алгоритм базируется на выявлении сходства между элементами (например, между читателями блога) посредством вычисления их расстояния от других элементов в пространстве признаков (признаком в пространстве признаков может, например, быть количество прочитанных статей в наборе блогов). Количество независимых признаков определяет размерность пространства признаков. Если элементы "близки" друг к другу, то их можно объединить в один кластер. Существует множество алгоритмов кластеризации. Самым простым из них является алгоритм k-средних, который разделяет элементы на k кластеров. Первоначально элементы распределяются по этим кластерам в произвольном порядке. Затем для каждого кластера вычисляется центр масс (или просто центр) как функция его членов. После этого проверяется расстояние каждого члена кластера от центра этого кластера.

Список литературы / References

1. *Большакова Е.И.* Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. 269 с.
2. *Васильев В.Г., Кривенко М.П.* Методы автоматизированной обработки текстов. М.: ИПИ РАН, 2008. 301 с.
3. *Гладкий А.В.* Синтаксические структуры естественного языка в автоматизированных системах общения. М.: Наука, 1985. 144 с.

СЛОЖНОСТИ МОДЕЛИРОВАНИЯ ЕСТЕСТВЕННОГО ЯЗЫКА

Юргель В.Ю. Email: Jurgel677@scientifictext.ru

Юргель Владислав Юрьевич - магистрант,
кафедра программного обеспечения информационных технологий,
Белорусский государственный университет информатики и радиоэлектроники,
г. Минск, Республика Беларусь

Аннотация: *появление сети Интернет и стремительно быстрый рост текстовой информации значительно ускорили необходимость и непосредственно развитие научной области, существующей уже несколько десятков лет и известной как автоматическая обработка текстов и компьютерная лингвистика. В области этой идеи было предложено довольно большое количество идей по автоматической обработке текста на естественном языке. Сфера приложений компьютерной лингвистики постоянно расширяется, появляются новые задачи, которые успешно решаются.*

Ключевые слова: *естественные языки, обработка текста, компьютерная лингвистика, моделирование естественного языка.*

COMPLEXITIES OF NATURAL LANGUAGE MODELING

Jurgel V.Yu.

Jurgel Vladislav Yurievich - Undergraduate,
INFORMATION TECHNOLOGY SOFTWARE DEPARTMENT,
BELARUSIAN STATE UNIVERSITY OF INFORMATICS AND RADIOELECTRONICS,
MINSK, REPUBLIC OF BELARUS

Abstract: *the emergence of the Internet and the rapidly rapid growth of textual information significantly accelerated the need and, directly, the development of the scientific field, which has existed for several decades and is known as natural language processing and computational linguistics. In the field of this idea, quite a number of ideas for natural language text processing have been proposed. The scope of applications of computational linguistics is constantly expanding, there are new problems that are successfully solved.*

Keywords: *natural languages, text processing, computational linguistics, natural language modeling.*

УДК 004.021

Сложность моделирования в компьютерной лингвистике возникает в связи с тем, что естественные языки - это большой открытая многоуровневая система знаков, возникшая для

обмена информацией в процессе практической деятельности человека и постоянно изменяющаяся в связи с этой деятельностью.

Текст на естественном языке составлен из отдельных единиц, и возможно несколько способов разбиения текста на единицы, относящиеся к разным уровням.

Существуют следующие уровни:

— синтаксический уровень: уровень предложений (высказываний);

— морфологический уровень: уровень слов (словоформ);

— фонологический уровень: уровень фонем (отдельных звуков).

Фонологический уровень выделяется для устной речи, а для письменных текстов в языках с алфавитным способом записи (в частности, в европейских языках) он соответствует уровню символов (фонемы приблизительно соответствуют буквам алфавита).

Уровни, по сути, есть подсистемы общей системы естественного языка, и в них самих могут быть выделены подсистемы. Так, морфологический уровень включает также подуровень морфем. Морфема - это минимальная значащая часть слова (корень, приставка, суффикс, окончание, постфикс).

Вопрос о количестве уровней и их перечне в лингвистике до сих пор остается открытым. Как отдельный может быть выделен лексический уровень - уровень лексем.

Лексема - это слово, как совокупность всех его конкретных грамматических форм (допустим, лексему лист образуют формы лист, листа, листу, листом). Точнее, лексема - семантический инвариант всех словоформ. В тексте встречаются словоформы (лексемы в определенной форме), а в словаре естественного языка - лексемы, точнее, в словаре записывается каноническая словоформа лексемы, называемая также леммой.

В рамках синтаксического уровня может быть выделен подуровень словосочетаний - синтаксически связанных групп слов (видел лес, синий шар), и надуровень сложного синтаксического целого, которому примерно соответствует абзац текста. Сложное синтаксическое целое, или сверхфразовое единство - это последовательность предложений (высказываний), объединенных смыслом и лексико-грамматическими средствами. К таким средствам относятся в первую очередь лексические повторы и анафорические ссылки - ссылки на предшествующие слова текста, реализуемые при помощи местоимений и местоименных слов (они, этот, там же и т.д.).

Иерархия уровней проявляется в том, что единицы более высокого уровня разложимы на единицы более низкого (например, словоформы на морфы); более высокий уровень в большой степени обуславливает организацию нижележащего уровня - так, синтаксическая структура предложения в значительной мере определяет, какие должны быть выбраны словоформы.

Можно также говорить ещё об одном уровне - уровне дискурса, под которым понимается связный текст в его коммуникативной направленности. Под дискурсом понимается последовательность взаимосвязанных друг с другом предложений текста, обладающая определенной смысловой целостностью, за счет чего он выполняет определенную прагматическую задачу. Во многих типах связных текстов проявляется традиционная схематическая (дискурсивная) структура, организующая их общее содержание, например, определенную структуру имеют описания сложных технических систем, патентные формулы, научные статьи, деловые письма и др.

Особым является вопрос об уровне семантики. В принципе, смысл есть всюду, где есть знаковые единицы языка (морфемы, слова, предложения). Подтверждением самостоятельности уровня семантики считается то, что человек обычно запоминает смысл высказывания, а не его конкретную языковую форму. До сих пор не ясна организация этого уровня, предполагается, что существует универсальный набор элементарных семантических единиц (называемых семами), примерно 2 тысячи, при помощи которых можно выразить смысл любого высказывания.

Кроме многоуровневости системы естественного языка сложность его моделирования связана с постоянно происходящими в нем изменениями (что вполне ощутимо по прошествии одного-двух десятилетий). Изменения касаются не только словарного запаса языка (новые слова и новые смыслы старых), но также синтаксиса, морфологии и фонетики. Как следствие, принципиально невозможно единожды разработать формальную модель конкретного естественного языка и построить соответствующий лингвистический процессор. Одной из самых больших сложностей при обработке текстов на естественном языке является неоднозначность (многозначность) его единиц, проявляющаяся на всех его уровнях, что выражается в явлениях полисемии, омонимии, синонимии.

Полисемия - наличие у одной единицы языка нескольких связанных между собой значений, в частности, полисемия слов, например: земля - суша, почва, конкретная планета.

Синонимия - полное или частичное совпадение значений разных единиц, например: синонимия слов: негодяй и подлец, синонимия приставок (морфов) пре- и пере- (прекрасный, пересохший).

Омонимия - совпадение по форме двух разных по смыслу единиц (в отличие от полисемии нет смысловой связи между совпавшими по форме единицами). Различают следующие виды омонимии:

— лексическая омонимия: означает одинаково звучащие и пишущиеся слова, не имеющие общих элементов смысла, например;

— морфологическая омонимия: совпадение форм одного и того же слова (лексемы), например словоформа карандаш соответствует именительному и винительному падежам.

— лексико-морфологическая омонимия: возникает при совпадении словоформ двух разных лексем;

— синтаксическая омонимия: означает неоднозначность синтаксической структуры, что приводит к нескольким интерпретациям.

Заключение

Компьютерная лингвистика демонстрирует вполне осязаемые результаты в различных приложениях по автоматической обработке и анализу текстов на естественном языке. В большинстве приложений используются простые и редуцированные модели естественного языка, которые однако дают приемлемые или даже хорошие результаты; нередко качество результатов достигает экспертного уровня - обычно там, где мнения экспертов могут расходиться.

Дальнейший прогресс в области компьютерной лингвистики связан как с более точным учетом лингвистических особенностей текстов на различных этапах его обработки и применением более детальных лингвистических моделей, так и с развитием методов машинного обучения и поиском более эффективных методов и их комбинаций для каждой прикладной задачи.

Список литературы / References

1. *Большакова Е.И.* Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. 269 с.
2. *Васильев В.Г., Кривенко М.П.* Методы автоматизированной обработки текстов. М.: ИПИ РАН, 2008. 301 с.
3. *Гладкий А.В.* Синтаксические структуры естественного языка в автоматизированных системах общения. М.: Наука, 1985. 144 с.