

ИСПОЛЬЗОВАНИЕ ДВУХУРОВНЕВЫХ ШАБЛОНОВ ГЕНЕРАЦИИ ОСМЫСЛЕННОЙ РЕЧИ

Цепкова М.В., Личаргин Д.В., Николаева Н.В.,
научный руководитель канд. техн. наук Личаргин Д.В.
Сибирский федеральный университет

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 выполнить задачи по построению экспертных систем, систем электронного обучения, систем автоматического перевода, программ для поддержки диалога с пользователями, созданию естественно-языковых интерфейсов.

Данная работа посвящена нахождению понятийного описания единиц естественного языка, способу задания и нахождения критериев осмысленности фраз на естественном, в частности, английском языке. Описываются способы представления учебных заданий и текстов на основе деревьев генерации учебных заданий и текстов. Новым понятием являются шаблоны добавления и экстракции семантического шума, т.е. декомпрессии и компрессии предложений соответственно.

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 семантической декомпрессии до сленга с чрезвычайно высокой степенью семантической декомпрессии – говоря упрощенно, с высоким уровнем семантического шума. Семантические шаблоны декомпрессии являются формальными описаниями, соответствующими выражению эмоций и пониманию глубины предмета носителем языка. Компьютер, в частности, может считать, что 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 трансформации. Например, слово «любить» соответствует вектору признаков:

[ОТНОШЕНИЕ-СУЩЕСТВО-Х \ СУЩНОСТЬ \ ПОЗИТИВНОСТЬ]

Слово «красивый» соответствует вектору признаков:

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

В результате, можно использовать возможность для перегруппировки сем естественного языка в семантической сети для каждого из слов. Например, фраза «The apple is beautiful» может быть преобразована во фразу «I like the form of the apple». В этом случае, понятие «beautiful» относится к группе сем со значением «see» и группе

сем общим для слов со значением «well», «love to» или «good», т.е. семей позитивности.

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 ...»).

Порождающей грамматике или реляционной шаблоны, используемые для семантической компрессии и декомпрессии предложений на основе трансформации по шаблону, могут быть разделены на различные подклассы, как показано далее в иерархическом списке.

1. Обиходный стиль;
 - 1.1. Сленг;
 - 1.1.1. Табуированный подязык;
 - 1.1.2. Уголовное аргю;
 - 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. ...

Выше представлена общая схема приведения литературной фразы сложного вида к упрощенному синонимическому эквиваленту посредством добавления логического, семантического, грамматического, морфологического, звукового 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 других.

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. Это очень полезно 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 (см. таблицу 1). 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: Пассивный + Получение + Металлический – Длинный - Гибкий); на основе такой фразы может быть сформировано учебное задание: «the user + takes / eats / wears / lives in + some wire». Может быть дано соответствующее задание: «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 user's taking the wire was necessary» или «the wire after taking by the user was given to me» (см. таблицу 1). Могут быть сгенерированы соответствующие задания на основе декомпрессированных предложений в не приведенной форме, например, «the wire after taking / eating / wearing / living by the user was given to me». Необходимо создать базы данных семантических шаблонов второго уровня для применений в области лингвистического программного обеспечения для электронного обучения.

Таблица 1

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

В заключении необходимо отметить, что возможно и целесообразно применение многомерной модели естественного языка, в частности, семантический векторизованной классификации слов и понятий естественного языка для классификации стилистических структур на основе множества правил порождающей грамматики или с 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.