

Личаргин Дмитрий Викторович, Чубарева Елена Борисовна, Николаева Наталья Владимировна
**РЕШЕНИЕ ПРОБЛЕМЫ ВАРИАТИВНОСТИ ПРИ ГЕНЕРАЦИИ ФРАЗ ЕСТЕСТВЕННОГО ЯЗЫКА
НА ОСНОВЕ ДВУХУРОВНЕВЫХ ШАБЛОНОВ**

В статье описана модель добавления шаблонов семантической декомпрессии. Освещается проблема моделирования естественных языков. Разработанная модель позволяет выполнять алгоритмы добавления шаблонов семантической декомпрессии. Данные шаблоны должны разнообразить генерацию учебных заданий лингвистическим программным обеспечением. В качестве формы представления грамматики и семантики слов используется вектор грамматических и семантических признаков слова. Необходима разработка базы данных шаблонов уровня порождения и уровня усложнения фраз языка.

Адрес статьи: www.gramota.net/materials/2/2013/11-1/31.html

Источник

Филологические науки. Вопросы теории и практики

Тамбов: Грамота, 2013. № 11 (29): в 2-х ч. Ч. I. С. 110-112. ISSN 1997-2911.

Адрес журнала: www.gramota.net/editions/2.html

Содержание данного номера журнала: www.gramota.net/materials/2/2013/11-1/

© Издательство "Грамота"

Информация о возможности публикации статей в журнале размещена на Интернет сайте издательства: www.gramota.net

Вопросы, связанные с публикациями научных материалов, редакция просит направлять на адрес: voprosy_phil@gramota.net

УДК 81`373.612

Филологические науки

В статье описана модель добавления шаблонов семантической декомпрессии. Освещается проблема моделирования естественных языков. Разработанная модель позволяет выполнять алгоритмы добавления шаблонов семантической декомпрессии. Данные шаблоны должны разнообразить генерацию учебных заданий лингвистическим программным обеспечением. В качестве формы представления грамматики и семантики слов используется вектор грамматических и семантических признаков слова. Необходима разработка базы данных шаблонов уровня порождения и уровня усложнения фраз языка.

Ключевые слова и фразы: генерация естественного языка; семантические признаки; классификации слов и понятий языка; генерация учебных материалов; порождение естественного языка; генерация осмысленных предложений.

Личаргин Дмитрий Викторович**Чубарева Елена Борисовна****Николаева Наталья Владимировна***Сибирский федеральный университет, г. Красноярск**lichdv@hotmail.ru; chubareva08@mail.ru; nikolnat@mail.ru*

РЕШЕНИЕ ПРОБЛЕМЫ ВАРИАТИВНОСТИ ПРИ ГЕНЕРАЦИИ ФРАЗ ЕСТЕСТВЕННОГО ЯЗЫКА НА ОСНОВЕ ДВУХУРОВНЕВЫХ ШАБЛОНОВ[©]

Проблема автоматизации систем устного и письменного перевода является актуальной для любого языка. Для ее решения ведутся работы в области семантики, лингвистики, дискретной математики и искусственного интеллекта. Автоматическая генерация учебных материалов и создание осмысленных подмножеств естественного языка позволят выполнить построение экспертных систем, систем электронного обучения, систем автоматического перевода.

Уровень разработки. Для решения проблемы генерации осмысленной речи 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, инструментарий системного анализа. Также используются порождающие грамматики и грамматики сложения деревьев, семантические сети, гибридные системы, компонентный анализ, парадигматический метод и др. [1-11; 13].

Основная идея данной работы состоит в представлении единиц естественного языка в виде множества векторов признаков, которым соответствует многомерное пространство векторизованных данных. Важно задать иерархию единиц языка и понятийное векторное описание для каждого из них, предоставить классификацию слов и основанную на ней классификацию предложений естественного языка. Слова выбираются из этого пространства на основе шаблонов генерации учебных заданий и текстов. Необходимо описать, как используются шаблоны компрессии и декомпрессии предложений.

Цель работы – построить модель генерации фраз и текстов для учебных курсов по английскому языку. **Задачи работы** заключаются в использовании классификации семантических категорий и слов естественного языка на основе шаблонов лингвистических трансформаций.

Для генерации семантически и грамматически осмысленных фраз естественного языка необходимо построить классификацию предложений естественного языка в виде многомерного пространства данных. Данная классификация является пересечением лексического и грамматического пространства слов – точек понятийного пространства. Грамматическое подмножество этой классификации имеет определенные координаты вектора классификации, например:

[Части речи {«Артикль», «Прилагательное», «Существительное», «Глагол»...},

Члены предложения {«Определитель», «Определение», «Подлежащее», «Сказуемое»...};

Категории. Лица {«1-е», «2-е», «3-е», «Не определено»},

Категории. Аспект {«Неопределенный», «Продолженный», «Совершенный», «Совершенный продолженный», «Не определен»};

Категории. Степень сравнения {...} и т.п.].

Грамматические конструкции включаются в ячейки многомерного массива. Пересечение таких координат вектора, как, например, V[ГЛАГОЛ / ОПРЕДЕЛЕНИЕ / СОВЕРШЕННЫЙ / ...], определяет ячейку многомерного массива с грамматической конструкцией: «having + ГЛАГОЛ-ed» Вектор V[Прилагательное / Сказуемое / 1-е лицо / Превосходная степень / Длинное слово / ...] = «am the most + ПРИЛАГАТЕЛЬНОЕ».

Лексическое подмножество этой классификации имеет определенные координаты вектора классификации:

[Порядок слов и члены предложения {Субъект, Предикат, Объект...},

Объекты по тематике изучения {идеи {науки, представления, чувства...}, предметы {одежда, еда, части тела, здания, транспорт...}, существа...},

Варианты подстановок слов в предложение {{позитивное, нейтральное, негативное}, {...}}].

Такое многомерное пространство включает комбинаторно сочетающиеся группы слов. Например, группа слов {варить, жарить, тушить, пробовать, подавать...} относится к ячейке многомерного пространства [ОТНОШЕНИЕ – СУЩЕСТВО – ОБЪЕКТ / ПИЩА / СОЗДАНИЕ]. При этом группа слов {морковь, капуста, свекла, картофель, лук...} относится к ячейке многомерного массива [ОБЪЕКТ / ПИЩА], что дает порождаемые фразы «варить картофель», «жарить лук», «тушить капусту» и т.п. Данное пространство становится критерием семантической и грамматической осмысленности речи. Функции определенного вида, определенной геометрии над данными группами слов образуют осмысленные фразы с хорошей вероятностью [1].

При решении проблемы обработки реальных текстов на естественном языке, например, программами-генераторами и анализаторами осмысленной речи важно 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, в частности с высоким уровнем семантического шума. Семантические шаблоны декомпрессии являются формальными описаниями, соответствующими выражению эмоций и пониманию глубины предмета носителем языка. В частности, компьютер 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 подходит метод составления семантических шаблонов декомпрессии, разработанный посредством анализа смысловых структур понятий для их трансформации.

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 цепочке векторов признаков слов для каждого из слов. Например, фраза «The flower is beautiful» может быть преобразована во фразу «I like the colour of the flower». В этом случае понятие «beautiful» относится к группе сем с внутренним значением, соответствующим слову «see» и элементам его значений в форме слов «well», «love to» или «good», т.е. семой позитивности.

Например, предложение «They are building the bridge» может быть трансформировано во фразу «Their building the bridge...», and «The bridge being built by them ...», and «It was... for them to build the bridge», and «The bridge after their building...».

В связи с тем, что обработка глубоких эмоциональных коннотаций является очень сложной задачей для лингвистического программного обеспечения, принципы автоматической генерации учебных заданий выбраны с учетом необходимости применения шаблонов для добавления семантического шума [10]. 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 + Реципиент, Варианты подстановок слов с их признаками: 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. 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».

Могут быть сгенерированы соответствующие задания на основе декомпрессированных предложений, например «the wire after taking / eating / wearing / living by the user was given to me». Необходимо создание базы данных семантических шаблонов второго уровня для применений в области лингвистического программного обеспечения для электронного обучения.

В заключение необходимо отметить, что возможно и целесообразно применение многомерной модели естественного языка, в частности, семантической 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 английского 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.

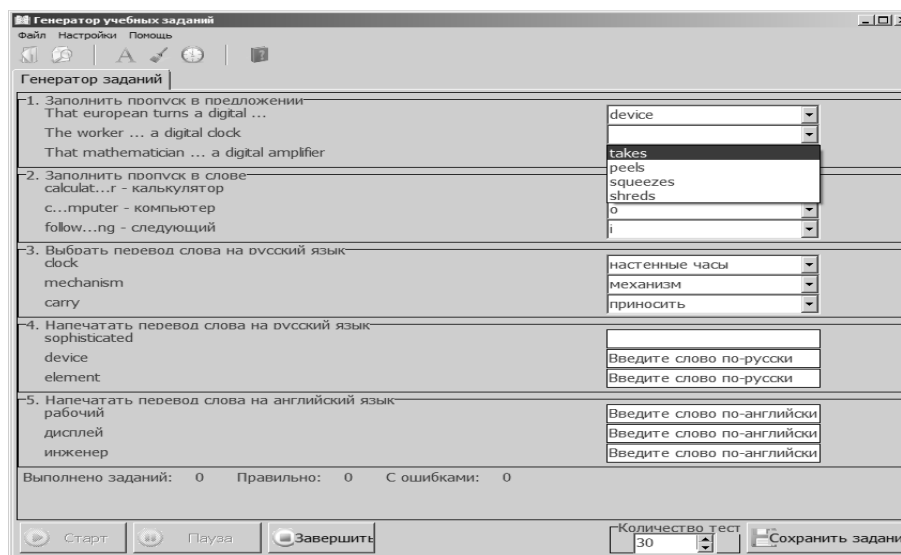


Рис. 1. Программное обеспечение для автоматической генерации учебных заданий

Список литературы

1. Агамджанова В. И. Контекстуальная избыточность лексического значения слова. М.: Высшая школа, 1977. 150 с.
2. Алексева И. С. Текст как доминанта перевода // Журнал СФУ. Красноярск: Издательство СФУ, 2011. Т. 4. № 10. С. 1375-1384.
3. Апресян Ю. Д. Идеи и методы современной структурной лингвистики. М.: Наука, 1966. 301 с.
4. Вердиева З. Н. Семантические поля в современном английском языке. М.: Высшая школа, 1986. 119 с.
5. Зыкова Т. В., Кытманов А. А., Цибульский Г. М., Шершнева В. А. Обучение математике в среде Moodle на примере электронного обучающего курса // Вестник Красноярского государственного педагогического университета имени В. П. Астафьева. 2012. № 1. С. 60-63.
6. Личаргин Д. В. Методы и средства порождения семантических конструкций естественно языкового интерфейса программных систем: дисс. ... к.т.н. Красноярск, 2004. 154 с.
7. Личаргин Д. В. Операции над семами слов естественного языка в машинном переводе // Труды конференции молодых ученых. Красноярск: ИВМ СО РАН, 2003. С. 23-31.
8. Личаргин Д. В. Порождение дерева состояний на основе порождающих грамматик над деревьями строк // Вестник СибГАУ. 2010. № 1. С. 57-58.
9. Личаргин Д. В., Бачурина Е. П. Обобщенная иерархическая структура учебного электронного курса и рассмотрение на ее основе электронного курса обучения английскому языку РИЯ ИКИТ СФУ // Информатизация образования и науки. 2012. № 3. С. 20-36.
10. Личаргин Д. В., Таранчук Е. А. Иерархическая структура учебного электронного курса и его вариабельность для обучения иностранному языку // Дистанционное и виртуальное обучение. 2011. № 4. С. 56-75.
11. Личаргин Д. В., Суманеева Я. А., Юрьева Е. В. Метод подстановочных таблиц и его применение в сфере обучения русскому языку для иностранцев // Вестник Сургутского государственного педагогического университета. 2012. № 6. С. 179-187.
12. Поветкина Ю. В. Моделирование как метод лингвистического исследования // Филологические науки. Вопросы теории и практики. Тамбов: Грамота, 2012. № 6 (17). С. 132-136.
13. Сдобников В. В. Новый взгляд на стратегию перевода: коммуникативно-функциональный подход // Журнал СФУ. Красноярск: Издательство СФУ, 2011. Т. 4. № 10. С. 1444-1453.

SOLVING PROBLEM OF VARIATION WHEN GENERATING NATURAL LANGUAGE PHRASES BASED ON TWO-LEVEL PATTERNS

Lichargin Dmitrii Viktorovich
 Chubareva Elena Borisovna
 Nikolaeva Natal'ya Vladimirovna
 Siberian Federal University, Krasnoyarsk
 lichdv@hotmail.ru; chubareva08@mail.ru; nikolnat@mail.ru

In the article a model of adding semantic decomposition patterns is described. The article is devoted to the problem of natural language modeling. The developed model provides the execution of algorithms for semantic decomposition patterns addition. The patterns allow diversifying the generation of educational tasks by linguistic software. A vector of word grammar and semantic features is used as a form of presenting the words grammar and semantics. The development of patterns database for language phrases generation and complication is necessary.

Key words and phrases: natural language generation; semantic features; language words and notions classification; educational materials generation; natural language generation; meaningful sentences generation.