

УДК 378.147:681.3

АНАЛИЗ СВОБОДНО КОНСТРУИРУЕМОГО ОТВЕТА ПО КЛЮЧЕВЫМ СЛОВАМ

Новикова Н.А.

Рассматривается способ анализа ответа обучаемого на контрольное задание путем выявления в ответе ключевых слов, их абсолютного и относительного расположения и сопоставления с эталоном ответа.

В процессе общения на естественном языке часто достаточно услышать только несколько слов из фразы для приблизительного понимания смысла сказанного. Известны различные системы анализа текста по ключевым словам [1,2]. Однако их задачей является, как правило, отнесение анализируемого текста к той или иной предметной области. Дополнив поиск ключевых слов, некоторыми операциями, можно создать достаточно эффективную методику автоматизированного распознавания ответов обучаемого, использующую упрощенный анализ семантики ответа на основе поиска слов, заданных преподавателем в эталоне ответа (ключевых слов).

Анализ текста предусматривает выполнение следующих операций:

- определение вхождения ключевого слова в анализируемый текст (одиночного или множественного);
- определение абсолютной позиции ключевого слова относительно начала, конца текста или другого ключевого слова (слева или справа);
- выделение и занесение в список ключевых слов некоторого выделенного из текста слова в зависимости от его абсолютной позиции (обычно это должны быть некоторые имена, вводимые обучаемым);
- определение недопустимых в тексте слов, если необходимо, то с указанием границ недопустимого вхождения;
- определение относительной позиции вхождения ключевых слов (порядка их следования в тексте);
- определение синонимов ключевых слов.

Формирование эталона ответа. Для формирования эталона ответа предлагаются три варианта решения.

А) Формирование эталона в виде строки, записанной на специальном метаязыке эталона.

Б) Формирование эталона в диалоговом режиме, когда метаязык в явном виде не фигурирует, но автоматически строится внутреннее представление эталона.

В) Запись предполагаемого правильного ответа на естественном или искусственном языке с последующей обработкой в диалоговом режиме с целью выделения ключевых слов и формирования эталона ответа. В этом случае внутреннее представление эталона также строится автоматически [3].

Поскольку в явной, либо в неявной форме метаязык присутствует во всех трех случаях, рассмотрим его описание.

Метаязык для записи эталона. Эталон на метаязыке состоит из одного или нескольких правил. Для разделения правил используется символ.

Символы “<” и “>” используются в качестве границ ключевого слова.

Если ключевое слово имеет синонимы (с точки зрения предметной области и сущности поставленного вопроса), то в рамках границ ключевого слова они разделяются символом “%”, например: <integer%byte>.

Ключевое слово может иметь атрибуты, которые записываются непосредственно за его правой границей.

Атрибут абсолютной позиции вхождения представляется одним символом и может принимать значения:

- b – слово должно занимать первую позицию в ответе;
- e – слово должно занимать последнюю позицию в ответе;
- l – слово должно непосредственно предшествовать другому ключевому слову, например: <магнитный>l<диск>;
- r – слово должно располагаться непосредственно за другим ключевым словом, например: <язык><программирования>r;

Атрибут количества вхождений задается одной цифрой, например: <function>1;

Вхождение ключевого слова один или более раз определяется символом “+”, а недопустимое вхождение – символом “-“, например: <procedure>-. После указанных атрибутов можно использовать атрибут границы дозволенного или запрещенного вхождения в виде целого числа, являющегося порядковым номером ключевого слова, до которого действует данное правило. Например, фрагмент правила <var>-15 запрещает использовать ключевое слово var в границах от появления этого слова до ключевого слова с номером 15 в текущем правиле.

Для включения в список ключевых слов одного из слов анализируемого ответа, записывается конструкция вида <n>, где n – номер определяемого ключевого слова. В дальнейшем это слово в эталоне будет иметь данный номер.

Для определения «веса ошибки» в ответе в правило может быть введен атрибут “vn”, где n – цифра, определяющая «вес ошибки».

Особое правило вида ^&^ и ^&<слово1%слово2%.....>^ ограничивает вхождение в ответ слов. В первом случае это могут быть только слова, разрешенные в ответе предыдущими правилами, во втором – к ним добавляются еще и те слова, которые непосредственно приведены в данном правиле.

Рассмотрим пример формирования эталона для анализа ответа обучаемого на задание: «Дайте описание некоторой функции на языке Паскаль»:

<program>-^<procedure>-^

<function>1b<1>l<:><integer%byte%real%string><;><begin><var>-
9<end><;>e.

Внутренняя форма эталона. Внутренняя форма нужна для непосредственного анализа ответа. Она образуется в результате синтаксического анализа эталона ответа на метаязыке либо в диалоговом режиме и представляет собой следующую таблицу:

Ключевое слово	Номер слова в правиле	Код слова	Слово – альтернатива	Вхождение слова	Граница вхождения	Позиция слова	Ссылка на ответ	Вес слова
----------------	-----------------------	-----------	----------------------	-----------------	-------------------	---------------	-----------------	-----------

Все колонки таблицы, кроме последней, заполняются в процессе обработки эталона ответа. Ссылка на ответ дает возможность связать ключевое слово со словом в ответе, что можно использовать при анализе ответа.

Выделение слов в ответе. Выделение слов в ответе производится по принципу определения разделителей. Для этого предлагается сформировать две строки. Первая содержит односимвольные разделители, вторая – двухсимвольные. Необходимость в двухсимвольных разделителях появляется при анализе ответа, записанного на формализованном языке, например языке программирования. Каждый разделитель, кроме пробела, в свою очередь интерпретируется как слово и попадает во внутреннюю форму ответа. Ниже приведен пример строки односимвольных разделителей и строки двухсимвольных разделителей для разбора ответа на некотором языке программирования:

“.,;,” “<!=>=”.

Сопоставление слов. Идентификация ключевых слов в ответе может выполняться различными способами. В некоторых случаях можно выбрать способ полного совпадения, например, при работе с формализованными языками. При работе с естественными языками полное совпадение не применимо, и можно предложить выделение основ слов, что само по себе является трудоемкой задачей и требует наличия в системе больших словарей, а также реализации правил словообразования. Учитывая приведенные доводы, предлагается вероятностное сравнение слов, реализующее тот известный факт, что смысл слова в основном определяется его первыми символами (основой), суффикс и окончание вносят, конечно, изменения в семантику, но обычно эти изменения не носят кардинального характера. Даже если речь идет о специальных терминах и условных обозначениях, их авторы обычно стараются сохранить начальную часть слова, если речь идет о родственных понятиях.

Предлагаемая схема сравнения предписывает назначение каждому символу слова своего «веса», который уменьшается по мере удаления от начала слова. Ниже приведен один из возможных алгоритмов сравнения: выделяется слово V; определяется его длина n; определяется «вес» слова S

$= 1+2+...+n$; определяется «вес» каждого символа $V_i = n-i+1$; производится сравнение с эталоном O_j ; при совпадении символов их веса суммируются, то есть, если $O_j = V_i$, то $S_j = S_j + V_i$; вычисляется вероятность совпадения выделенного слова с эталонным $P_j = S_j / S$; выбирается эталон O_m с наибольшей вероятностью совпадения символов. Если эта вероятность больше пороговой, то принимается решение о совпадении слов.

Внутреннее представление ответа и его анализ. Слова, выделенные из текста ответа, заносятся в таблицу, приведенную ниже.

Слово из ответа	Сообщение о результате анализа	Вес ошибки
-----------------	--------------------------------	------------

В процессе анализа ответа происходит его сравнение с эталоном. Ошибки, имеющие локальный характер, т.е. типа «недопустимое слово в ответе», «недопустимое слово в данном фрагменте», «нет предыдущего слова», «нет последующего слова» записываются в таблицу во вторую колонку в строку с соответствующим ошибочным словом из ответа. Приведенная таблица может использоваться в качестве сообщения обучаемому в режиме автоматического обучения. Ошибки, имеющие глобальный характер, типа «нет слова xxx в первой позиции», «избыточное число вхождений слова xxx в ответ», выдаются отдельным сообщением.

В режиме контроля оценка может определяться как на основе подсчета отношения количества выполненных правил к общему числу правил в эталоне, так и дополнительно путем определения рейтинга каждого правила.

В настоящее время создан действующий макет программы анализа свободно сконструированного ответа. Выполняется работа по созданию инструментальной системы, содержащей изложенные решения как вариант обработки ответов обучаемого.

Приведенная методика анализа ответа не претендует на универсальность и, безусловно, не может обеспечить стопроцентное качество анализа в любой предметной области. Однако во многих приложениях, особенно связанных с точными науками, она может быть достаточно эффективной. Достоинством ее также является простота подготовки эталона ответа, что является привлекательным с точки зрения преподавателя.

Библиография

1 Brown J.S. Situated cognition and the culture of learning / J.S.Brown, A.Collins, P.Duid // Edic. Resercher. — 1989. — № 18(1). — P. 32 – 42.

2 Anderson J.R. Cognitive psychology and intelligent tutorial / J.R.Anderson // Proceeding of the cognitive science society conference, Balder, Colorado, 1984. — Balder, 1984. — P. 37 – 43.

3 Хантер Р. Проектирование и конструирование компиляторов / Р.Хантер. — М., 1984. — С. 87 – 91.

Поступила в редакцию 25.09.2000 г.