Машинное понимание текстов в общей задаче распознавания образов.

Михайлов Д. В., Емельянов Г. М.

Новгородский Государственный Университет имени Ярослава Мудрого

Выделение классов Смысловой Эквивалентности (СЭ) есть неотъемлемая составляющая задачи машинного понимания текста Естественного Языка (ЕЯ). В терминологии распознавания образов понимание текста есть его соотнесение с одним из заданных смысловых эталонов и оценка меры близости смысла текста найденному эталону.

В данном докладе рассматриваются три вопроса, представленные *задачами* на *Плакате 1*:

- формирование смыслового эталона, который в конечном итоге и соответствует классу СЭ;
- структура и наполнение базы данных (БД) смысловых эталонов для заданной предметной области;
- количественная мера семантической близости ЕЯ-высказываний относительно БД эталонов.

Основой формирования смыслового эталона в настоящей работе служит Ситуация Языкового Употребления (СЯУ). Содержательно СЯУ есть описание некоторого факта действительности множеством СЭ-фраз. Сказанное позволяет отождествлять СЯУ с ситуацией СЭ для заданного естественного языка. Произвольность форм описания СЯУ дает возможность использовать в качестве таких форм деревья синтаксического подчинения, а сам прецедент класса смысловой эквивалентности (смысловой эталон) формировать по результатам синтаксического разбора фраз СЭ-множества. Поскольку установить факт СЭ означает доказать идентичность ролей сходных понятий относительно сходных ситуаций, описываемых сравниваемыми текстами, то наиболее приемлемым вариантом множества объектов для класса эталона будет множество основ слов, синтаксически подчиненных другим словам. При этом признаки содержательно характеризуют тип связи главного и зависимого слова, который определяется сочетаниями основ и флексий зависимого и главного слова, а также связями «основа–флексия» для зависимого и главного слова, соответственно.

Выделением указанных связей между объектами и признаками класса СЭ формируется набор характеристических функций, которые и задают смысл каждой из СЭ-фраз, определяющих СЯУ. Следует отметить, что именно характеристические функции смысла следует рассматривать в качестве отношений, непосредственно определяющих данные в БД эталонов. Для формирования и кластеризации указанных отношений предлагаемым в работе методом языковому контексту СЯУ ставится в соответствие формальный контекст, представленный на *Плакате 2* и используемый в теории Анализа Формальных Понятий (АФП). При этом множеству объектов класса эталона ставится в соответствие множество G, множеству признаков класса эталона — множество M. Сам АФП как расширение теории решеток является инструментом концептуальной кластеризации, поскольку Формальные Понятия (ФП) решетки являются классами с заданной в виде содержания понятий интерпретацией. При этом, как и в классической задаче распознавания образов,

выявляемые классы СЭ различаются степенью абстракции, которая зависит от частоты употребления главных слов анализируемых сочетаний в различных синтаксических контекстах.

Следует отметить, что для оценки близости ЕЯ-высказывания эталону значимы классы одного уровня абстракции, соответствующие подчинению существительных, обозначающих участников ситуации, тем словам, которые ее называют и не входят в Расщепленные Предикатные Значения (РПЗ). Содержательно РПЗ есть совокупность вспомогательного глагола (связки) и некоторого существительного, называющего ситуацию. Представленное в виде *теоремы* на *Плакате 3* правило исключения объектов и признаков РПЗ из состава формального контекста СЯУ очевидным образом вытекает из свойств базиса импликаций последнего.

После удаления информации РПЗ формальный контекст СЯУ отражает классы смысловых отношений, определяемых ролями участников описываемой ситуации действительности по отношению к ней самой. Тем не менее, число форм языкового описания СЯУ изначально не оговаривается. Фактически это означает, что слова-синонимы могут обозначать понятия с различной степенью абстракции. На практике данная степень тем выше, чем больше число СЯУ, относительно которых понятие фигурирует в некоторой фиксированной роли. Указанный факт может служить основой определения меры близости для СЯУ, порождаемых независимо друг от друга. При этом сама мера основана на использовании теоретико-решеточного представления ситуации в качестве информационной единицы тезауруса предметной области. На *Плакате 4* показана модель тезауруса в виде формального контекста, объекты которого соответствуют ситуациям языкового употребления для заданной предметной области. В признаковое множество формального контекста тезауруса войдут признаки формального контекста для каждой СЯУ в совокупности с указаниями на объекты формальных контекстов отдельных СЯУ, связями «основа–флексия» для синтаксически зависимого слова и сочетаниями основ зависимого и главного слова. Предполагается, что информация РПЗ из множеств объектов и признаков каждой ситуации удалена заранее.

На *Плакате 5* приведено формальное *определение* отношения схожести между СЯУ, базирующееся на введенной модели тезауруса. Данное отношение будет иметь место, если для каждого объекта в составе формального контекста СЯУ анализируемого высказывания найдется объект-прообраз (слово-синоним) в формальном контексте эталона, характеризующийся сходством флективной и лексический сочетаемости. Причем указанные виды сочетаемости рассматриваются как относительно формальных контекстов эталона и анализируемой СЯУ (*Условие 1)*), так и с привлечением формального контекста тезауруса (*Условия 2–4)*)). Рассматриваемое определение схожести СЯУ отражает случаи синонимии среди слов, синтаксически главных по отношению к сравниваемым (*Условие 2) и 3)*), в том числе с учетом родо-видовых отношений (*Условие 4)*) и, следовательно, учитывает степень абстракции понятий, обозначаемых словами-синонимами. При этом анализ схожести СЯУ включает сравнение последовательностей двух и более соподчиненных слов. Пример: *«средняя ошибка на обучающей выборке»*$\Leftrightarrow$*«эмпирический риск»*. Синонимические преобразования ЕЯ-фраз не меняет состав таких последова-

тельностей. Выполнимость *Определения 4* анализируется только для главных слов (в примере это *«ошибка»* и *«риск»*). Последовательности считаются взаимно заменяемыми, если возможно их построение по формальному контексту тезауруса на наборе признаков «главное-основа:» для одной и той же СЯУ. При этом главные слова последовательностей должны одинаково подчиняться одному и тому же слову, что проверяется по сочетанию флексий.

Для СЯУ, формальные контексты которых отвечают *Определению 4*, мера близости вычисляется по *формуле (5)* из представленных на *Плакате 6*. Содержательно мера близости СЯУ будет определяться числом признаков, которые разделяются объектами сравниваемых СЯУ относительно формального контекста тезауруса. Чем больше слов могут быть синтаксически главными по отношению к каждому из слов сравниваемой пары, тем выше значение указанной меры. При наличии в структуре формального контекста анализируемой СЯУ хотя бы одного объекта, для которого нет выполнимых условий *Определения 4*, мера ее близости эталонной СЯУ считается равной нулю.

В качестве примера рассмотрим ЕЯ-описание факта связи между *переобучением* и *эмпирическим риском*. Факты предметной области «Математические методы обучения по прецедентам», использованные для генерации тезауруса, приведены в *Таблице 1* на *Плакате 7*. Пусть заведомо корректное («эталонное») ЕЯ-описание связи *переобучения* и *эмпирического риска* описывается четырьмя синонимичными простыми распространенными предложениями русского языка *(Таблица 2, Плакат 8)*. Предложения 1 и 2: *«Переобучение (=переподгонка) приводит к заниженности эмпирического риска»*. Предложения 3 и 4: *«Заниженность эмпирического риска связана с переподгонкой (=переобучением)»*. Допустим, имеются четыре анализируемых варианта СЯУ. Первые три из них связаны отношением схожести с эталоном согласно *Определению 4* и описывает тот же самый факт связи *переобучения* и *эмпирического риска*, но посредством одного предложения. Первый вариант: *«Заниженность средней ошибки на обучающей выборке связана с переобучением»*. Второй вариант: *«Заниженность средней ошибки на обучающей выборке связана с переподгонкой»*. Третий вариант: *«Переобучение приводит к заниженности средней ошибки на обучающей выборке»*. Четвертый же вариант не только не излагает сути рассматриваемого факта, но и является некорректным с предметной точки зрения: *«Заниженность средней ошибки на обучающей выборке приводит к эмпирическому риску»*. На основе синтаксического разбора ЕЯ-фраз программой «Cognitive Dwarf» (ООО «Когнитивные технологии», http://cs.isa.ru:10000/dwarf), выделяются основы, флексии и их сочетания, строятся формальные контексты эталонной и анализируемых СЯУ. Как видно из *Таблицы 3* на *Плакате 9*, наибольшее значение близости эталону имеет *Вариант 1* из анализируемых СЯУ. Причина в том, что признаки объектов формального контекста для этого варианта разделяются большим количеством объектов формального контекста эталона, чем признаки у объектов формальных контекстов для *Вариантов 2 и 3*. Иными словами, признаки для *Варианта 1* являются более стереотипическими по отношению к эталону, чем признаки у других вариантов. У *Варианта 4*, как и следовало ожидать, значение близости эталону равно нулю.

В заключении следует отметить, что предложенная модель тезауруса поз-

воляет за счет иерархического представления информации сократить как размер базы данных эталонов, так и время поиска в ней. При этом сжатие информации будет тем выше, чем более релевантным заданной предметной области является каждое представленное в решетке ЕЯ-описание некоторого факта. Количественные оценки полноты охвата языкового описания предметных знаний в решетке тезауруса заслуживают отдельного прикладного исследования.