5. Гаврилова Т.А., Червинская К.Р. Извлечение и структурирование знаний для экспертных систем. – М.: Радио и связь, 1992.
6. Головко В.А. Нейроинтеллект: Теория и применения. Книга 1. Организация и обучение нейронных сетей с прямыми и обратными связями - Брест:БПИ, 1999, - 260с.
7. Ермаков А.Е., Плешко В.В. Синтаксический разбор в системах статистического анализа текста. // Информационные технологии. – 2002. – N7.
8. Солтон Дж. Динамические библиотечно-информационные системы. – М.: Мир, 1979 – 230с.

Статья поступила в редакцию 06.03.2007

УДК 801.73:681.3

Антонов С.Г.

ИЕРАРХИЧЕСКАЯ МОДЕЛЬ ТЕКСТА ДЛЯ ЗАДАЧИ КОРРЕКЦИИ ОШИБОК

Введение

Традиционно основными показателями эффективности систем автоматической переработки текста являются показатели полноты и точности. Поскольку такая переработка в общем случае является многостадийной, т.е. предполагает прохождение всех уровней глубины языка, то, учитывая его природу, итоговые значения указанных показателей в реальных промышленных системах, как правило, невелики. Анализ показал, что одной из наиболее существенных причин этого является наличие различного рода ошибок в самом входном тексте, т.е. нарушение его лингвистической адекватности.

Процедура устранения нарушений лингвистической адекватности текста в общепринятом понимании включает: поиск ошибок, построение вариантов коррекции и устранение неоднозначности. Наиболее сложным этапом является устранение неоднозначности вариантов коррекции, поскольку их построение, как и поиск ошибок, являются в большей степени техническими задачами.

Устранение неоднозначности вариантов коррекции, очевидно должно базироваться на формализованных лингвистических знаниях, выраженных посредством модели текста и ее содержания. В основу такой модели, как оказалось, целесообразно положить естественную иерархичность графического и лингвистического представления текстов любых алфавитных естественных языков: буква (графический уровень) – морфема – словоформа - сегмент предложения – предложение [1].

Общая схема построения иерархической модели

Рассмотрим общую схему построения иерархической модели текста до синтаксического уровня включительно.

При построении модели сначала выделяются элементы текста каждого уровня иерархии и принципы перехода с уровня на уровень. В результате определяется тип модели, например, n-граммная, морфологическая, синтаксическая и т.п., и далее строится непосредственно сама модель [2].

Основной составной частью модели является система простых элементов модели или элементов первого уровня. Обозначим их множеством $Z^{(1)} = \left\{ z_i^{(1)} \right\}_{i=\overline{1,r_1}}$. Это элементы, не имеющие деления на более мелкие в пределах данного уровня. Выбор системы простых элементов определяет основу всей модели, являясь при этом связующим звеном между моделью и реальными объектами окружающей действительности.

Другой составной частью являются правила построения элементов следующего уровня $Z^{(2)} = \left\{ z_i^{(2)} \right\}_{i=\overline{1,r_2}}$ на языке простых элементов, заданные или списочно, или посредством некоторого другого описания. Дальнейшее наращивание модели происходит по такой же схеме.

Поскольку система простых элементов текста может быть выделена на любом уровне его иерархического представления, уровнем модели будем считать уровень иерархии текста, на котором базируется система простых элементов. Например, модель на уровне букв означает, что простыми элементами модели являются буквы, из которых строятся другие элементы текста, а в модели на уровне слов, соответственно, текст строится из словоформ.

Наибольшую сложность представляет формализация закономерностей построения элементов текста каждого уровня. Если ввести в состав сложных элементов любого уровня выше первого так называемый пустой элемент $\varnothing$, т.е. элемент, не соответствующий никакому объекту текста, то тогда можно описать правило построения элементов h-го уровня из элементов (h-1)-го в виде некоторого отображения

$$\pi_{h-1} : \bigcup_j \left(Z^{(h-1)} \right)^j \to Z^{(h)},$$

где $\left(Z^{(h-1)} \right)^j$ есть j-я декартова степень (т.е. все последовательности длины j) множества элементов (h-1)-го уровня. Вообще говоря, j есть любое натуральное число. Однако текстам естественных языков всегда присущи ограничения на существующие длины последовательностей, т.е. $\exists J_{h-1} : \forall \bar{z} \in \left(Z^{(h-1)} \right)^j, j > J_{h-1}, \pi_{h-1}\left(\bar{z}\right) = \varnothing$. Тогда нам достаточно рассматривать ограничение π_{h-1} на $\bigcup_{j=1}^{J_{h-1}} \left(Z^{(h-1)} \right)^j$. Указанное ограничение будем называть правилами построения элементов следующего уровня

$$\pi_{h-1}^* : \bigcup_{j=1}^{J_{h-1}} \left(Z^{(h-1)} \right)^j \to Z^{(h)}.$$

Если модель имеет p уровней, то тогда должны быть описаны правила $\pi_h^*, h = \overline{1, p-1}$. Набор $\pi^* = \left(\pi_h^* \right)_{h=\overline{1,p-1}}$ можно назвать набором правил модели, а суперпозицию $\pi' = \pi_{p-1}\left(\pi_{p-2}\left(\ldots \left(\pi_1 \right) \ldots \right) \right)$ правилами построения $Z^{(p)}$ из $Z^{(1)}$.

Очевидно, что для построения модели необходимо описать элементы всех уровней и правила их построения, основанные на неких признаках этих элементов.

Множество элементов модели h-го уровня должно содержать все объекты формализованного текста, которые могут встретиться на соответствующем уровне. Эти объекты можно назвать «максимальными объектами», соответствующими данному уровню модели. Иерархическое представление уровней модели предопределяет, что максимальные объекты надо

***Антонов Сергей Георгиевич**, к.т.н., старший научный сотрудник, научный консультант ООО «Крос 2000», Россия, г. Москва.*

составить из более мелких элементов, которые, в свою очередь, являются максимальными объектами для предыдущего (h-1)-го уровня.

Для выражения максимальных объектов текста для каждого уровня модели, кроме первого, предлагается ввести разрешенные правилами построения текста естественного языка максимальные формы, т.е. шаблоны элементов высшего уровня, состоящие из элементов низшего уровня, или, иными словами, допустимые в языке упорядоченные последовательности элементов данного уровня максимальной длины [1,2,3]. Переход с графического уровня на следующий уровень иерархической модели, т.е. словообразование, в терминах морфологических элементов должно описываться одной максимальной формой, а образование сегментов предложений из словоформ и предложений из сегментов – набором форм, отражающих различные типы сегментов и предложений.

Правила построения элементов различных уровней модели текста должны описывать два взаимосвязанных процесса взаимодействия данных. Один процесс можно назвать «горизонтальным» или «внутриуровневым», отвечающим за согласование элементов одного и того же уровня при их объединении. Реализация этого процесса происходит при использовании признаков, которые будем называть «согласующими» признаками для данного уровня модели. Второй процесс можно обозначить как взаимодействие «снизу – вверх», т.е. формирование элементов более высокого уровня из элементов предыдущего уровня с вычислением характеристик вновь построенного уровня. Признаки элементов текста предыдущего уровня, обеспечивающие вычисление признаков следующего уровня, будем называть «образующими».

Очевидно, что согласующие признаки элементов каждого уровня взаимодействуют только внутри этого уровня модели. Некоторые образующие признаки могут использоваться не только для образования признаков следующего уровня, но способны отражать согласование элементов низшего уровня. Сама максимальная форма также накладывает ряд ограничений на согласование элементов модели текста.

Проблема разработки систем согласующих и образующих признаков является наиболее важной для процесса конструирования модели текста естественного языка, поскольку выделение других компонентов модели – собственно элементов текста, а именно, морфем, словоформ, фрагментов предложений, происходит в основном (но не всегда) в соответствии со сложившимися в традиционной грамматике правилами.

В качестве практической реализации изложенной выше схемы формального представления лингвистической информации с помощью иерархической модели была проведена попытка построения модели текста французского языка. Модель действует в рамках пяти первых указанных выше уровней.

Минимально значимыми с лингвистической точки зрения единицами текста в предлагаемой модели являются морфологические единицы (в некотором роде формальные аналоги морфем). Поэтому первая из задач связана с переходом от букв (графического уровня) к уровню морфологических единиц (МЕ). Это может быть сделано лишь фиксацией таких МЕ в специальных таблицах, т.к. не существует алгоритма порождения МЕ из букв [4]. Однако сами МЕ обладают определенной многозначностью, вследствие этого они не могут служить элементами второго уровня иерархической модели. Поэтому уже на этом этапе возникает необходимость формализации, т.е. замены реальных МЕ наборами некоторых категориальных характеристик. По сложившемуся опыту построения иерархических моделей часть характеристик относится к форме представления данных, а не к содержанию. В рассматриваемом случае отдельно составляются словари префиксов, корней, суффиксов и флексий, а также полные списки специфичных для языка предлогов, союзов, служебных глаголов, имён собственных и т.д. Всё это, по существу, описывает таблично часть следующего уровня иерархии – уровня словоформ. Осуществленный в [5] анализ показал, что выделение морфологических единиц было проведено правильно не только с лингвистической, но и с информационной (по Шеннону) точек зрения.

Остальные характеристики, требуемые для описания текста создаваемой моделью, должны обеспечивать, как было сказано выше, два процесса анализа данных. Для перехода с уровня МЕ к уровню словоформ в категориальных характеристиках должны быть формально представлены морфологические и словообразовательные закономерности для согласующих признаков и синтаксические закономерности для образующих признаков.

Анализ французской морфологии и словообразования показал, что без использования семантических характеристик для иерархической модели можно использовать только фонологические ограничения в виде согласующих признаков. К таким ограничениям относятся, например, употребление префиксов и корней, начинающихся с гласной или немого h после префикса RE-, или начинающихся с букв p, m, b после префиксов, заканчивающихся на m (IM-, UM- …).

К согласующим признакам также можно отнести формально-позиционные характеристики МЕ, такие как: всегда начальный префикс; прикорневые префикс и суффикс; оканчивающий словоформу суффикс; промежуточный суффикс (требующий после себя обязательно ещё один суффикс). В этом случае легко и достаточно прозрачно решался вопрос универсальности описания и его алгоритмической реализуемости. Ради этого пришлось иначе интерпретировать некоторые МЕ. Например, флексия мужского рода –EUX была причислена к оконечному суффиксу.

Частеречную информацию для использования в качестве согласующих признаков не удалось каким-то образом обобщить, поэтому она используется только при проверке допустимости продолжения сформированной части словоформы какой-либо флексией, т.е. характеристика части речи для флексии должна совпадать с уже вычисленной характеристикой части речи для словоформы.

Лишь частично использовались при таком переходе характеристики «род/число». Совместно с позиционными характеристиками суффиксов для некоторых суффиксов употребляется прямой запрет на использование после них флексий рода и/или числа.

В результате оказалось, что в описании перехода с графического уровня на следующий уровень иерархической модели морфология и словообразование в классическом понимании такого перехода лингвистикой, нашли прямое отражение только в виде максимальной формы, т.е. в форме представления данных, и в меньшей мере - в самих данных.

Следующим уровнем описываемой модели является уровень словоформ, точность вычисления категориальных характеристик единиц которого имеет наибольшее значение. Поиск и описание различных лингвистических характеристик, необходимых для формализации представления сегментов предложений и предложений из словоформ являются важными не только для моделирования предложений, но и для построения образующих признаков и алгоритма вычислений при переходе от МЕ к словоформам. Таким образом, третий уровень иерархической модели, описывающий словоформы, оказывается главным ядром всей модели, в целом согласующимся с распространенным мнением, что центральным элементом естественного языка является слово.

Основная часть категориальных характеристик словоформ французского языка была использована в качестве образующих признаков, которые были приписаны к конкретным МЕ при описании данных морфологического уровня. Это следующие характеристики: часть речи, принимающая 53 различ-

ных значения; род/число – 3 значения (мужской и женский род, множественное число); 3 лица; 5 типов артиклей; 6 видовременных форм глагола. На первый взгляд может показаться странным очень большое количество частей речи и небольшое количество остальных характеристик, однако это вполне объяснимо характером создаваемой модели. Характеристики, уточняющие части речи на уровнях «словоформа – сегмент – предложение» являются согласующими признаками. В максимальных формах для переходов между этими уровнями используется характеристика части речи, а её уточнения описываются как параметры, требующие или не требующие согласования внутри объекта, описываемого максимальной формой.

Из вышеизложенного становится понятным, почему среди 53 частей речи - 6 различных прилагательных, 18 местоимений, 9 глаголов, 6 причастий.

Сегменты предложений (как самостоятельные, так и в виде рекурсивных вставок в другие сегменты) описываются 27 максимальными формами. Приведем пример описания максимальной формы одного из сегментов, а именно числительного (С):

C = (IN)(IN)(PI)(PI'')(A)(IN) C (P'')(J)(J'')(IN)
0 0 0 0 2 0 2 2 2 2 0

Ряд чисел под последовательностью частей речи означает необходимость их согласования по некоторым характеристикам, в данном случае – по роду/числу в позиционированном представлении лексико-грамматических классов слов [3]. В скобках указаны необязательные компоненты формы, т.е. в конкретных реализациях они могут отсутствовать.

Среди максимальных форм выделены одна номинативная (существительного), 10 местоименных, 1 - числительного, по 3 причастных и деепричастных, 9 глагольных форм.

Анализ целесообразности использования предложенного распределения лингвистической информации между понятиями модели «часть речи» и «уточняющие характеристики» приводит к выводу, что существуют закономерности в текстах любого естественного языка, которые можно описать простой линейной зависимостью. Простота этой зависимости проявляется только после более широкого толкования рассматриваемых лингвистических характеристик по отношению к классическому их определению в лингвистике.

Очевидно, что полноценное представление всех свойств текста в рамках линейной модели невозможно. Однако даже легко формализуемые закономерности в рамках этой линейной модели требуют достаточно существенных усилий по приведению общепризнанных лингвистических характеристик к несколько искусственной группировке. Это, по всей видимости, относится к любым попыткам построить реляционную модель текста, пусть это будет статистическая модель типа марковской либо иерархическая, кажущаяся многоуровневой, однако таковой по сути не являющаяся [6,7]. Моделью, адекватной человеческому восприятию, может быть лишь синтетическая модель, состоящая из набора различных моделей реляционного и сущностного характера, сочетание которых определяется только в зависимости от конкретной проблемы, решаемой при анализе того или иного фрагмента текста. С этой точки зрения совокупные категориальные характеристики, как оконечные признаки, описывающие ту или иную единицу текста, можно считать достаточно универсальным средством представления такого описания. Именно поэтому их используют в различных статистических моделях, не учитывающих структуру текста.

Заключение

В целом иерархическую модель можно считать некоторым «скелетом», на который, в зависимости от решаемой задачи, тем или иным образом наращиваются «мягкие ткани», представляющие собой различное отражение классических лингвистических понятий. Структурирование линейного представления лингвистической информации в виде совокупных категориальных характеристик, введение новых классов, представляющих уже известные объекты, помогают учитывать закономерности, присущие тексту с точки зрения цели конкретной задачи.

СПИСОК ЦИТИРОВАННЫХ ИСТОЧНИКОВ

1. Antonov S., Nikishin A. Elements of Hierarchic Model of a Text. 12th World Congress of Applied Linguistics. AILA'99 Tokyo, August 1-6, 1999, Program &Abstracts p.389.
2. Антонов С.Г., Никишин А.Г. Метод построения иерархической модели текста. // Материалы международной конференции «Когнитивное моделирование в лингвистике», Переславль-Залесский, 2000. / Обработка текста и когнитивные технологии, №5 – М., 2001, с. 35-42.
3. Антонов С.Г. Позиционная модель именного сегмента в тексте немецкого языка. Вестник Минского государственного лингвистического университета. Серия 1. Филология. №14 – Мн.: МГЛУ. – 2004, с.158-162.
4. Яковишин В.С. Формальный язык. Теория. Грамматика. Применение. – Минск: НАНРБ, 2000.
5. Антонов С.Г., Никишин А.Г., Пересыпкин В.А., Степанов А.В. Об информационных свойствах морфологической модели текста. Актуальные проблемы компьютерной лингвистики. Сб. научных статей. – Мн., 2005. – С.16-24.
6. Зубов А.В. Статистика и моделирование как основа систем искусственного интеллекта // Международный семинар "Язык и технология" – Санкт-Петербург, 1996, с.16.
7. Антонов С.Г., Богушевич Д.Г., Зубов А.В., Нехай О.А., Никишин А.Г. К проблеме построения формальных моделей естественных языков. Вестник Минского государственного лингвистического университета. Серия 1. Филология. №4 – Мн.: МГЛУ. – 1998, с.165-169.

Статья поступила в редакцию 21.05.2007

УДК 004.421.2

Афонин В.Г.

О НЕКОТОРЫХ ПРОЦЕДУРАХ ВЫЧИСЛИТЕЛЬНОЙ ПРАКТИКИ

Введение. Важнейшим итогом решения любой вычислительной задачи является степень доверия к полученным результатам, оценка их погрешности и устойчивость вычислительного процесса. Эта устойчивость (корректность) является также важнейшей характеристикой всей математической модели реальной задачи. В п. 1.1 данной работы предлагается весьма простой и высокоэффективный приём установления степени корректности широкого класса вычислительных алгоритмов.

Как правило, в математических моделях присутствуют параметры, которые получаются в результате измерений. Эти измерения содержат погрешности, зачастую немалые. Кроме

Афонин Владимир Гаврилович, *к.ф.-м.н., доцент кафедры информатики и прикладной математики БрГТУ.*
Беларусь, Брестский государственный технический университет, 224017, г. Брест, ул. Московская 267.