Автоматизация накопления знаний о синонимии и семантическая схожесть текстов предметного языка.

Михайлов Д. В., Емельянов Г. М.

Новгородский Государственный Университет имени Ярослава Мудрого

Актуальная *глобальная задача*, которой посвящена настоящая работа — автоматизация накопления знаний о взаимодействии семантики, синтаксиса и морфологии при установлении семантической эквивалентности (СЭ) текстов предметно-ориентированного подмножества естественного языка (ЕЯ).

Выделение класса СЭ высказывания является *важнейшей составляющей* компьютерного анализа его смысла. В общих чертах СЭ означает идентичность ролей сходных понятий относительно сходных ситуаций, описываемых сравниваемыми текстами. Следует отметить, что вне зависимости от степени абстракции класс СЭ описывается тройками «объект–признак–значение». Наиболее естественным является выделение указанных сочетаний на основе синтаксического контекста существительного относительно предикатного слова (глагола). При этом объекты будут соответствовать зависимым словам, признаки — синтаксически главным словам, значения признаков — типу связи главного и зависимого слова, который определяется:

- сочетанием изменяемых (флективных) частей слов в совокупности с предлогом, посредством которого главное слово связывается с зависимым;
- сочетанием неизменяемых частей зависимого и главного слова.

Разработка и исследование методов формирования и кластеризации знаний о синонимии с определением меры семантической схожести между ЕЯ-текстами является конечной *целью* настоящей работы.

Основные задачи, решение которых предполагает достижение указанной цели, представлены на *Плакате 1*. В нашем докладе мы более подробно остановимся на *формировании семантических отношений* как *основы знаний о синонимии*, а также *методе анализа семантической схожести* текстов.

Предлагаемое решение *первой* из задач основано на закономерностях выражения смысла носителем ЕЯ. Базовым здесь является понятие ситуации употребления ЕЯ как основы его генезиса. В общем случае под ситуацией языкового употребления (СЯУ) понимается описание некоторого факта заданной предметной области множеством СЭ-фраз. Формально языковой контекст, фиксируемый СЯУ, задается тройкой, представленной на *Плакате 2*.

Произвольность формы описания СЯУ дает возможность использовать в качестве таких форм деревья синтаксического подчинения. Известно, что синтаксические отношения задают синтагматические зависимости, которые определяют возможность сосуществования словоформ в линейном ряду. Следовательно, синтаксические деревья могут быть представлены в линейной форме, что позволяет идентифицировать отношения между словесными обозначениями понятий посредством выделения в буквенном составе каждого слова каждого текста неизменной и изменяемой (флективной) части. На множестве символов последний выражаются синтагматические зависимости, которые в силу показанной выше особенности признаков классов СЭ и составляют основу выявления *семантических отношений*. При этом модели линейных структур ЕЯ-фраз заданного синонимического множества строятся

введением *индексного множества* для неизменных частей всех слов, употребленных более чем в одной фразе. Порядок индексов в модели идентичен порядку следования соответствующих слов в предложении, что позволяет однозначно восстановить ЕЯ-фразу на множестве всех слов из всех фраз синонимического множества. И наоборот, для любой из синонимичных ЕЯ-фраз на заданном индексном множестве можно однозначно построить модель.

Для формирования искомого множества отношений между словесными обозначениями понятий в заданной СЯУ *требуется найти* совокупность указанных моделей, удовлетворяющих *требованиям проективности. Модель линейной структуры ЕЯ-фразы* 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 эту позицию не должны накрывать стрелки, выходящие из позиций других индексов. С учетом *линейной природы синтагм* вышеуказанные требования дополняются следующим образом. Будем считать, что *модель* линейной структуры ЕЯ-фразы *проективна* относительно множества синтаксических отношений в заданной СЯУ, 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 *допустимой* для модели, если в рассматриваемом синонимическом множестве существует пара фраз, модели линейных структур которых содержат в качестве подпоследовательности либо саму пару индексов, для которых определяется связь, либо ее же, но записанную в обратном порядке.

На *Плакате 3* представлены выделенные нами свойства моделей линейных структур ЕЯ-фраз, актуальные для идентификации:

- слов-синонимов;
- словесных обозначений непосредственных участников ситуации, отождествляемой с предикатным словом;
- опорных существительных в составе генитивных конструкций;
- слов-наречий и слов-прилагательных

с учетом возможного неприсутствия слова во всех ЕЯ-фразах рассматриваемого синонимического множества. При этом удвоенная длина общей неизменной части пары слов всегда больше суммы длин флективных частей.

Сформированное таким образом множество связей относительно моделей линейных структур синонимичных ЕЯ-фраз отражает сочетаемость:

- основ синтаксически главных и зависимых слов. Данный вид отношений необходим для выделения объектов и признаков во всех видах синонимии;
- флексий главных и зависимых слов;
- слова и его производных в рамках лексико-функциональной синонимии, обусловленной варьированием абстрактными словами и их сочетаниями.

Сами семантические отношения при этом составляют основу классификации и вычисления меры схожести СЯУ.

Задача классификации и анализа схожести СЯУ наиболее эффективно решается методами анализа формальных понятий (АФП). Модель СЯУ в виде *формального контекста (ФК)* на *Плакате 4* естественным образом согласуется с рассмотренным описанием класса СЭ тройками «объект–признак–

значение». При этом в решетке формальных понятий (ФП) для формального контекста СЯУ выделяются классы семантических отношений по сходству:

- основы главного слова, что особенно актуально для исследования сочетаемости в рамках лексических функций (ЛФ)-параметров, посредством которых описываются расщепленные предикатные значения (РПЗ);
- 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.

Указанные классы формальных понятий в решетке СЯУ различаются степенью абстракции, которая зависит от частоты употребления главных слов анализируемых сочетаний в различных синтаксических контекстах. Заметим, что для количественной оценки СЭ значимы классы одного уровня абстракции, соответствующие подчинению существительных, обозначающих участников ситуации, тем словам, которые ее называют и не входят в РПЗ. Содержательно само РПЗ есть совокупность вспомогательного глагола (связки) и некоторого существительного, называющего ситуацию. Редукция ФК СЯУ основана на сформулированной нами *теореме*, представленной на *Плакате 5*.

Пример исходной решетки формальных понятий СЯУ для описания факта связи между переобучением и эмпирическим риском, представлен на *Плакате 6*. Исключая объекты и признаки слов РПЗ, получаем редуцированный ФК, решетка формальных понятий для которого представлена на *Плакате 7*.

После удаления информации РПЗ формальный контекст СЯУ отражает классы 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, синонимичные по *Лемме 1*, могут обозначать понятия с различной степенью абстракции. На практике указанная степень тем более, чем больше число СЯУ, относительно которых понятие фигурирует в некоторой фиксированной роли.

Возьмем указанный факт за основу определения меры схожести для ситуаций языкового употребления, порождаемых независимо друг от друга.

Согласно определению, синтаксические зависимости как частный случай семантических отношений выражаются определенными сочетаниями флексий. Сказанное позволяет в ряде случаев выделять основы и их сочетания на базе указанных морфологических зависимостей. Эти зависимости могут быть либо выявлены ранее для других СЯУ, либо найдены с помощью программ синтаксического анализа, реализующих стратегию разбора на основе наиболее вероятных связей слов. Фактически данные связи и выделяет модель, представленная на *Плакатах 2* и *3*.

Ставится *задача* накопления и систематизации знаний, представляемых формальными контекстами СЯУ. При этом предполагается, что из множеств объектов и признаков каждой ситуации удалена информация расщепленных предикатных значений. Если указанные знания формируются на основе *независимых* ЕЯ-описаний различных *фактов* некоторой *предметной области*, то получаемая структура будет соответствовать *тезаурусу* этой *предметной*

области. Модель такого тезауруса в виде *ФК* представлена на *Плакате 8*. На этом же *плакате* дан пример *формального понятия* в решетке тезауруса, которое соответствует *формальному контексту СЯУ* на *Плакате 7*.

Теоретико-решеточное представление тезауруса позволяет определить *отношение схожести между СЯУ*. Приведенное на *Плакате 9* 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 последовательностей. Выполнимость условий *Определения 6* анализируется только для главных слов (в приведенном примере это *«ошибка»* и *«риск»*). Последовательности считаются взаимно заменяемыми, если возможно их построение по формальному контексту тезауруса на наборе признаков с префиксом «главное-основа:» для одной и той же СЯУ. При этом главные слова последовательностей должны быть одинаково подчинены одному и тому же слову, что проверяется по сочетанию флексий.

С учетом выполняемого согласно *Определению 6* сопоставления формальных контекстов, мера схожести СЯУ вычисляется по *формуле (5)* из представленных на *Плакате 10*. При этом схожесть СЯУ тем выше, чем большее число признаков разделяются объектами сравниваемых СЯУ относительно формального контекста тезауруса (чем большее число слов могут быть синтаксически главными по отношению к каждому из слов для сравниваемой пары). *Формула (6)* на *Плакате 10* отражает взаимную специфичность понятий, обозначаемых сравниваемыми словами в СЯУ.

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 11*. Полученная модель тезауруса в виде решетки формальных понятий представлена на *Плакате 12*. В целях сравнения уровней абстракции примеры классов, определяемых сочетаниями флексий, выделены прямоугольниками.

Пусть заведомо корректное («эталонное») ЕЯ-описание связи *переобучения* и *эмпирического риска* задается четырьмя синонимичными простыми распространенными предложениями русского языка *(Плакат 13)*. Предложения 1 и 2: *«Переобучение (=переподгонка) приводит к заниженности эмпирического риска»*. Предложения 3 и 4: *«Заниженность эмпирического риска связана с переподгонкой (=переобучением)»*. Выполнив синтаксический разбор программой «Cognitive Dwarf» (ООО «Когнитивные технологии», http://cs.isa.ru:10000/dwarf), выделяем основы, флексии и их сочетания. Получаем формальный контекст, представленный решеткой на *Плакате 14*.

Теперь предположим, что мы имеем три анализируемых независимых варианта СЯУ. Каждый из них связан отношением схожести с эталоном согласно *Определению 6* и описывает тот же самый факт связи *переобучения* и *эмпирического риска*, но посредством одного предложения *(Плакат 13)*.

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»*. Как и для «эталонного» варианта, ФК СЯУ здесь строятся на основе результатов синтаксического разбора предложений программой «Cognitive Dwarf». Полученные решетки ФП представлены на *Плакатах 15, 16* и *17*.

Как видно из *Таблицы 3* на *Плакате 18*, наибольшее значение схожести с эталоном из анализируемых ЕЯ-описаний заданного факта предметной области имеет *Вариант 1*. Причина состоит в том, что признаки объектов формального контекста для этого варианта разделяются большим числом 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 двух других вариантов.

Немаловажную роль при вычислении схожести СЯУ играет полнота и непротиворечивость описания предметных знаний при формировании тезауруса. Теоретико-решеточная модель последнего позволяет задействовать, в частности, базис импликаций ФК для изучения взаимозаменяемости абстрактных слов в синтаксических контекстах предметных существительных. Пример: *(«связана с переобучением»⇔«переобучение приводит (к)»)*.

Для количественной же оценки полноты охвата языкового описания предметных знаний могут быть использованы коэффициенты сжатия информации по основам и флексиям, представленные на *Плакатах 19*, *20* и *21*. Как видно из приведенных на *Плакатах 22*, *23* и *24* примеров результатов расчета указанных коэффициентов, сжатие текстовой информации в решетке будет тем выше, чем более релевантным заданной предметной области является каждое представленное в решетке ЕЯ-описание некоторого факта. Требуемая точность вычисления схожести СЯУ предложенным в работе методом достигается добавлением синонимичных перифраз для каждой из рассматриваемых СЯУ с максимизацией указанных коэффициентов. Причем пополнять множество форм языкового описания совсем не обязательно именно для СЯУ, соответствующих сравниваемым ЕЯ-высказываниям. Унифицируемое теоретико-решеточное представление анализируемых СЯУ и тезаурусной информации позволяет предельно просто пополнять сам тезаурус с максимизацией коэффициентов сжатия информации. Сказанное особенно актуально для задач тестирования знаний, когда разработчик теста открытой формы формулирует один или несколько вариантов «правильного» ответа, опираясь на собственные знания по заданной предметной области. При этом факт, описываемый «правильным» ответом, может отсутствовать в тезаурусе.

В заключении отметим, что модель тезауруса в виде решетки формальных понятий обеспечивает сжатие текстов (в первую очередь) за счет тех предикатных слов, которые обозначают ситуации, сходные в той или иной мере по составу участников и характеру выполняемых ими действий, а также за счет абстрактной лексики. Качественные же особенности сжатия текстов различных жанров заслуживают отдельного прикладного исследования.