ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ СТАТИСТИЧЕСКИХ МЕТОДОВ РАСПОЗНАВАНИЯ РЕЧИ ДЛЯ ОПРЕДЕЛЕНИЯ БЛИЗОСТИ ЯЗЫКОВ

Анс БЕРЗИНЬ, Рига, Латвия
Рижский технический университет

Аннотация. *В докладе сделана попытка использования скрытых моделей Маркова, применяемых при распознавании речи, для определения степени близости устных языков. В отличие от распознавания речи, где фонограмма сначала разбивается на единицы речи, мы применяем методы с использованием скрытых марковских моделей к целым, продолжительным фонограммам разных информантов.*

Ключевые слова: *обработка естественного языка, компьютерная лингвистика, компьютерная диалектология, диалектометрия, распознавание речи, статистические методы, скрытые марковские модели, фонограмма, близость языков, расстояние, латыский, латгальский, наречия.*

Abstract. *In the article it is told about possibilities of use of hidden Markov models' speech recognition's methods for definition of distance between natural spoken languages .In difference from speech recognition, where sound tracks are broken into speech units, we apply hidden Markov models methods to the whole, long sound tracks of dif ferent informants.*

Keywords: *natural language processing, computational linguistics, computational dialectology, dialectometrics, speech recogni tion, statistical methods, hidden Markov models, recording, distance between languages, Latvian, Latgalian, tongues, dialects.*

ВВЕДЕНИЕ

На протяжении нескольких лет мы интересуемся возможностями автоматизированного измерения близости языков. Исследования ведутся на материале балтийских наречий, хотя эти методы применимы и к другим языкам. В 2004 году мы представили доклад на кон-ференции „Корпусная лингвистика“, в котором рассказывалось об измерении близости наречий по частотным спискам n-грамм [6], а в 2006 году – на конференции „Диалог“, в котором рассказывалось об определении расстояния на параллельном словаре путём использо-вания расстояний редактирования [7].

Оба упомянутых метода предназначены для работы на материале в фонетической транскрипции, т.е., уже расшифрованных фонограмм, причём расшифрованных в одинаковой записи, не зависящей от черт и взглядов расшифровщика. Формирование подобных материалов чрезвычайно трудозатратно, поэтому очевидна наша заинтересованность в создании метода, позволяющего работать на прямую с фонограммами. В данной статье мы расскажем о первом этапе подобной попытки, увенчавшимся успехом.

ПРИМЕНЕНИЕ СКРЫТЫХ МОДЕЛЕЙ МАРКОВА

В области распознавания речи широко применяются скрытые марковские модели (СММ). В данной статье мы не будем подробно останавливаться на идее применения скрытых марковских моделей, так как с ней можно ознакомиться в публикациях других авторов [4], в том числе и на русском языке [5], а всего лишь опишем, чем наша идея отличается от традиционно применяемой при распознавании речи.

Так как при распознавании речи базовой единицей, интересующей разработчика, является слово, то в качестве анализируемых речевых единиц, в зависимости от характера задачи, например, размера словаря или склоняемости языка, задаются либо сами слова, либо их сочетания, либо составляющие – фонемы или их сочетания. Т.е., в понимании распознавания речи, объектами описания скрытыми марковскими моделями являются вышеуказанные единицы. Модели обучаются на некотором количестве разных записей произнесения данных единиц, и таким образом, при сравнении распознаваемой единицы с заданными моделями, будет найдено наиболее вероятное соответствие.

В отличии от распознавания речи, в котором, в принципе, имеется интерес в транскрибировании устной речи, нас интересуют устные языки в целом – в широком смысле этого слова, т.е., в том числе и наречия. Поэтому логично было бы обьектами описания СММ вместо вышеуказанных единиц считать некие более протяжённые объекты, характеризующие язык в целом. В таком случае обучение СММ должно проводиться не на фонограммах данного слова (или набора, или составляющих), а на фонограммах языка в целом, т.е., вместо слова должна анализироваться фонограмма целиком. Конечно, независимость от говорящего должна достигаться теми же средствами, т.е., обучение модели должно проводиться на достаточно большом количестве фонограмм разных информантов-носителей языка. Но, так как нас интересует язык в целом, то мы предполагаем, что фонограммы, при достаточно больших их продолжительности и количестве, могут быть произвольными, т.е., дабы лучше всего отражать живой язык – экспедиционными записями спонтанной речи.

МАТЕРИАЛ

В принципе, описанные эксперименты применимы для любых человеческих языков в широком смысле этого слова, т.е., для языков и наречий, в том числе и не обладающих письменной формой. Но так как латышские и латгальские наречия нам более доступны, то мы эксперименты проводим на них.

В Латвии с 50-х по 80-е годы в рамках диалектологических экспедиций интенсивно собирались аудиозаписи наречий. Однако, качество данных записей настолько плохо, что даже на слух их воспринять очень трудно, а автоматизированный анализ ввиду высокого уровня помех вообще представляется невозможным.

В последние годы достаточно часто проводятся экспедиции другого профиля, например, по собиранию фольклора и жизнеописаний. Однако, собранные в таких экспедициях звукозаписи тоже не соответствуют нашим требованиям и по содержанию (собиратели в большинстве случаев разговаривают с информантами на литературном языке, таким образом поощряя их отвечать не на наречии, а на (ломаном) литературном), и по качеству (запись ведётся в сжатом формате, разными микрофонами, при наличии внешнего шума, голова информанта перемещается относительно микрофона и т.п.).

Поэтому, после продолжительного ознакомления с доступными материалами, мы пришли к выводу, что удовлетворяющий нас материал мы, к сожалению, можем собрать только сами. Осенью 2008 года мы провели несколько экспедиций в Латгалии и Курляндии. В результате был собран материал 4 латгальских и 1 курляндского наречий: 30 говорящих на вилекском, 23 – на бальтиновском, 29 – на рудзатском, 14 – на ауленском и 17 – на дундажском наречии. Все информанты рассказывали о своём жизненном пути: родителях, семье, школе, работе, женитьбе, детях, хозяйстве и т.п. Ввиду нехватки времени мы для подготовки этой публикации воспользовались лишь частью собранного материала, но в ближайшем будущем планируем провести эксперименты и на остальном материале.

ЭКСПЕРИМЕНТ

В рамках проверки идеи было проведено несколько экспериментов. Все они проводились, пользуясь программным пакетом HTK [1], т.е., у нас не было необходимости вникать в техническую реализацию алгоритмов в программном обеспечении. Суть эксперимента заключалась в создании нескольких скрытых марковских моделей на соответствующих наборах фонограмм и попытке распознавания вида других, неиспользованных при обучении, фонограмм, т.е., нахождении наиболее вероятностной близости по отношению к моделям.

Первый эксперимент был проведён над зачитанным одним и тем же человеком одинаковым текстом на трёх языках – латышском, латгальском и русском. На каждом языке были зачитаны четыре фонограммы – три в среднем темпе речи, а одна – в ускоренном. На данных фонограмм, зачитанных в среднем темпе, были созданы скрытые марковские модели для каждого языка. Далее при помощи утилиты HVite (реализация алгоритма Витерби в пакете HTK) проводился поиск наиболее близкой модели для записей ускоренного темпа. При небольшом числе компонентов Гауссовой смеси (мод) результаты были неудовлетворительными, но начиная с четырёх и выше, язык записи ускоренного темпа определялся безошибочно. Положительные результаты этого эксперимента подтолкнули нас провести второй, уже на настоящих, интересующих нас данных.

Из собранных нами данных мы выбрали два наречия – рудзатское и вилекское, т.е., оба латгальских, но с разных концов Латгалии – с Севера и с Запада. Таким образом рассматриваемые языки являлись очень близкими (что усиливает значение результатов в случае положительного исхода), но всё же мы могли рассчитывать, что их различия не будут смазаны ввиду социальных контактов носителей. Из каждого наречия мы случайным образом взяли по восемь информантов женского пола (из таких же соображений – дабы задача была более сложной), которых так же случайным образом поделили на два подмножества: по пять – для создания моделей, и по три – для проверки. Результаты оказались такими-же, как в предидущем эксперименте: при небольшом количестве мод язык определялся ошибочно (при чём по разному), а начиная с четырёх – безошибочно.

ВЫВОДЫ

Хотя данная работа находится лишь в начальной стадии, мы всё-таки решили доложить о положительных результатах, которые несомненно свидетельствуют о том, что надо продолжать работу в данном направлении.

Так как описанные эксперименты проводились вручную, то требуется создать программное обеспечение для автоматизации процесса проведения экспериментов на разных наборах фонограмм.

Кроме того, необходимо продолжить экспедиционную работу, чтобы иметь возможность работать с достаточно большим и разнообразным материалом.

Но самый главный вопрос – это определение расстояния между СММ, что в нашем случае тождественно определению расстояния между обьектами нашего исследования, т.е., языками – в данной публикации мы описали нахождение наиболее близкой модели, но не

задали расстояние между моделями как таковыми (по сути, мы должны были обучить модели на всех имеющихся данных и рассчитать расстояние между моделями). Наиболее часто используемое для оценки близости СММ расстояние Кульбака-Лейблера по сути расстоянием не является (поэтому некоторые его называют более корректно – расхождением), так как не является симметричным, т.е., не соответствует одной из аксиом метрики. В нашем случае симметричность расстояния существенно и мы не можем себе позволить её опустить. Известны попытки задания симметричного расстояния между СММ путём улучшения расстояния Кульбака-Лейблера, например [3], но наиболее интересным и новаторским является принципиально новое описание расстояния между скрытыми марковскими моделями, удовлетворяющего всем аксиомам метрики, опубликованное американскими учёными китайского происхождения Лю и Хуаном в 2000 году [2]. Это расстояние полностью удовлетворяет нашим требованиям, и мы планируем им воспользоваться. Но, так как программный код Лю и Хуана является собственностью частного предприятия и поэтому другим не доступен, то в ближайшем будущем планируем написать ПО по описанному ими алгоритму.

ЛИТЕРАТУРА

1. *Young S., Evermann G., Gales M., Hain Th., Liu X., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland Ph.* The HTK Book (for HTK Version 3.4). Cambridge: Cambridge University Engineering Department, 2009.

2. *Liu, Z., Huang, Q.* A new distance measure for probability distribution function of mixture type // IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. – Vol. 1 – P. 616-619.

3. *Johnson D., Sinanovic S.* Symmetrizing the Kullback-Leibler Distance. Computer and Information Technology Institute, Department of Electrical and Computer Engineering, Rice University. – Houston, 2001.

4. *Rabiner L.* A Tutorial on Hidden Markov Models ans Selected Applications in Speech Recognition. // Proceedings of the IEEE. – February 1989. – Vol. 77. – No 2 – P. 262-286.

5. *Кушнир Д.А.* Алгоритм формирования структуры эталона для пословного дикторонезависимого распознавания команд ограниченного словаря.// «Штучный інтелект» – № 3 – Київ, 2006.

6. *Берзинь А.У.* Сравнение балтийских языков методом n-грамм // Труды международной коференции „Корпусная лингвистика - 2004“. – СПб: Издательство С.-Петербургского университета, 2004. – С. 65-71.

7. *Берзинь А.У.* Измерение фономорфолексического расстояния между латышскими наречиями путём применения расстояния Вагнера-Фишера // Труды международной конференции «Диалог 2006». – М.: Издательство РГГУ, 2006. – С. 65-72.