

УДК 81'322

НЕКОТОРЫЕ ПРИНЦИПЫ АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ УЧЕБНЫХ МАТЕРИАЛОВ НА ОСНОВЕ ДВУХУРОВНЕВЫХ СЕМАНТИЧЕСКИХ ШАБЛОНОВ

Личаргин Д.В., Николаева Н.В., Чубарева Е.Б.

ФГАОУ ВПО «Сибирский федеральный университет», Красноярск, e-mail: lichdv@hotmail.ru

В работе предлагается модель добавления шаблонов семантической декомпрессии на основе принципов добавления семантического шума на основе двухуровневых шаблонов. Рассмотренные принципы позволяют обеспечить выполнение алгоритмов добавления шаблонов семантической декомпрессии в целях порождения более широкого множества фраз естественного языка, например, в форме учебных заданий с применением лингвистического программного обеспечения. Метод представления данных в форме многоуровневых шаблонов позволяет генерировать фразы естественного языка на основе слов, соответствующих векторам семантических признаков, описываемых в работе, что позволяет использовать критерий их соответствия, определяющих осмысленность конечных фраз процесса генерации. Многоуровневые шаблоны развиваются на различные по результатам генерации стилистические подмножества. Результаты апробированы на основе программы генерации учебных заданий, имеющей целью повышение эффективности подготовки материалов к учебным занятиям. Делается вывод о необходимости продолжения данных исследований, в частности на основе многоуровневых шаблонов с большей мощностью генерируемых подмножеств естественного языка.

Ключевые слова: генерация естественного языка, семантические признаки, порождение естественного языка, генерация осмысленных предложений

SOME PRINCIPLES OF AUTOMATIC GENERATION OF LEARNING MATERIALS BASED ON TWO-LEVEL SEMANTIC TEMPLATES

Lichargin D.V., Nikolaeva N.V., Chubareva E.B.

Siberian Federal University, Krasnoyarsk, e-mail: lichdv@hotmail.ru

The paper proposes a model of adding semantic decompression templates, based on the principles of adding semantic noise on the basis of two-level templates. The considered principles allow execution of the algorithms of adding semantic decompression pattern for generating a wider set of phrases of a natural language, for example in the form of learning tasks using linguistic software. The method of data presentation in the form of multi-level templates to generate sentences of the natural language based on words that correspond to the vectors of semantic features, described in this work that allows using the criterion of their correspondence, defining meaningfulness of the end phrases in the generation process. The multi-level patterns are separated into different stylistic subsets as a result of the generation. The results were tested based on the educational tasks generation software, which aims to improve the effectiveness of training materials for the training sessions. The conclusion is that it is needed to continue these studies, in particular, on the basis of multilevel templates with bigger power of the subsets of the generated natural language.

Keywords: the generation of natural language, semantic features, natural language generation, the generation of meaningful sentences

В настоящее время является актуальной проблема автоматизации систем письменного и устного перевода для различных языков, экспертных, поисковых систем и систем реферирования. Для ее решения существуют многочисленные теории, концепции и программные системы, ведутся работы в области семантики, дискретной математики, лингвистики и искусственного интеллекта. На современном этапе также важно решить проблему автоматической генерации учебных материалов как частного случая текстов на естественном языке и проблему создания осмысленных подмножеств естественного языка в различных приближениях. Все это позволит более эффективно решать задачи по построению экспертных систем, систем электронного обучения, систем автоматиче-

ского перевода, программ для поддержки диалога с пользователями, созданию естественно-языковых интерфейсов.

Для решения проблемы генерации осмысленной речи на сегодня используется широкий инструментарий, как семантики, так и искусственного интеллекта в рамках понятийного аппарата и различных моделей математической семантики. В частности, для анализа естественного языка традиционно используются модели и средства, такие как метод онтологий, метод лингвистической классификации, метод многомерного представления данных, OLAP системы, реляционные базы данных, фреймы, инструментарий системного анализа. Также используются порождающие грамматики, в частности порождающие грамматики Монтегю и грамматики сложения деревьев,

семантические сети, теория графов и метод резолюций, гибридные системы, а также лингвистические методы, такие как компонентный анализ, валентностное представление слов языка, парадигматический метод, методы американского структурализма и др. [1–6].

Новизна данной работы состоит в описании способов сохранения осмысленности фраз на естественном, в частности английском языке при их трансформациях с применением двухуровневых шаблонов семантических данных.

Основная идея работы состоит в принципе генерации подмножества фраз естественного языка в виде двухуровневых шаблонов, которым соответствует ранжирование фраз естественного языка по критерию добавления семантического шума.

Цель работы – описать принцип генерации текстов для учебных курсов по иностранному, в частности английскому языку. Задачи работы состоят в применении данной модели на основе принципа генерации осмысленной речи на основе двух- и (в перспективе) многоуровневых шаблонов и определение места этого принципа в системе языка при анализе принципов добавления и экстракции семантического шума.

При решении проблемы обработки реальных текстов на некотором языке, в частности программами-генераторами и анализаторами осмысленной речи, необходимо решить проблему построения шаблонов семантической декомпрессии. Семантические шаблоны декомпрессии могут оперировать информацией для построения текстов, написанных в разных стилях языка: от академического стиля с низкой степенью семантического шума до сленга с чрезвычайно высоким уровнем добавления семантического шума. Семантические шаблоны декомпрессии являются формальными описаниями, соответствующими выражению эмоций и пониманию глубины предмета носителем языка. Компьютер, в частности, может считать это несущественным для задачи генерации осмысленных единиц естественного языка: слов (например, неологизмов), предложений и текстов.

Для генерации предложений с использованием шаблонов семантической декомпрессии можно также использовать метод составления семантических шаблонов декомпрессии разработчиками посредством анализа смысловых структур понятий для их последующей трансформации.

Слова естественного языка могут быть представлены в форме векторов семантиче-

ских признаков, например слово «любить» соответствует вектору признаков:

[ОТНОШЕНИЕ-СУЩЕСТВО-X \ СУЩНОСТЬ \ ПОЗИТИВНОСТЬ]

Слово «красивый» соответствует вектору признаков:

[ОТНОШЕНИЕ \ СУЩНОСТЬ \ ОТНОШЕНИЕ-СУЩНОСТЬ-X \ ИДЕЯ \ НА (НЕ) ЖИВОМ \ ПОЗИТИВНОСТЬ].

Слово «смотреть» соответствует вектору семантических признаков:

[ОТНОШЕНИЕ-СУЩЕСТВО-X \ СУЩНОСТЬ \ ОТНОШЕНИЕ-СУЩЕСТВО-X \ ИДЕЯ \ НА (НЕ) ЖИВОМ \ ПОЗИТИВНОСТЬ].

В результате можно использовать возможность для перегруппировки сем естественного языка в семантической сети для каждого из слов на основе определенного вида трансформаций. Например, предложение «I cook dinner» может быть трансформировано во фразу («My cooking dinner» ... & «The dinner being cooked by me ...» & «It was ... for me to cook dinner» & «My dinner after cooking ...»). Ниже представлен пример параллельной генерации по другой теме: «We build the museum» → («Our building the museum ...» & «The museum being built by us ...» & «It was ... for us to build the museum» & «Our museum after building ...»), аналогично «They listen to the music» → («Their listening to the music ...» & «Music being listened to by them ...» & «It was ... for them to listen the music» & «Their music after listening ...»).

На рисунке представлена общая схема приведения литературной фразы сложного вида к упрощенному синонимическому эквиваленту посредством добавления логического, семантического, грамматического, морфологического и других шаблонов добавления шума / шаблонов декомпрессии.

Предполагается, что этот подход может использоваться для создания различных видов лингвистического программного обеспечения, например для систем реферирования, систем электронного перевода, экспертных систем, систем извлечения данных из текстов на естественном языке и т.п. Сами правила таких трансформаций, имеющие вид двухуровневых шаблонов, описываемых ниже, делятся на подмножества в соответствии со стилем изложения.

- | | |
|--|--|
| 1. Обиходный стиль;
1.1. Сленг;
1.1.1. Табуированный подъязык;
1.1.2. Уголовное арго;
1.2. Нейтральный стиль;
1.3. Игра слов, юмор;
1.4. Стиль СМИ;
1.5. ...
2. Художественный стиль;
2.1. Поэзия;
2.2. Проза;
2.1. Сказка;
2.2. ... | 3. Научный стиль;
3.1. Академический;
3.2. Общенаучный;
3.3. Научно-популярный;
3.4. ...
4. Религиозный стиль;
4.1. Православие;
4.2. Буддизм;
4.3. Ислам;
4.4. ...
5. Нейтральный стиль;
5.1. Публицистический стиль;
6. Смешанный стиль. |
|--|--|



Схема генерации осмысленных учебных заданий к текстам

Приведем пример дерева генерации предложений с семантическим шумом, одной из составляющих декомпрессии предложений естественного языка.

1. «The driver carries the cans...» – «Водитель привозит консервы...».

1.1. «The cans carried by the driver...» – «Консервы, привозимые водителем...».

1.2. «The driver, who carries the cans...» – «Водитель, который привозит консервы...».

1.3. «Carrying the cans by the driver...» – «Перевозка консервов водителем...».

1.3.1. «Carrying the cans from the side of the driver...» – «Перевозка консервов со стороны водителя...».

1.3.2. «Carrying the cans by the efforts of the driver...» – «Перевозка консервов усилиями водителя...».

1.3.2.1. «The process of carrying the cans by the efforts of the driver...» – «Процесс перевозки консервов усилиями водителя...».

1.3.2.2. «The task of carrying the cans by the efforts of the driver...» – «Задача перевозки консервов усилиями водителя...».

1.3.2.3. «Carrying the goods, for example, cans by the efforts of the driver...» – «Перевозка товаров, на пример, консервов усилиями водителя...».

1.3.2.3.1. «Carrying the consumer goods – cans by the effort of the driver...» – «Перевозка потребительских товаров – консервов усилиями водителя...».

1.3.2.3.2. «Carrying the goods – round cans by the effort of the driver...» – «Перевозка товаров – круглых консервов усилиями водителя...».

1.3.2.3.3. «Carrying the goods – metal cans by the effort of the driver...» – «Перевозка товаров – металлических консервов усилиями водителя...».

Далее рассмотрим шаблоны трансформации фраз приведенного вида в направлении декомпрессии. В связи с тем, что обработка глубоких эмоциональных коннотаций является очень трудной задачей для лингвистического программного обеспечения, задача создания программного обеспечения, алгоритмы и подходы для автоматической генерации учебных заданий выбраны с учетом цели применения шаблонов для добавления семантического шума. Этот принцип целесообразно использовать в системах электронных учебных курсов. Подобная задача была реализована на основе принципов языковой комбинаторики, на сегодняшний момент такие системы генерации учебных заданий находятся в процессе

усовершенствования. Один из последующих шагов должен заключаться в использовании шаблонов второго уровня для генерации естественного языка.

Таким образом, шаблоны второго уровня могут быть использованы, например, для генерации учебных заданий, основанных на принципах сочетаемости слов: «the user + takes + some wire» (Тема: Детали оборудования, Позиция в предложении: Делатель + Действие + Принимающий действие, Варианты подстановок:

user's taking the wire was necessary» или «the wire after taking by the user was given to me» (таблица). Могут быть сгенерированы соответствующие задания на основе декомпрессированных предложений в неприведенной форме, например «the wire after taking / eating / wearing / living by the user was given to me». Необходимо пополнять существующие базы данных семантических шаблонов второго уровня для применений в области лингвистического программного обеспечения для электронного обучения.

Шаблоны компрессии-декомпрессии текста

Doer	Action	Object	Substance
I	Eats	The ...	With / without ...
We	Cooks	Dish	Beef
Bob	Roasts	Potatoes	Fish
I	Sews	The ...	From ...
They	Knits	Jacket	Wool
The girl	Irons	Shirt	Cotton

Subject	Predicate	Object	Modifier
My DOER's	ACTION.MAKING-ing	Needs / requires / ...	Good / nice / ... + SUBSTANCE
My cook's	roasting	needs	(good) beef
My mother's	sewing	refers to	(brilliant) silk
This / the / the given + SUBSTANCE	Is good / nice / ideal / ... for	For my / his / her / ... + DOER	To + ACTION
Silk	Is good for	My mother	To sew
Fish	Is ideal for	My brother	To cook
ACTION-ing	Cannot / will not + go on / continue / be done / be all right	–	Without + such / this / ... > like this + SUBSTANCE
Cooking	Will not be done	–	Without beef
Knitting	Will not be all right	–	Without wool

Пассивный + Получение + Металлический – Длинный – Гибкий); на основе такой фразы может быть сформировано учебное задание: «the user + takes / eats / wears / lives in + some wire». Для генерации учебных заданий на основе результата такой генерации может быть дан следующий контекст: «выбрать правильное слово из списка» или «заполнить пробелы с одним из вариантов, предложенных ниже» и т.д.

Но также может быть выполнена дальнейшая декомпрессия фразы посредством добавления семантического шума: «the

В заключение необходимо отметить, что возможно и целесообразно применение многомерной модели естественного языка и классификации слов естественного языка [1] для классификации стилистических структур использованием множества реляционных шаблонов пословного построения предложений, используемых для генерации естественного языка. В работе была проанализирована структура такого рода правил. Последние могут быть использованы для компрессии и декомпрессии предложений и текста языка. Были

предложены методы использования декомпрессии на основе речевых шаблонов второго уровня для порождения языка в целях улучшения лингвистического программного обеспечения для автоматической генерации учебных заданий, например, на уроках иностранного языка. Классификация понятий естественного языка может служить источником лексических единиц, для составления шаблонов генерации осмысленных текстов, которые можно усложнять посредством добавления семантического шума на основе расширенных порождающих грамматик над лесом строк и деревьев разнородных данных. Необходимо дальнейшее исследование в этой области.

Список литературы

1. Личаргин Д.В. Многомерное представление данных по лексике и грамматике английского языка / Сибирский федеральный университет. – Красноярск, 2010.

2. Личаргин Д.В., Бачурина Е.П. Обобщенная иерархическая структура учебного электронного курса и рассмотрение на ее основе электронных курсов обучения английскому языку // Информатизация образования и науки. – 2012. – № 15. – С. 20–36.

3. Личаргин Д.В., Таранчук Е.А. Иерархическая структура учебного электронного курса и ее вариабельность для обучения иностранному языку // Дистанционное и виртуальное обучение. – 2011. – № 4. – С. 56–75.

4. Сафонов К.В., Личаргин Д.В. Некоторые принципы автоматической генерации учебных материалов на основе баз знаний и лингвистической классификации // Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева. – 2012. – № 2 (42). – С. 72–77.

5. Сафонов К.В., Личаргин Д.В. Разработка векторизованной семантической классификации над словами и понятиями естественного языка // Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева. – 2009. – № 4 (25). – С. 33–37.

6. Safonov K.V., Lichargin D.V. Elaboration of a vector based semantic classification over the words and notions of the natural language // Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева. – 2009. – № 5 (26). – С. 52–56.