DOI 10.37882/2223–2966.2021.04.40

МЕТОДЫ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА В ПРИЛОЖЕНИИ ДЛЯ ЯЗЫКОВОЙ ПРАКТИКИ[1]

NATURAL LANGUAGE PROCESSING TECHNIQUES IN A LANGUAGE PRACTICE APPLICATION[2]

A. Shiryaev
A. Kapitanov

Summary. Globalization affects each of us in one way or another. Communication with colleagues around the world is one of the factors that affects the acceleration of progress and the increase in advantages compared to those who are in an insurmountable language barrier. The main problems in learning a foreign language: internal fears for communicating with a living person; low efficiency in learning a language without fixing the material in practice; lack of time or opportunity to attend conversational clubs or courses for practice; loss of skills in the absence of constant practice. The main idea of creating an application for language practice is to improve communication skills in a foreign language during the discussion of the chosen topic. In this paper, we study the accuracy of the choice of learning paths for various methods of preprocessing text messages.

Keywords: NLP, Finite-state machine, machine learning, Glove.

Ширяев Алексей Павлович
Руководитель службы технической поддержки, ООО МУЛЬТИПАС
alex-sh2@yandex.ru
Капитанов Андрей Иванович
Ассистент, Национальный исследовательский университет «МИЭТ»
andrey@kapdx.ru

Аннотация. Глобализация в той или иной мере сказывается на каждом из нас. Коммуникация с коллегами по всему миру является одним из факторов, который влияет на ускорение прогресса и увеличение преимуществ по сравнению с теми, кто находится в непреодолимом языковом барьере. Основные проблемы при изучении иностранного языка: внутренние страхи при общении с живым человеком; низкая эффективность при изучении языка без закрепления материала на практике; отсутствие времени или возможности посещать разговорные клубы или курсы для практики; потеря навыков при отсутствии постоянной практики. Основная идея создания приложения для языковой практики — повышение коммуникативных навыков владения иностранным языком в процессе обсуждения выбранной темы. В рамках данной работы проводится исследование точности выбора траекторий обучения для различных методов предварительной обработки текстовых сообщений.

Ключевые слова: NLP, конечный автомат, машинное обучение, Glove.

Введение

По данным опроса Всероссийского центра изучения общественного мнения всего лишь 5% россиян свободно владеют английским языком [1]. При этом владение иностранным языком чаще всего помогает россиянам в следующих случаях:

1. Чтение инструкций и этикеток — 42%.
2. Чтение интернет-сайтов — 38%.
3. Обучение — 14%.

Среди основных проблем при изучении иностранного языка можно выделить следующие:

- стеснение уровнем владения иностранным языков и внутренние страхи при общении с живым человеком;
- низкая эффективность при изучении языка без закрепления материала на практике;
- отсутствие времени или возможности посещать разговорные клубы или курсы для практики;
- потеря навыков при отсутствии постоянной практики.

Основная идея создания приложения для языковой практики «Мой иностранный собеседник» — повышение коммуникативных навыков владения иностранным

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19–37–90144.
[2] Acknowledgments: The reported study was funded by RFBR, project number 19–37–90144.

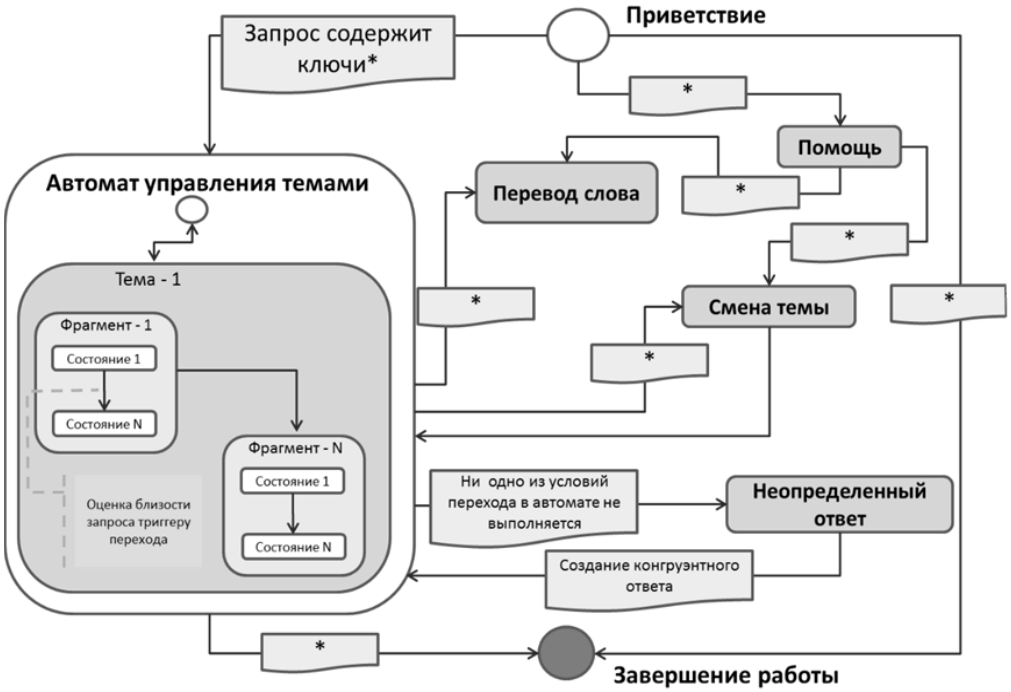

Рис. 1. Иерархический расширенный конечный автомат

языком в процессе обсуждения выбранной темы. Совмещение различных моделей на основе правил с моделями машинного обучения позволяет вести вариативные обучающие диалоги. А интерактивные диалоги в формате чата или голосового общения обеспечивают эффективное получение и применение навыков ведения беседы на иностранном языке.

Постановка и формализация задачи

Системы адаптивного обучения являются инструментом классификации обучающихся на различные когнитивные уровни. Однако, каждая система имеет уникальные алгоритмы адаптации, основанные на методике построения контента, стратегиях, анализе предпочитаемых стилей обучения, реакции системы на ответы пользователей, что позволяет формировать уникальные способы обучения.

Пусть система адаптивного обучения S формализована в виде конечного автомата [2, с. 1089]:

$$S = \{U, Y, X, x_0, \Lambda, H\},$$

где U — конечные набор входов $\{u_i\}_{i=1\dots I}$;
Y — конечные набор входов $\{y_i\}_{i=1\dots I}$;

X — конечный набор состояний $\{x_i\}_{i=1\dots I}$;
x_0 — начальное состояние системы;

Λ — набор функций перехода $U \times X \to X, \{\lambda_i\}_{i=1\dots I-1}$;
H — набор функций выхода $U \times X \to \mathrm{H}, \{\eta_i\}_{i=1\dots I}$.

Представленную модель можно интерпретировать следующим образом:

$\{u_i\}_{i=1\dots I}$ — ответы пользователя на поставленные задачи;
$\{y_i\}_{i=1\dots I}$ — реакция системы на ответы пользователя;
$\{x_i\}_{i=1\dots I}$ — когнитивный уровень в соответствии с таксономией Блума;
x_0 — начальный уровень знаний пользователя;
$\{\lambda_i\}_{i=1\dots I-1}$ — вопросы, задаваемые системой пользователю для перехода на другой уровень компетенций с генерацией соответствующей реакции;
$\{\eta_i\}_{i=1\dots I}$ — объяснения, предлагаемые системой на данном уровне компетенций.

Модель должна быть конкретизирована как иерархический конечный автомат [3, с. 29; 4, с. 254; 5, с. 1732], в первую очередь, с целью компактности представления.

Обработка естественного языка

Алгоритм общения с пользователем в чате приложения можно описать двумя конечными автоматами:

1. Работа с командами, выбор темы и т.д.
2. Перемещение в рамках темы + смена темы.

Таблица 1. Точность алгоритмов при различных начальных условиях

Условия * / Алгоритмы **	Точность алгоритма,%			
	№ 1	№ 2	№ 3	№ 4
A	89,09	80,83	68,41	59,65
B	96,97	77,46	87,60	83,76
C	95,72	81,09	87,31	81,13
D	x	64,05	x	37,08
E	85,91	80,83	68,91	51,78

Каждое сообщение пользователя проходит предварительную обработку. В процессе обработки из текста производится удаление неинформативных частей (*графематический анализ*). Для этого используется список стоп-слов, не представляющих ценности при данном типе обработки: местоимения, предлоги, союзы, междометия и пр. Наряду с графематическим анализом также выделяются: лексический, морфологический, синтаксический, семантический и прагматический. Также одним из этапов предобработки текста является *noun chunking*, или выделение «базовых словосочетаний», состоящих из существительного и определяющих его слов, таких как артикль и прилагательные.

После предварительной обработки оценивается вероятность перехода (по набору слов-триггеров), далее на основе косинусной меры выполняется переход по наиболее близкому набору. Диалог разбивается на фрагменты, которые состоят из шаблонов (*конечный автомат*); если пользователь спрашивает или говорит о чем-то за пределами заранее заданных тем, то сервис отвечает уточняющим вопросом. Передача управления между автоматами осуществляется на основе определения языка запроса.

Реализованы 4 состояния высокого уровня: «начало», «конец», «тема» и «помощь». У каждой темы существуют состояния низкого уровня. На рисунке представлена схема работы алгоритма в виде иерархического расширенного конечного автомата (рис. 1).

Использование данной схемы позволяет находить оптимальные переходы для формирования уникальных траекторий обучения.

Эксперимент. Был проведен следующий эксперимент: генерировался массив сообщений (предложений) и с помощью экспертной оценки размечались возможные варианты переходов.

В таблице 1 описаны результаты точности перехода в ожидаемое состояние в зависимости от начальных условий (предобработки текста) и выбранного алгоритма.

x — выбранный алгоритм не предусмотрен для обработки базовых словосочетаний.

*** Расшифровка условий:**

A. Коррекция грамматики, токенизация и удаление знаков препинания, лемматизация, удаление стоп-слов, noun chunking.

B. Токенизация и удаление знаков препинания, лемматизация.

C. Токенизация и удаление знаков препинания.

D. Токенизация и удаление знаков препинания, noun chunking.

E. Токенизация и удаление знаков препинания, удаление стоп-слов, noun chunking.

**** Расшифровка алгоритмов:**

№ 1. Векторные представления Glove (KeyedVectors) + косинусная мера.

№ 2. Wordnet + косинусная мера.

№ 3. Doc2vec + косинусная мера.

№ 4. Glove + матрица расстояний + косинусная мера.

Выводы

По результатам моделирования наиболее эффективные методы предобработки сообщений основываются на токенизации и удалении знаков препинания. При этом дополнительная обработка в виде лемматизации и удалении стоп-слов сильно зависит от выбранного алгоритма.

В последующих работах планируется провести сравнение иерархического расширенного конечного автомата и метода автоматического анализа: деревья принятия решений.

ЛИТЕРАТУРА

1. ВЦИОМ. Новости: Иностранный язык: перспективная инвестиция? [Электронный ресурс] — Режим доступа: https://wciom.ru/analytical-reviews/analiticheskii-obzor/inostrannyj-yazyk-perspektivnaya-investicziya
2. Прохоров С.А., Куликовских И.М. Система адаптивного обучения на основе иерархических конечных автоматов // Известия Самарского научного центра Российской академии наук. 2015. Т. 17, № 2–5.
3. Кузьмин Е.В. Иерархическая модель автоматных программ // Моделирование и анализ информационных систем. 2006. Т. 13, № 1.
4. Lemch E.S., Caines P.E. On the existence of hybrid models for fi nite state machines // Systems & Control Letters. 1999. Vol. 36.
5. Spinke V. An object-oriented implementation of concurrent and hierarchical state machines // Information and Software Technology. 2013. Vol. 55.

Журнал «Современная наука: актуальные проблемы теории и практики»

МИЭТ