

Приведенный анализ рекурсии позволяет свести количество базовых операций к минимуму. Путем ввода осторожной стратегии ограничения

проверяемых классов сложности вычислений уменьшается до обуславливаемого экспериментальными исследованиями уровня, достигая ускорения работы алгоритма, кратного, например, корню квадратному из общего числа классов. Учитывая постоянное расширение текстовой базы, введение многоязычия и увеличения порядка лингвистической модели, необходимо дальнейший рост быстродействия, что возможно за счет распараллеливания вычислений.

Связь классов слов лингвистической модели с лингвистическими категориями. При анализе результатов процедуры кластеризации возникает закономерный вопрос, каким образом соотносится содержание построенных кластеров с категориями, которыми оперирует лингвистическая наука, т.е. существует ли взаимосвязь автоматически сформированных классов слов с их грамматическим (части речи и формы слов), фонетическим и семантическим наполнением.

Предположим, что кластеризация происходит по признакам, определяемым перечисленными лингвистическими категориями. Оценим, потенциальное количество таких кластеров.

Анализируя семантическое наполнение, исходим из того, что имеется некоторое множество семантических признаков. В табл. 1 приведено одно из возможных разбиений на категории, включающие один или больше признаков, с примерами слов. Слова, не обладающие ни одним из введенных признаков, выделяются в отдельную категорию.

Таблица 1

Примеры семантических признаков с разбиением по категориям

Семантические признаки		**Примеры слов**
Имена собственные, топонимы и их производные (всего 8)	Имена	Валя, Миколині, Антононівна, Джоне
	Фамилии	Петренко, Обамою, кучмівської
	Названия	ООН, Укрнафта, Артеку, НАНУ, Динамо
	Территории	Україна, Крим, Арізона, райський
	Города	Київ, Харківське, Вашингтону
Человеческая деятельность	Прочее (3)	Трубіж, Егейське; Еверест; Великодні
	Общественная	виборці, команда, парламентські
	Интеллектуальная	знання, говорив, інтеграл, створено
	Физическая	ходити, пити, сон, чистий
	Пространство	метр, близько, праворуч, над, літрів
Количество, мера (всего 14)	Время	секунд,завтра,раніше,після,десятирічний
	Вес	грам, тонн, легко, фунтів
	Безразмерные	відсотків,бали,ступеню,більше,половина
	Числительные (5)	один; десятий; двісті; мільйони; багато
	Прочее (5)	гривень; байт; градуси; вольт; карат

Обращаясь к лексикографической базе данных и знаний [8], можно оценить количество грамматических форм слов для различных частей речи с учетом введенных семантических признаков (табл. 2). Подсчеты показывают справедливость оценки, превышающей 1000 лингвистических категорий в целом.

Отдельный интерес вызывает обработка слов-омографов. Эти слова имеют одинаковое орфографическое написание (возможно, с точностью до ударения), но их отнесение к лингвистическим (грамматическим и

семантическим) категориям является неоднозначным. Например, слова *руки* и *ноги*, которые могут быть отнесены как к категории существительного родительного падежа единственного числа, так и именительного падежа множественного числа, а слово *мило* может принадлежать к одной из трех грамматических категорий.

Таблица 2

Разнообразие лингвистических категорий в грамматико-семантической плоскости

Часть речи	Количество форм (оценка)	Имена собственные, топонимы и их производные (всего 8)		Человеческая деятельность		Количество, мера (всего 14)	
		категории	форм	категории	формы	категории	формы
Сущ.	13	5	10	3	12	10	12
Прил.	20	5	20	3	24	10	24
Гл.	40	-	-	2	40	1	40
Числ.	18	-	-	-	-	5	15
Нескл.	1	5	1	3	1	10	1

Логично предположить, что омографы должны выделяться в отдельные классы согласно набору лингвистических категорий, который порождается неоднозначностью. Таких наборов на основе [8] выявлено несколько сотен.

Результаты кластеризации показали чувствительность к фонетическому наполнению слов. Например, в украинском языке перевод слова, обозначающего союз *и*, передается одним из трех слов: *і*, *й*, *та*. Будучи одинаковыми по смыслу, каждое из этих трех слов имеет регламентированное употребление в зависимости от фонетического окружения. Так, обычно союз *і* употребляется после слов, оканчивающихся на согласный, *й* ставится между словами, в конце и начале которых стоит гласный. Подобным образом обусловлено употребление ряда слов, относящихся как к служебным частям речи (*у/в/уві/вві*, *з/зі/із/зо*, *б/би*, *ж/же*), так и к существительным, прилагательным и глаголам (*вчений/учений*, *ітu/їти*). Однако для последних примеров разделение на классы с учетом фонетического признака не наблюдалось.

Таблица 3

Примеры автоматически сформированных кластеров

Слова	Частота	Слова	Частота	Слова	Частота
багато	134590	**які**	590681	**заявив**	163547
чимало	24482	котрі	24499	вважає	99803
безліч	7696	яки	465	повідомив	80043
немало	2191	**де**	246376	заявила	32795
якнайбільше	760	куди	31966	заявляє	31965
багацько	255	звідки	15373	розповів	30504
богато	123	звідкіль	120	говорить	29756

В табл. 3 приведено полное содержание нескольких классов за исключением последнего класса. Наиболее встречаемое слово в классе выделено полужирным шрифтом. Нетрудно заметить, что объединенные в

классы слова содержат общие грамматические и семантические признаки. Написанные с ошибками слова (*богато* и *яки*) были, тем не менее, отнесены к наиболее соответствующим классам. В примере последнего класса общими являются грамматические категории 3-го лица глагола и двух времен и родов. Кроме приведенных примеров, следует отметить достаточно четкое выделение в классы имен собственных и числительных по признакам, указанным в таблицах 1 и 2.

Выводы. Идея отнесения слов к классам и оперирования скорее классами слов, чем словами является конструктивным развитием лингвистического компонента систем распознавания речи. Впервые сформированные классы слов для украинского языка представленными в работе автоматическими средствами обобщают грамматические, семантические и фонетические признаки слов.

Углубленный анализ существующего алгоритма кластеризации показал возможность существенного ускорения его работы, что позволяет приступить к созданию лингвистических моделей славянских языков на основе текстовых корпусов, представляющих все разнообразие языка.

1. *Винцюк Т.К.*, Анализ, распознавание и смысловая интерпретация речевых сигналов. — Киев: Наукова думка, 1987. — 264 с.
2. *Сажок М.М., Селюх Р.А., Юхименко О.А.* Адаптація до голосу диктора на основі гендернозалежних акустичних моделей фонем для української мови // Десята Всеукраїнська міжнар. конф. УкрОбраз’2010. — Київ, 2010. — С.59–62.
3. *Робейко В., Сажок М.* Розпізнавання спонтанного мовлення на основі акустичних композитних моделей слів у реальному часі // Штучний інтелект. — Донецьк, 2012. — № 4. — С. 253–263.
4. *Whittaker E.W.D.* Statistical Language Modelling for Automatic Speech Recognition of Russian and English. PhD thesis. — Cambridge University, 2000. — 140 p.
5. *Gales M., Young S.* The Application of Hidden Markov Models in Speech Recognition // Foundations and Trends in Signal Proc. — 2007. — **1**(3). — P. 195–304.
6. The HTK Book Version 3.4 / S.J. Uoung, G. Evermann, M. Gales et al. — Cambridge University, 2006. — 360 p.
7. *Martin S., Liermann J., Ney H.* Algorithms for bigram and trigram word clustering // Proc. of Eurospeech. — Madrid, 1995. — Vol. 2, — 1293–1256 p.
8. *Широков В.А., Манако В.В.* Організація ресурсів національної словникової бази // Мовознавство. — 2001. — № 5. — С. 3–13.

Международный научно-учебный центр
информационных технологий и систем
НАН Украины и Министерства образования
и науки, молодежи и спорта Украины, Киев

Получено 27.11.2012