УДК 81`322

НЕКОТОРЫЕ ОСОБЕННОСТИ РЕКОНСТРУКЦИИ ЗНАЧЕНИЙ СЛОВ ЕСТЕСТВЕННОГО ЯЗЫКА С ИСПОЛЬЗОВАНИЕМ СЕМАНТИЧЕСКОЙ КЛАССИФИКАЦИИ НА ОСНОВЕ САМООБУЧАЮЩИЩСЯ СИСТЕМ

Матанин Г.А., Петров А.С., Личаргин Д.В.
научный руководитель канд. тех. наук Личаргин Д.В.
Сибирский федеральный университет

На сегодняшний день широко распространены и разрабатываются разнообразные системы, направленные на изучение языков и их (реже) непосредственный и (чаще) опосредованный промежуточными языками и данными перевод. Многие языки по природе своей насыщены очень богатой лексикой. Плюс ко всему языки все время развиваются, порождая новые языковые единицы, новые словосочетания и конструкции в результате, как возникновения новых элементов в жизни ввиду непрерывного прогресса, так и в результате замещения изначальных значений слов. В итоге возникающая проблема состоит в том, что люди физически не способны обработать вручную такое количество информации. Современные вычислительные мощности позволяют создавать самообучающиеся системы, которые способны на основе семантической классификации уже известной базы слов определять значения слов ранее неизвестных.

Проблема создания таких самообучающихся систем является актуальной для большинства программ-переводчиков в связи с необходимостью обеспечения правильного, наиболее близкого по значению перевода, который подобные программы в настоящий момент не всегда могут дать, ввиду невозможности автоматического пополнения базы данных актуальными словами. Это также справедливо для семантически ориентированных поисковых систем, преследующих цели корректного отображения результатов на запросы пользователей.

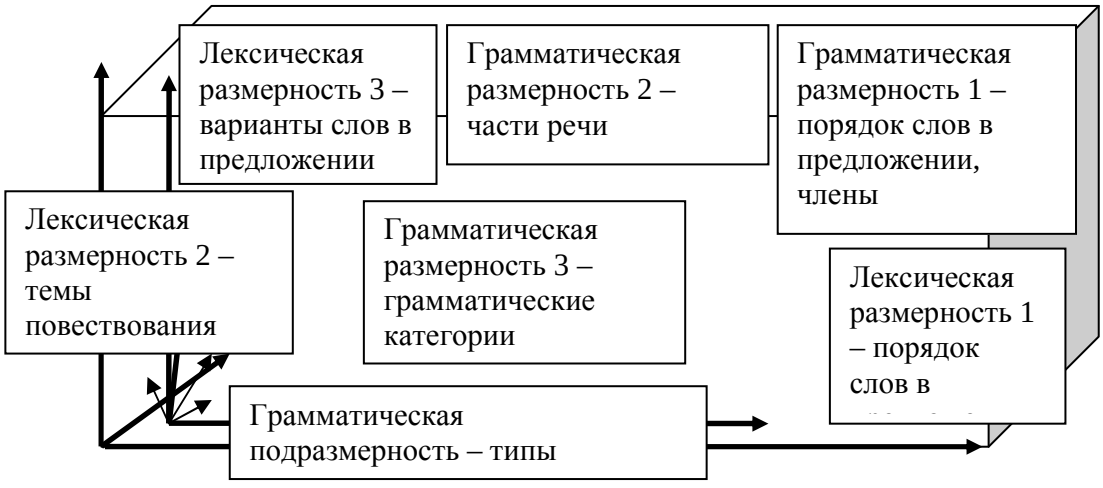

Рис. 1. Координаты многомерного лексико-грамматического подпространства леса данных естественного языка

Для решения этой проблемы в данный момент существуют многочисленные теории, концепции и программные системы, вопрос подвергается анализу в области семантики, дискретной математики, информатики, лингвистики и искусственного интеллекта.

Основная цель работы – построить и описать некоторые принципы самообучающейся системы с возможным привлечением понятийного описания единиц естественного языка и нахождения критериев осмысленности фраз на естественном языке.

Единицы естественного языка включаются в классы единиц естественного языка, пересечение и комбинаторика которых порождает его парадигматическую систему – фрагменты реляционной базы данных как срезов многомерной базы данных (см. рисунок 1).

Парадигмы естественного языка являются подмножествами многомерных пространств, или (что эквивалентно) древесных иерархий естественного языка, представляемых на основе векторов признаков древесной классификации, т.е. векторов координат многомерного пространства для состояний единиц естественного языка.

Следующим образом выглядит подмножество класса «Food» в виде древесной иерархии естественного языка, на основе векторов признаков древесной классификации:

Something
- **Essence As**
  - *Thing*
    - Food (In-Alive)
      - Dishes
        - Cutlet
        - Sausage
        - Salad
      - Building (In-Which-Alive)
        - For Action
          - Over Thing
            - Food (In-Alive)
              - Café (for not much object of exchange: money)
              - Bar (with a liquid effecting mind: alcohol)
              - Restaurant (for many dishes)
  - *Creature*
    - Creature (According to Thing-In-Alive (Food))
      - Eater
      - Gourmand
      - Cook
      - Vegetarian
  - *Idea*
    - Thought
    - Opinion
  - *Information*
- **Aspect**
  - *Of Thing*
    - Food (In-Alive)
      - Spices
        - Pepper
        - Coriander
      - Salt
  - *Of Information*
    - On Essence
      - On Thing
        - On Food
          - Menu
          - Bill
  - *Of Idea*
    - Taste (In-Alive)
      - Salty
      - Spicy
      - Bitter
- **Property**
  - *Of Object*
    - Of Food
      - Nourishing
      - Tasty
      - Appetizing
- **Action**

With Object
With Food
Have (Usage)
Eat
Swallow
Digest
Cook (Making)
Fry
Bake
Roast
Boil
With Idea
Food (In-Alive)
Taste
Enjoy
Presentation
Of Food
Treat … with

Опираясь на вектор, представленный в таблице 1, приведем упрощенный пример, предположим, что нашей системе не известно слово «Barman» и нам нужно определить относится ли оно к теме Food и в каком месте в предложении должно находиться. Пример: «The barman has just served my food in this pub». На основе уже известных системе семантически связанных единиц естественного языка, фактически по шаблону, слово Barmen будет отнесено к теме «food», в колонку «Делатель» наряду со своими дочерними понятиями («waiter», «cook», «vegetarian», etc.) Этот простой пример показывает принцип, по которому может работать самообучающаяся система.

Таблица. 1. Семантические шаблоны по теме «Food».

the ... **этот ...**	to ... **<что делать> ...**	food **пища**	in the ... **в этом...**
good-eater **обжора**	have **есть, пить**	cuisine **кухня**	building **здание**
gourmand **гурман, лакомка**	taste **пробовать**	snack **закуска**	bar **бар**
vegetarian **вегетарианец**	swallow **глотать**	course **блюдо**	snack bar **закусочная**
diabetic **диабетик**	absorb **поглощать**	first course **первое**	cafe **кафе**
drunkard **пьяница**	digest **переваривать**	second course **второе**	cafeteria **кафетерий**
cook **повар**	warm up **разогревать**	third course **третье**	restaurant **ресторан столовая**
waiter **официант**	cool **охладить**	dessert **десерт**	pub **пивная**

Для реальных же условий, для эффективной самообучающейся системы необходимы семантически индексированный словарь не менее чем с 10 000 единицами естественного языка, разложенными в соответствии со структурой вектора многомерного лексико-грамматического подпространства леса данных, а так же базы знаний морфологического, семантического, синтаксического анализа слов языка. Формируют такую систему с помощью индексированных лингвистических текстов на заданном базовом языке. При использовании готовых баз возможно автоматическое обучение системы любому из заданных языков, при условии наличия необходимой информации в лингвистических

текстах по соответствующему целевому языку при наличии эквивалентных преобразований указанных языковых текстов. Происходит образование автоматически и полуавтоматически индексируемых семантических структур, и формирование логических выводов на этих семантических структурах для формирования ответов, которые соответствуют автоматическим запросам, используемым для формирования баз знаний по морфологическому, семантическому, синтаксическому анализу для конкретного языка. Благодаря этому происходит извлечение знаний, то есть самообучение на основе текстовых документов на заданном языке.

Для осуществления процесса самообучения системы правилам морфологического анализа, в автоматически индексируемом тексте выделяют определенное количество словесных форм каждого слова. Основы слова и наборы, имеющихся в тексте, окончаний или предлогов получают стохастические индексы, осуществляют произвольный доступ по указанным индексам к таким лингвистическим текстам. Они выделяют в них все необходимые фрагменты, которые связывают имеющийся набор окончаний слов или предлогов с (соответствующей данному слову) частью речи, включая предлоги и окончания, получившиеся в результате склонения или спряжения. Далее такие фрагменты преобразуются в формат правил, путем их стохастического индексирования, корректность этих правил достигается ввиду их формирования на основе нескольких фрагментов из соответствующих лингвистических текстов.

Автоматическое обучение системы правилам синтаксического анализа осуществляют по подобной схеме, только в индексированных лингвистических текстах осуществляется поиск фрагментов, описывающих не словоформы, а порядок синтаксического разбора предложений. Полученные правила заносят в базу знаний синтаксического анализа, по мере заполнения которой осуществляют ее автоматическое индексирование и представление в виде таблицы индексов.

В результате при появлении ранее не встречавшегося слова во время индексирования текстовых документов, которое не содержится в словаре вероятностно индексированных слов и лингвистических текстах, в словаре находится однокоренное слово с указанным новым словом, далее найденное однокоренное слово преобразуется в новое слово, согласно правилам, содержащимся в базе знаний морфологического анализа. При этом по виду эквивалентного преобразования определяют так же все его словоформы, включая те, что получаются при склонении или спряжении. Если же однокоренных слов в словаре нет, то из текста выбирают определенный набор словоформ нового слова, и определяют часть речи, к которой оно относится, осуществляется полный разбор словоформ при склонении, спряжении, по предлогам и окончаниям, как с помощью вероятностно индексируемого словаря, так и с помощью правил морфологического анализа

Средства автоматической адаптации системы на лексическом и морфологическом уровнях позволяют системе самой установить грамматические характеристики незнакомых языковых объектов, обнаружить и исправить некоторые ошибки в словах. Они позволяют выявить специфические для пользователя отклонения от модели языка, используемой системой, в случае несоответствия слов текста множеству грамматических производных слов словаря.

Выводы. В работе выполнен анализ некоторых проблем по использованию представления значения слов в самообучающихся системах. Рассмотрены некоторые подходы по обеспечению самообучения с учетом разложения знаний по векторам многомерного лексико-грамматического подпространства леса данных.

Рассматриваются возможности создания программной системы для обеспечения автоматического сбора знаний о языке в целях повышения качества работы лингвистического программного обеспечения. Подчеркивается важность продолжения исследований по теории автоматического извлечения знаний из корпусов тестов на естественном языке.