УДК 681.3.06

**Д.В.Михайлов, Г.М.Емельянов**

# К ВОПРОСУ АВТОМАТИЗАЦИИ ПОПОЛНЕНИЯ БАЗЫ ДАННЫХ ЛЕКСИЧЕСКИХ ФУНКЦИЙ В ЗАДАЧЕ УСТАНОВЛЕНИЯ СМЫСЛОВОЙ ЭКВИВАЛЕНТНОСТИ ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА

*Институт электронных и информационных систем НовГУ*

The problem of machine learning to identify the situations of semantic equivalence of statements in natural language, each of which is described with an involment of standard lexical functions apparat is presented here. Special attention is given to automation of identification of dependences between arguments and lexical functions semantics.

**Постановка проблемы**

Одним из наиболее лингвистически перспективных путей решения задачи установления смысловой (семантической) эквивалентности высказываний естественного языка (ЕЯ), в частности русского, является привлечение знаний о той синонимии языковых конструкций [1], которая описывается с привлечением аппарата стандартных лексических функций (ЛФ, ЛФ-синонимия) [2]. Действительно, богатое словесное варьирование присуще только небольшому числу смыслов, которые и выделяются в качестве стандартных лексических функций-параметров [3]. Как следует из данного И.А.Мельчуком [4] определения, этот вид синонимии выражений ЕЯ характеризуется следующими особенностями.

1. Глубинным синтаксическим структурам (ГСС) сравниваемых высказываний соответствуют одни и те же (или эквивалентные) семантические представления (СемП) [5].

2. В семантическом графе (СГ) СемП выделяются подграфы (пучки) и каждому подграфу СГ будет соответствовать свое поддерево ГСС каждого из сравниваемых высказываний.

3. Существует как минимум один подграф СГ, который будет по-разному отображаться в глубинных синтаксических структурах каждого из сравниваемых высказываний. Иными словами, один и тот же смысл в разных ГСС выражается разными обобщенными лексическими единицами [6] рассматриваемого ЕЯ. Но при этом перераспределение смысла между лексемами сводится к минимуму [4], а смысловые соотношения между цельными лексическими единицами описываются с помощью аппарата стандартных ЛФ.

В силу регулярности стандартных ЛФ и операций над ними ЛФ-синонимические отношения между ГСС оказываются более регулярными и однотипными, нежели произвольные синонимические отношения между ГСС [4]. ЛФ-синонимические отношения между ГСС могут быть описаны с помощью специального исчисления в виде системы правил, которая любой данной ГСС ставила бы в соответствие все другие ГСС, ЛФ-синонимичные с ней. ЛФ-синонимические преобразования ГСС являются частным случаем преобразований деревьев. К настоящему моменту отечественными исследователями разработан корректно формализуемый математический аппарат для формального описания процесса переработки помеченных деревьев — Δ-грамматики [7] и их разновидности [8]. В частности, в [8] был доказан ряд свойств расширенных лексико-синтаксических Δ-грамматик, актуальных для моделирования рассматриваемых преобразований глубинных синтаксических структур ЕЯ.

Тем не менее, практическая реализация описанной в [2] системы правил ЛФ-синонимических преобразований при разработке основ теории «Смысл⇔Текст» была изначально задумана для синтеза фраз по заданному СемП. Обратная процедура или семантический анализ была упомянута [9] как получение по исходной последовательности ГСС фраз анализируемого текста его СемП и не рассматривалась вовсе. Система перифразирования представлялась как генератор, выдающий для заданной ГСС конечное множество ГСС, ЛФ-синонимичных с ней [10]. При этом для каждой глубинной синтаксической структуры, получаемой непосредственно из СемП будущего текста и выбираемой в качестве базового представителя множества ЛФ-синонимичных ГСС — базовой ГСС (БГСС) [11], с помощью имеющихся правил и наличного словаря строится множество деревьев глубинного синтаксиса, ЛФ-синонимичных с исходной БГСС [12].

Следует отметить, что постановка рассматриваемой нами задачи анализа двух фраз на ЛФ-синонимию как построение для каждой анализируемой фразы ЛФ-синонимического множества ГСС с последующим поиском ненулевого пересечения полученных множеств наименее оптимальна с точки зрения вычислительной сложности. При таком подходе будут актуальны все проблемы, характерные для решения задач методом поиска в пространстве состояний. Помимо этого, в силу отсутствия по-настоящему формализованного описания системы, затруднительным будет доказательство конечности процесса перифразирования. Конечность рассматриваемого процесса может быть обоснована при текущем состоянии вопроса ссылками на некоторые не вполне формальные соображения, делающие свойство конечности процесса перифразирования весьма вероятным (прежде всего — с учетом результатов эксперимента), но не обязательным при любых условиях [13]. Кроме того, значительную трудность при практической реализации указанных преобразований представляет формализация особого компонента правила, именуемого условием его применимости. В содержательном плане условие применимости лексического правила представляет собой совокупность требований к синтаксическим и семантическим свойствам лексических единиц исходной ГСС, входящих в заменяемое правилом поддерево [14]. Подобные ограничения отражают особенности лексики конкретного ЕЯ и выполняют функции фильтров [15], задерживающих синтез определенной фразы из множества семантически эквивалентных, если конечный продукт синтеза дает нарушение лексического значения, сочетаемости или стилистических норм. Многие фильтры были сформулированы в работах И.А.Мельчука, И.А.Жолковского. Однако, как отметил акад. Ю.Д.Апресян [16], проблема нуждается в дальнейшей разработке. Тем более, что, по оценке [2], специальных исследований по данному вопросу не проводилось, а сами правила синонимических преобразований ГСС с применением аппарата стандартных ЛФ описаны в первом приближении. Как писал сам И.А.Мельчук, «во многих случаях лексическое правило не является точной эквивалентностью: между его левой и правой частями может иметься ощутимое смысловое различие, которым мы пренебрегаем» [17].

Рассмотрим содержательную особенность

применения системы правил ЛФ-синонимических преобразований для нашей задачи. Как следует из сформулированного в [2] определения БГСС [18], для доказательства ЛФ-синонимичности двух произвольных фраз необходимо и достаточно построить ГСС каждой из них и доказать, что им соответствует одна и та же БГСС как базовый представитель некоторого ЛФ-синонимического множества. В силу своей каноничности [19] БГСС в общем случае ближе к СемП, чем произвольная ГСС. В настоящей работе путем использования свойств базовых ГСС, определяемых их близостью СемП, за счет формализации условия применимости лексического правила, мы рассмотрим возможность уйти от используемого в [8] перебора при установовлении ЛФ-синонимии фраз ЕЯ.

**Два этапа машинного обучения**

Поставим *задачу доказательства ЛФ-синонимии фраз как частный случай* ранее рассмотренной нами *задачи концептуально-ситуационного моделирования процесса перифразирования* [20]. *Переформулируем* общую задачу обучения распознаванию ситуаций смысловой эквивалентности текстов ЕЯ как *двухэтапное обучение*. Первый этап — *обучение распознаванию ЛФ-синонимии* при *использовании лексических синонимических конструкций* (ЛСК) [21], заменяемых лексическими правилами, *в качестве форм поверхностного выражения* ситуации синонимической замены. Сама ситуация формально описывается лексическими правилами синонимических замен относительно заданного ключевого слова [22].

В ходе машинного обучения мы будем использовать информацию СемП при формировании того, что именуется условием применимости правила. Фактически условие применимости правила определяет результат обобщения того, что нами в [20] было рассмотрено как ситуация языкового употребления. С учетом показанных выше свойств БГСС формирование множества прецедентов как известных системе фактов смысловой эквивалентности [20] следует начинать именно с тех прецедентов, которые соответствуют «переходу» к БГСС на уровне глубинного синтаксиса. Как показано И.А.Мельчуком [23], базовые ГСС при синтезе текста строятся на основе СемП фраз с применением семантико-языкового словаря (о содержательной стороне наполнения этого словаря см. [24]). Это позволяет просто и естественно реализовать переход ГСС⇒Концептуальное представление (при рассмотрении ГСС как формы поверхностного выражения ситуации [20]) с учетом показанного в [25] соответствия между множеством выражений концептуальных языков и единицами информационного уровня, именуемыми в лингвистике семами и образующими базовый алфавит СемП [26]. При рассмотрении пар деревьев, преобразуемых правилами и рассматриваемых как прецеденты, описанное в [20] обучение следует вести именно в направлении от базовой ГСС, рассматривая в «обратном» направлении все пути ее получения.

Второй этап задачи обучения распознаванию ситуаций смысловой эквивалентности текстов ЕЯ — *обучение распознаванию более сложных видов синонимии ГСС*, которую не удается описать с помощью ЛФ и которая выявляется при установлении соответствия БГСС⇔СемПФ [27]. В настоящей работе нами изучается первый этап как попытка по-настоящему формализовать описание системы перифразирования для конкретной практической задачи — доказательства наличия/отсутствия ЛФ-синонимии двух произвольных фраз ЕЯ. Следует отметить, что для этого этапа мы анализируем именно рассматриваемую в [2] ситуацию смысловой эквивалентности относительно варьирования ЛФ (семантику самой ситуации ЛФ-синонимии), сведя к необходимому минимуму привлечение концептуальных знаний. Второй этап — тема отдельного исследования и в данной работе не затрагивается. Как и в наших предыдущих работах, мы ведем рассмотрение на материале русского языка как одного из наиболее употребляемых, для которых характерно отсутствие жесткого ограничения на синтаксическое строение предложения.

Другая немаловажная задача, которой мы касаемся в настоящей работе, заключается в уточнении и конкретизации по итогам машинного обучения общих требований к базовой лексике, сформулированных в [2].

**Формулирование требований к правилам синонимических замен**

Рассмотрим ключевые свойства глубинной лексики, используемые как критерии при отборе последней в качестве базовой и актуальные для выделения тех правил ЛФ-синонимических преобразований, посредством которых теоретически может быть описан переход к базовой ГСС. Из свойств базовой лексики наибольший интерес для нас представляют следующие:

— базовые лексические единицы интуитивно должны быть как можно более элементарными;

— значение базовых лексических единиц не допускает омонимии и полисемии;

— отсутствие расщепленных глаголов; это требование может быть удовлетворено благодаря наличию среди базовых лексем глубинных фиктивных глаголов, посредством которых создается глагольное оформление предикатного (ситуационного) значения, для которого в рассматриваемом ЕЯ имеется только именное значение. Примеры: *ВАХТИТЬ = стоять (быть) на вахте, *КОМПРОМИССИРОВАТЬ = идти на компромисс (см. [11]).

Естественным при отборе правил также является требование сбалансированности дерева-результата: высоты каждого поддерева одинаковы или отличаются от высот других поддеревьев рассматриваемого узла не более чем на 1. Как показано в [8], основная идея генерации БГСС состоит в попытке внести упорядоченность в синтаксическую структуру фразы. Из всех ЛФ-синонимичных ГСС базовой считается такая, у которой в порядке очередности заполнены узлы, соответствующие одному из шести возможных актантов: вершина дерева интерпретируется как сказуемое, первый актант — как подлежащее, второй актант — как прямое дополнение, третий и четвертый актанты представляют дополнения в роли поясняющих компонен-

тов, пятый актант — определение или обстоятельство, шестой актант — соседний член однородного ряда.

Благодаря близости базовых ГСС семантическому представлению фраз использование базовой лексики на упомянутом выше втором этапе задачи обучения позволяет значительно сократить объем обучающей выборки за счет упомянутого разделения синонимии: знания, относящиеся к ЛФ-синонимии оказываются уже полученными машиной на предыдущем этапе.

**Семантика лексических функций и ситуации смысловой эквивалентности**

Во многих случаях использование БГСС решает немаловажную проблему предикативного (глагольного [28]) обозначения имени ситуации в анализируемой ГСС одной лексемой при формировании обучающей выборки для упомянутого выше второго этапа обучения распознаванию ситуаций смысловой эквивалентности. Это достигается, в частности, за счет использования в правилах фиктивных лексем. В исходной для перифразирования глубинной синтаксической структуре мы фактически сопоставляем лексической синонимической конструкции как особому словосочетанию уровня глубинного синтаксиса относящиеся к ситуации семантические актанты [29].

Будем рассматривать ЛФ-синонимию с позиций введенной и рассмотренной нами в [20] модели ситуации языкового употребления:

$$S = (O, P, V), \qquad (1)$$

где $S$ — ситуация, фиксирующая однозначный языковой контекст. Здесь в качестве $S$ мы подразумеваем «определенное лексическое отражение некоторого куска действительности» [30], связанного с ключевым словом ЛФ-синонимической замены; $O$ — множество подразумеваемых объектов, которые соответствуют в конкретной ситуации языкового употребления денотатам. Каждый объект при этом описывает семантику лексемы в узле преобразуемого дерева глубинного синтаксиса принятым в теории К-языков способом — посредством семантических координат (так, как это делается в описанном в [25] лексико-семантическом словаре). В качестве отношений из множества $P$ выступают смысловые (= семантические) отношения, задаваемые следующими функциями, требованиями, зависимостями.

1. Представленными в узлах ЛСК лексическими функциями. Эти отношения имеют место между сущностью, обозначаемой ключевым словом ЛФ-синонимической замены (аргументом этих лексических функций), и понятиями, обозначаемыми на поверхностном уровне значениями указанных ЛФ. Более сложный вид указанных отношений — между понятиями, соответствующими значениям лексических функций от ключевого слова, с которым связывается ситуация $S$. Пример: лексические функции, которые могут иметь склеенные и несклеенные значения. Для несклеенного значения характер синтаксической деривации определяется как лексической функцией, так и самим ключевым словом ЛФ-синонимической замены [31].

2. Требованиями к распределению между элементами ЛСК актантов ситуации, обозначаемой посредством рассматриваемой ЛСК. Фактически здесь задаются типы семантических отношений между понятиями, обозначаемыми элементами ЛСК, и понятиями, соответствующими каждому из актантов рассматриваемой ситуации. В содержательной лингвистической интерпретации в роли указанных требований могут выступать грамматические запреты в области семантической сочетаемости [32].

3. Ролевыми зависимостями между элементами ЛСК и грамматически подчиненными им словами, которые не выражают актантов ситуации $S$. Здесь следует упомянуть грамматические запреты в области морфо-синтаксической сочетаемости [33].

Следует отметить особенность формирования множества $V$ альтернативных форм поверхностного выражения ситуации $S$. В [20] мы отмечали приводимость различных вариантов представления $V$ к естественному для поверхностного уровня ЕЯ представлению в линейной форме. Из упомянутых там форм поверхностного выражения содержательный интерес для нашей задачи представляют деревья глубинного синтаксиса, замену которых описывают правила, и соответствующие этим деревьям фразы ЕЯ. Сам переход между этими двумя уровнями представления форм из $V$ предполагает построение ГСС для исходных анализируемых фраз с применением информации справочников моделей управления и лексических функций. Практически во всех известных системах анализа текстов ЕЯ на основе подхода «Смысл⇔Текст» указанные справочники заполняются вручную.

В [34] нами предложена методика пополнения справочника моделей управления на основе данных лексикографического толкования предикатных слов. Поставим следующую задачу. Пусть для каждого слова ЕЯ мы имеем описание его семантики посредством лексико-семантического словаря [25], который ставит в соответствие слову единицу семантического уровня — сему [2]. Кроме того, для предикатных слов мы описываем смысловые отношения между значением глагольной формы и значениями зависящих от нее в предложении групп слов посредством словаря глагольно-падежных семантико-синтаксических фреймов. На основе семантической информации, представляемой указанными выше словарями, мы можем выявить зависимости между предикатами, описывающими связи между аргументами и значениями лексических функций. Выявленные зависимости удобнее всего представлять в виде предикатов, обобщающих указанные связи для каждой из используемых нами ЛФ. Полученные таким образом знания позволят для произвольной пары слов на основе их словарной семантической информации определить наличие зависимости, задаваемой той или иной ЛФ.

Аргументы указанных предикатов задаются упорядоченными наборами следующего вида:

$$(lec,\ pt,\ sem,\ st_1, \ldots, st_k), \qquad (2)$$

где $lec$ — элемент множества лексем заданного морфологического базиса; $pt$ — обозначение части речи лексемы $lec$; компонент $sem$ обозначает значение лексемы $lec$ (отождествляемое с ней понятие);

$st_1, \ldots, st_k$ — различные семантические координаты сущности, характеризуемой понятием *sem* .

Использование утверждений вида (2) для описания аргументов и значений лексических функций позволяет учесть тот факт, что ЛФ может быть определена только для слов [35]:

— определенной части речи (функции-параметры *Oper* , *Func* и *Labor* );

— определенной семантики (лексические функции *Magn* , *Cap* и *Equip* ),

а также то, что для ряда слов рассматриваемого ЕЯ лексическая функция может не иметь значения даже при соответствии семантических и синтаксических свойств аргумента.

Для обобщения предикатов, представляющих значения одной и той же ЛФ для разных слов, следует использовать методы анализа данных, основанные на таксономии знаний [36]. При этом будут выявлены закономерности, позволяющие найти сходства и различия первых и вторых аргументов предикатов и произвести группировку описываемых предикатами знаний в таксоны (=кластеры). Первым делом при решении задачи кластеризации следует *по-настоящему формализовать* компоненты концептуального базиса, используемые при построении упомянутых выше словарей и именуемые в [25] *сортовой системой* (она определяет совокупность $st_1, \ldots, st_k$ в (2)), и *множества типов*, которые ей порождаются. Применение математических методов формального концептуального анализа [37] (и реализующего эти методы специализированного ПО ToscanaJ [38]) позволит оценить, в частности, адекватность отношений общности и совместимости (толерантности) [25], задаваемых на используемом множестве сортов.

Тема отдельного рассмотрения — *унификация* используемых в лингвистической литературе *понятий «сорт» и «тип»* [39]. Здесь особого внимания заслуживает согласование терминологии, касающейся описания дифференциальных признаков значения слова. В частности, практический интерес представляет *сравнительный анализ* самого *понятия типа*, который рассматривается в [25] как семантическая характеристика сущности *sem* относительно некоторого концептуального базиса, и используемой в Русском общесемантическом словаре *семантической характеристики значения слова* [40]. Это позволит значительно облегчить использование существующих лингвистических информационных ресурсов при построении концептуальных базисов групп предметных областей для решения практических задач.

*Работа выполнена при поддержке РФФИ (проект №06-01-00028).*


1. Емельянов Г.М., Михайлов Д.В. // Искусственный интеллект. 2004. №2. С.86-90.
2. Мельчук И.А. Опыт теории лингвистических моделей «Смысл⇔Текст». Семантика, синтаксис. М.: Языки русской культуры, 1999. 345 с.
3. Там же. С.106.
4. Там же. С.147.
5. Там же. С.32.
6. Там же. С.178.
7. Гладкий А.В., Мельчук И.А. Грамматики деревьев. I. Опыт формализации преобразований синтаксических структур естественного языка // Информационные вопросы семиотики, лингвистики и автоматического перевода. Вып.1. М., 1971. С.16-41.
8. Emelyanov G.M., Krechetova T.V., Kurashova E.P. // Pattern Recognition and Image Analysis. 2000. Vol.10. №4. P.520-526.
9. Мельчук И.А. Цит. соч. С.177.
10. Там же. С.193.
11. Там же. С.148-149.
12. Там же. С.190.
13. Там же. С.192.
14. Там же. С.151.
15. Апресян Ю.Д. Избр. тр. Т.I. Лексическая семантика. Синонимические средства языка. М.: Языки русской культуры, 1995. С.335, 336.
16. Там же. С.336.
17. Мельчук И.А. Цит. соч. С.160.
18. Там же. С.148.
19. Там же. С.199.
20. Емельянов Г.М., Корнышов А.Н., Михайлов Д.В. // Искусственный интеллект. 2006. №2. С.72-75.
21. Emelyanov G.M., Mikhailov D.V., and Zaitseva E.I. // Pattern Recognition and Image Analysis. 2003. Vol.13. №3. P.447-451.
22. Мельчук И.А. Цит. соч. С.149.
23. Там же. С.178.
24. Там же. С.82.
25. Фомичев В.А. // Качество и ИПИ (CALS)-технологии. 2005. №3. С.30-38.
26. Мельчук И.А. Цит. соч. С.57.
27. Там же. С.177.
28. Там же. С.91.
29. Там же. С.86.
30. Там же. С.85.
31. Там же. С.151.
32. Апресян Ю.Д. Указ. соч. С.341.
33. Там же. С.343.
34. Емельянов Г.М., Михайлов Д.В. // Таврический вестник информатики и математики. 2005. №1. С.35-48.
35. Мельчук И.А. Цит. соч. С.103.
36. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: Ин-т математики СО РАН. 1999. 270 с.
37. Ganter B. and Wille R. Formal Concept Analysis — Mathematical Foundations. Berlin: Springer-Verlag, 1999. 284 p., 105 figs.
38. http://toscanaj.sourceforge.net
39. Partee B.H., Borschev V.B. Genitives, types, and sorts. In Possessives and Beyond: Semantics and Syntax / Eds. Ji-yung Kim, Yury A. Lander and Barbara H. Partee. Amherst, MA:GLSA Publications, 2004. P.29-43.
40. Леонтьева Н.Н. // НТИ. Сер.2. 1997. №12. С.5-20.