Анализ формальных понятий и сжатие текстовой информации в задаче автоматизированного контроля знаний.

Емельянов Г.М., Михайлов Д.В.

Новгородский государственный университет имени Ярослава Мудрого

Настоящая работа посвящена вопросам организации структуры, пополнения и использования знаний о синонимии в предметно-ограниченном естественном языке для численной оценки смысловой близости текстов при интерпретации тестов открытой формы в системах автоматизированного обучения и контроля знаний.

Тестовое задание открытой формы в системе контроля знаний предполагает ответ испытуемого в виде одного или нескольких предложений естественного языка. Как правило, разработчик теста формулирует свой вариант правильного ответа на основе собственных знаний по заданной предметной области. Традиционно интерпретация ответа испытуемого здесь заключается в простом поиске среди “правильных” вариантов (система “ТестЭкзаменатор”), либо в ручной обработке ответа. Как следует из работ Аванесова В.С., Красильниковой В.А., Майорова А.Н., сама постановка теста открытой формы зачастую сводится к простым заданиям на дополнение с ограничениями на ответы. Для правильного оценивания результатов выполнения открытого теста необходимо наличие подсистемы обработки естественного языка, которая, во-первых, способна обрабатывать высказывания с отклонениями от грамматической нормы, а во-вторых, характеризуется единообразием механизмов оперирования предметными и языковыми знаниями.

В общих чертах интерпретация результата выполнения задания открытой формы есть анализ степени близости ответа испытуемого заданному эталону, что предполагает доказательство идентичности ролей сходных понятий относительно сходных ситуаций, описываемых сравниваемыми текстами ответа испытуемого и “правильного” ответа, сформулированного разработчиком теста. Наиболее близка данной идее обработка текстов на основе коммуникативной грамматики. Хорошим примером является поисковая система “Exactus”. Тем не менее, в отличие от традиционного для поисковых систем взаимодействия “запрос – ответ”, интерпретация результатов теста открытой формы означает не столько отображение ответа на заданную предметную область, сколько оценку близости нужному эталону. По оценке Г. С. Осипова, свойства связей между минимальными семантико-синтаксическими языковыми единицами в самой коммуникативной грамматике требуют более детального изучения. В связи с этим автоматизация накопления знаний о взаимодействии семантики, синтаксиса и морфологии естественного языка при установлении смысловой эквивалентности текстов является чрезвычайно актуальной.

Учитывая вышесказанное, *цель* работы была сформулирована как *разработка теоретико-методологических основ организации обработки естественного языка для задачи автоматизированного обучения и контроля знаний на основе тестовых заданий открытой формы*.

Для достижения поставленной цели в качестве модели семантики конструкций естественного языка в работе предложено использовать модель ситуации языкового употребления. При этом наиболее естественно концептуальная модель ситуации языкового употребления реализуется на основе идей и методов анализа формальных понятий – в виде формального контекста. При этом в решетке формальных понятий выделяются классы семантических отношений по сходству:

- основы главного слова, что особенно актуально для исследования сочетаемости в рамках лексических функций-параметров, посредством которых описываются расщепленные предикатные значения;
- флексии зависимого слова, что необходимо для выделения и обобщения синтаксических отношений;
- лексической и флективной сочетаемости, что позволяет выявить зависимости, аналогичные смысловой связи между опорным словом и генитивной именной группой в составе генитивной конструкции русского языка.

Численной оценкой схожести ситуаций языкового употребления определяется смысловая близость высказываний. В дополнение к моделям ситуаций языкового употребления здесь используется теоретико-решёточная модель тезауруса предметной области. При этом отдельные ситуации языкового употребления соответствуют объектам формального контекста. Признаковое множество формального контекста тезауруса включает признаки из формальных контекстов отдельных ситуаций, признаки-указания на их объекты, связи “основа-флексия” для синтаксически зависимого слова, а также сочетания основ зависимого и главного слова.

Само отношение схожести между корректным описанием некоторого факта (ответ преподавателя) и анализируемым (ответ студента) будет иметь место тогда, когда для каждого объекта в составе одного формального контекста найдётся прообраз в другом формальном контексте по сходству флективной и лексический сочетаемости. Причем указанные виды сочетаемости рассматриваются как относительно сравниваемых формальных контекстов (*условие 1*), так и с привлечением формального контекста тезауруса (*условия 2–4*). Предложенное формальное определение схожести ситуаций языкового употребления отражает случаи синонимии среди слов, синтаксически главных по отношению к сравниваемым (*условие 2 и 3*), в том числе с учетом родовидовых отношений (*условие 4*). Численная оценка схожести ситуаций определяется количеством признаков, которые разделяются объектами сравниваемых ситуаций относительно формального контекста тезауруса. Чем больше слов могут быть синтаксически главными по отношению к каждому из слов сравниваемой пары, тем выше значение схожести. При наличии в структуре формального контекста анализируемой ситуации языкового употребления хотя бы одного объекта, для которого нет выполнимых условий определения схожести, значение последней считается равным нулю.

Заметим, что число фраз, задающих ситуацию языкового употребления, как и число ситуаций, представленных в тезаурусе, изначально не оговаривается. Сказанное влияет на точность вычисления оценок смысловой близости.

Решение указанной проблемы подразумевает:

- отбор фраз для формирования тезаурусных единиц. При этом каждая фраза должна максимально точно описывать свою ситуацию действительности (выражать смысл “на одном дыхании”);
- разделение знаний о сходных языковых формах описания различных ситуаций действительности (с одной стороны) и внешне различающихся формах наиболее “компактного” описания каждой из ситуаций в тезаурусе (с другой стороны).

Для решения задачи отбора фраз, представляющих тезаурусные единицы, вводится понятие смыслового эталона ситуации языкового употребления и рассматриваются два приближенных метода его построения с представлением в виде формального контекста.

Первый и наиболее естественный метод основан на подходе к выделению и классификации синтагматических зависимостей. По сути, этот метод и есть формальное определение понятия смыслового эталона для ситуации языкового употребления. Выделение указанных зависимостей основано на на рассмотрении текста с точки зрения символов, которые его составляют. При этом для любого текста из заданного синонимического множества выделяется изменяемую часть и некоторая неизменная часть, которая является общей для всех текстов множества. На множестве символов изменяемой части выражаются синтагматические зависимости, которые задаются с помощью синтаксических отношений и определяют возможность сосуществования словоформ в линейном ряду. На основе сочетания флексий выделяются морфологические зависимости. Поскольку указанные зависимости служат одним из способов реализации синтаксического отношения, то и само оно может быть выявлено попарным сравнением буквенного состава различных слов с выделением неизменной и флективной части.

Введя индексное множество для неизменных частей всех слов, употребленных во всех фразах заданного синонимического множества, можно определить модели линейных структур фраз. Каждая такая модель есть последовательность индексов неизменных частей слов, присутствующих в заданной фразе. Для формирования множества смысловых отношений в заданной ситуации языкового употребления необходимо найти совокупность указанных моделей, удовлетворяющих требованиям проективности. Модель линейной структуры следует считать проективной в содержательном смысле, если все стрелки выявленных синтаксических связей могут быть проведены без пересечений по одну сторону прямой, на которой записана модель. Кроме того, если из позиции некоторого индекса выходят несколько стрелок, то её не должны накрывать стрелки, выходящие из позиций других индексов. С учетом линейной природы синтагм вышеуказанные требования дополняются следующим образом. Модель линейной структуры фразы считается проективной относительно множества синтаксических отношений в заданной ситуации языкового употребления, если сумма длин всех связей относительно модели не превышает длины ее самой. При этом пара индексов, относительно которых задается связь, соответствует одной синтагме. Связь считается до-

пустимой для модели линейной структуры, если в рассматриваемом синонимическом множестве существует пара фраз, модели линейных структур которых содержат либо саму пару индексов, для которых определяется связь, либо её же, но записанную в обратном порядке.

Группировкой пар индексов, относительно которых определены связи для моделей линейных структур, формируется граф синтагм, на основе которого строится синтаксическое дерево-прецедент заданного синонимического множества. Использованием маршрутов в данном дереве закономерности сосуществования слов в линейном ряду могут быть выявлены на основе формального контекста сочетаемости флексий. Свойства модели линейной структуры фразы, актуальные для поиска места нераспознанного предикатного слова в структуре синтаксического дерева-прецедента при наличии расщеплённых предикатных значений либо конверсивов, сформулированы и доказаны в виде двух лемм и теоремы.

Второй метод основан на построении формального контекста эталона по совокупности формальных контекстов отдельных фраз, задающих ситуацию языкового употребления. При этом формальные контексты указанной совокупности строятся по результатам разбора этих фраз внешней программой синтаксического анализа. Для отбора объектов и признаков из формальных контекстов отдельных фраз вводятся коэффициенты сжатия информации относительно формального контекста ситуации языкового употребления. Помимо максимизации указанных коэффициентов, ключевым требованием при отборе признаков в результирующий формальный контекст эталона является то, что его объекты должны обладать признаками только из формируемого признакового множества, а каждый признак описывать минимум один объект. Данный метод актуален при наличии существенных смысловых ограничений на перифразирование (например, если от обучаемого требуется сохранить авторский язык при пересказе фрагмента художественного произведения в тестах по русской литературе).

Вне зависимости от способа формирования смыслового эталона ситуации языкового употребления точность решения оценивается средним числом невыделенных (опущенных) признаков на один объект формального контекста сформированного эталона. Значение данного показателя будет тем выше, чем меньше частота, с которой сочетания слов в основе отношения “объект-признак” для эталона совместно встречаются в различных фразах из определяющих заданную ситуацию языкового употребления.

Качественно процесс формирования смысловых эталонов в целом характеризуется соотношением размеров тезауруса при построении его на основе формальных контекстов для всех фраз каждой ситуации языкового употребления и на основе эталонов при заданном числе ситуаций в тезаурусе.

Далее приводится пример построения смыслового эталона в виде формального контекста ситуации языкового употребления.

В сформированном формальном контексте эталона все обозначения основ слов в составе имен объектов и признаков могут быть заменены переменными, а для каждой переменной задана конкретизация некоторой основой. В

этом случае преобразованный указанным образом формальный контекст есть шаблон формального контекста эталона. Аналогичные замены производятся для каждой фразы исходного множества, где отдельное слово при этом представлено парой “основа-флексия” по результатам выделения эталона.

Совокупность шаблонов формальных контекстов известных смысловых эталонов может быть использована для построения потенциально возможных эталонов на множестве ситуаций языкового употребления в заданной предметной области. При этом на основе каждого такого шаблона выделяется набор синтаксических отношений, в примере из презентации для описания отдельного отношения используется составной объект *d_synt_rel* языка Пролог. Кроме того, на основе сформированного набора синтаксических отношений строятся описания возможного присутствия в анализируемой фразе пар отношений, связывающих нераспознанное предикатное слово со словами, непосредственно зависимыми от него. В том же примере указанные связи представлены составным Пролог-объектом *d_no_marked*, где первые два компонента есть идентификационные номера синтаксических отношений, а последний – идентификационный номер шаблона СЯУ. Структуры *d_synt_rel* и *d_no_marked* при наличии соответствующих конкретизаций переменных служат основой рекурсивного построения формального контекста смыслового эталона, что актуально для “воссоздания” языковых описаний тех фактов предметной области, которые не были учтены разработчиками теста. Пример – смысловой эталон для предложения “Нежелательное переобучение служит причиной заниженности средней ошибки на тренировочной выборке”.

Далее рассматривается архитектура системы контроля знаний с применением тестовых заданий открытой формы. Каждое задание есть совокупность вопроса и шаблона формального контекста эталона “правильного” ответа плюс конкретизации для переменных. После ввода ответа система делает попытку применить шаблон “правильного” ответа с учетом конкретизаций в рамках задания. Если сопоставление неуспешно, делается попытка применить другие шаблоны из базы и в случае успеха – доказать схожесть ответа испытуемого и “правильного” ответа. При успешном доказательстве вычисляется оценка схожести, а полученное значение может быть использовано при выставлении испытуемым оценок, а также для сбора статистики.

Основной результат настоящей работы – метод минимизации базы знаний, используемых для численной оценки смысловой схожести высказываний предметно-ограниченного подмножества естественного языка. Применение предложенного метода позволяет уменьшить размер используемой базы в среднем на 40–50%. Синтаксический разбор на основе наиболее вероятных связей даёт высокую (менее 2% ошибок) точность выделения связей “объект – признак” независимо от ограничений на перифразирование.

Дальнейшие исследования связаны с более глубокой проработкой проблем информативности, полноты и репрезентативности исходного текста, анализа параметров формального контекста для отдельного текста и для тезауруса, выработкой критериев полноты и совершенности формируемых знаний, а также релевантности используемых лексико-синтаксических шаблонов.