УДК 004.912: 303.7

ПРОБЛЕМЫ ПОПОЛНЕНИЯ СЕМАНТИЧЕСКОГО СЛОВАРЯ

К.К. Боярский, Е.А. Каневский

Рассмотрены проблемы пополнения компьютерного семантического словаря новыми словами, встреченными в тексте при его анализе. Предлагаемая для этого система работает в полуавтоматическом диалоговом режиме. На первом этапе определяются морфологические характеристики нового слова, на втором – его синтактико-семантические параметры по аналогам, имеющимся в существующем словаре. Предлагаемые подходы обеспечивают высокий уровень точности. Впервые появилась возможность указания точной семантики новых слов с учетом не только семантических классов, но и аргументов, обеспечивающих связь с подсоединяемыми словами.

Ключевые слова: анализ текста, лексема, морфология, семантика, синтаксис, словарь, слово.

Введение

Задаче компьютерного анализа текста на естественном языке посвящено множество теоретических и практических работ. Эти задачи, а именно – поиск документов, рубрицирование и аннотирование документов, диалог с компьютером, машинный перевод и построение баз знаний, – решали и решают различными методами, используя или не используя ту или иную дополнительную информацию. Решение любых прикладных задач, связанных с анализом естественного текста, начинается с морфологического анализа. Такой анализ еще можно проводить без использования словаря [1]. Далее может проводиться синтаксический и семантический анализы, для которых словарь крайне необходим.

Используемый авторами семантический словарь В.А. Тузова [2] основан на расширенном морфологическом словаре А.А. Зализняка [3] и представляет собой список статей (лексем), каждая из которых соответствует одному слову русского языка. При этом одному слову может соответствовать несколько лексем, выражающих различный семантический смысл. Так, например, слову *коса* соответствуют три лексемы: *девичья коса*, *береговая коса* и *острая коса*. В настоящее время словарь насчитывает 165 тысяч лексем, соответствующих 145 тысячам слов общей нормативной лексики русского языка. Лексемы сгруппированы в 1650 классов, которые образуют иерархическую структуру, отражающую родовидовые отношения между лексемами [2]. Каждая из статей словаря содержит морфологическое, синтаксическое и семантическое описания лексемы. Так для лексемы «ЗОРИН» словарная статья имеет следующий вид:

ЗОРИН $12413/03000(S1>Hab(S1:ЧЕЛОВЕК$1241,S0:ФАМИЛИЯ$1241/11)) {м11о 298}

Морфологическое описание лексемы содержится в фигурных скобках, где «м11о» (морфологический описатель, аналогичный описателю в [3] обозначает существительное мужского рода, 11-го класса, одушевленное, а число «298» – адрес соответствующих падежных окончаний в файле окончаний лексем. Идентификатор $12413/03000 обозначает принадлежность к классу (ФО/Живой/Человек/Личность/ФИО/Фамилия). В круглых скобках расположено собственно синтаксическое и семантическое описание лексемы, которое в данном случае означает «человек имеет фамилию».

Проблема заключается в том, что какого бы объема ни был словарь, при анализе очередного текста всегда обнаруживаются новые слова (НС), в данном словаре отсутствующие. Это могут быть имена и фамилии, географические названия и образованные от них прилагательные, специальные термины и слова, употребленные автором в необычном значении, словоформы, противоречащие современным правилам грамматики (например, при передаче особенностей речи персонажей) и т.д. Так, по наблюдению Т.Ю. Кобзаревой, при анализе текстов Набокова, Мандельштама, Л.Н. Толстого, Гоголя часто встречались лексически продуктивные формы, неологизмы, «аномалии», не учтенные, например, в [3] и в компьютерном словаре пакета Word-2000 [4]. Рассмотрим подробнее структуру НС в романах Гончарова.

Новые слова

В процессе предварительного анализа текста романа авторами был получен полный перечень нераспознанной лексики по трем романам И.А. Гончарова. Ниже (таблица) приведены общие данные о словах, отсутствующих в семантическом словаре [5].

Романы И.А. Гончарова	Слово-форм	НС	Из них:				
			Имен	Сущ.	Глаг.	Прил.	Нареч.
Итого	467 тыс.	1300	545 39%	273 20%	188 14%	251 18%	84 6%

Таблица. Состав новых слов романов Гончарова

Как видно из таблицы, основную проблему представляют имена собственные. Количество их доходит до 40% от всех. Для данного списка характерно как раз наличие устаревших словоформ, использующих окончания с мягким знаком (*Артемий – Артемью*, *Василий – Васильем*) и уничижительных имен (*Аверка, Васька*).

Удельный вес новых существительных почти в два раза меньше. Часть из них использует окончание *-ье* вместо современного *-ие* (*вдохновенье*, *влеченье*). Некоторые существительные использованы в устаревшем написании в корне (*бусурман*, *нумер*). В ряде существительных используются устаревшие словоизменительные формы (*крыло – крылами, чулки – чулков*). Особо следует отметить существительные, пишущиеся через дефис (*наденька-изменница, старец-классик*). Обычно в случае отсутствия таких слов в словаре программа разбивает их на два слова, что является ошибкой.

Доля новых глаголов несколько меньше. Некоторые глаголы использованы в устаревшем написании (*дотрогиваться*, *испужаться*). В ряде глаголов используются устаревшие словоформы (*воздвигнуть – воздвигнул*, *гулять – гуляючи*).

Что касается новых прилагательных, то сравнительно большая их группа, начинающаяся с приставки *не-*, в современном языке почти не употребляется (*непогрешительный*, *неупотребленный*). Небольшая группа прилагательных начинается с приставки *пре-* (*превеселенький*, *преглубокий*). Особо следует отметить прилагательные, пишущиеся через дефис (*безвинно-угнетенный, бледно-чернильный)*.

Количество новых наречий невелико. Значительная их группа начинается с приставки *по-* (*по-вчерашнему*, *по-латыни*). Следует отметить небольшую группу наречий, пишущихся через дефис (*мало-мало*, *благородно-бесполезно)*.

Во всех трех романах авторам встретилось 20 новых междометий (*м-м, тс, фу-фу)*.

Вспомогательная система Adviser

В связи с вышеизложенным возникает задача пополнения словаря. Пополнение семантического словаря само по себе является сложной задачей, для решения которой предлагаются различные способы [6]. Простейший вариант, который может быть использован для этой цели, – использование образца [1]. Однако даже правильно указать морфологические параметры иногда оказывается весьма непростой задачей. Дело в том, что при сравнительно небольшом числе вариантов грамматических категорий, сопоставляемых определяемому слову, общее количество наборов окончаний приближается к тысяче. Описать же синтактико-семантические характеристики для неспециалиста по устройству данного конкретного словаря просто нереально. Авторами разработана система Adviser, позволяющая пополнять словарь НС в полуавтоматическом диалоговом режиме.

На первом этапе определяются морфологические характеристики НС. Пользователь задает часть речи, к которой относится НС – существительное, прилагательное, глагол, наречие или междометие. Возможно также дополнительное указание ряда характеристик – одушевленность и род для существительных, совершенный/несовершенный вид глагола и т.д. После этого система позволяет подобрать из имеющихся в словаре такое слово, окончания словоформ которого совпадают с окончаниями словоформ НС.

Поскольку в систему подсказки заложены все известные варианты словоизменений, то нужный вариант обязательно найдется [3]. Исключение составляет архаическое или нарочито искаженное написание слова, например, деепричастие *завидя* от глагола *завидеть*. В этом случае используется файл замен.

После установления морфологических характеристик НС нужно задать его семантику (синтактико-семантические параметры). Прежде всего нужно определить класс по классификатору. Эта структура предъявляется пользователю в виде дерева классов. Для облегчения работы предусмотрена возможность ускоренного указания таких часто встречающихся для НС классов, как имя, отчество, фамилия, различные названия (географических объектов, фирм, документов и др.) – всего 82 класса. Кроме того, можно просто найти синоним НС. Например, к слову *вдохновенье* указать синоним *вдохновение* и сразу полу-

чить, что это слово относится к классу «Физический_объект/Живой Человек/Психика/Душа/Чувство/Депрессия-Вдохновение».

После этого пользователю предъявляется полный список слов, принадлежащих данному классу. Из них выбирается наиболее близкое по значению, и его семантика приписывается к НС. При необходимости эта семантика может быть уточнена вручную. Теперь НС с правильной морфологией и семантикой готово для занесения в словарь.

Предлагаемая система Adviser апробирована на массиве более 1000 слов и показала прекрасные результаты. Следует отметить, что без системы такого рода составление подсловаря на 1300 слов потребовало бы значительно больше времени.

Определение морфологии НС

На первом этапе определяются морфологические характеристики НС. Методика их определения основана на использовании обратного словаря [7]. Как известно, наиболее приемлемой в данном случае является почти полностью автоматизированная процедура склонения и спряжения, реализованная в виде диалога с пользователем. Считается, что достаточно ограничиться четырьмя знаменательными частями речи: существительными, прилагательными, глаголами и наречиями.

Система определения морфологических характеристик НС достаточно проста. Пользователю предлагается три окна и набор кнопок (рис. 1). Вручную или из заранее подготовленного файла новое слово вводится в среднее окно. Затем пользователь выбирает часть речи, к которой относится НС – существительное, прилагательное, глагол, наречие или междометие (по нашему мнению, междометие также заслуживает того, чтобы быть включенным в систему). Как показала практика, среди других частей речи НС практически не встречаются. Возможно также дополнительное указание ряда характеристик: одушевленность и род для существительных, совершенный или несовершенный вид глагола и т.д.

Программа осуществляет получение обратного отображения заданного слова и поиск статьи из соответствующего файла, в которой имеет место совпадение заданного слова с первым словом соответствующей статьи этого словаря по максимальному количеству букв, начиная с трех. Если необходимое трехбуквенное сочетание вообще отсутствует в файле, ищется двухбуквенное сочетание или одна буква. После нахождения подходящей статьи введенное слово, морфологический описатель и адрес падежных окончаний отобранной лексемы передаются в морфологический анализатор. Последний по исходной форме введенного слова (единственное число, именительный падеж – для склоняемых частей речи) выполняет генерацию всей его парадигмы. Для решения этой задачи используется файл окончания лексем. Результат генерации всегда выводится в правое окно.

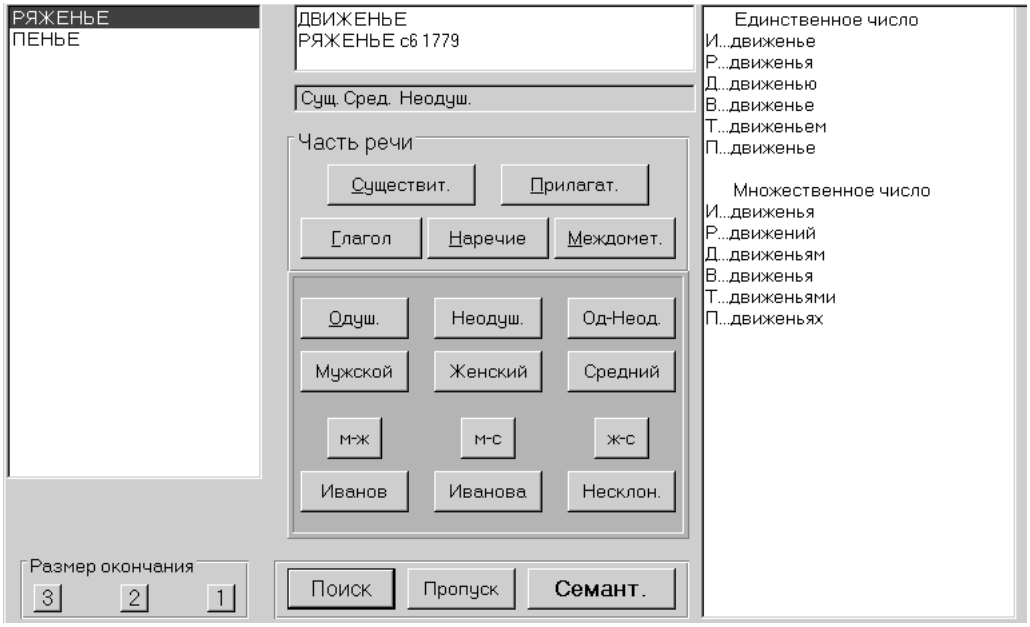

Рис. 1. Подбор морфологического аналога для существительного *движенье*

Для имен существительных (рис. 1) выводится склонение по падежам для единственного и множественного чисел. Здесь следует отметить два обстоятельства. Во-первых, очень многие русские фамилии, особенно такие, которые оканчиваются на -ов, -ев, -ин, склоняются по типу слова ТОПТЫГИН, так что целесообразно для ускорения обработки таких фамилий ввести особую кнопку. Во-вторых, в исходном

словаре Зализняка [3] вообще не оказалось ни одного слова, которое бы склонялось по типу фамилий СИДОРОВА, ИВАНОВА, так что пришлось в основной словарь добавить статью

ИВАНОВА ж11о 27937

и также ввести особую кнопку для обработки подобных фамилий.

Для имен прилагательных выводится склонение по падежам для мужского и женского родов (единственное число) и множественного числа, мужской и женский род для краткой формы и сравнительная форма. Здесь следует отметить важность вывода информации о наличии кратких и сравнительных форм. Так, например, нужно уметь отличать склонение прилагательного *аляповатый* (*аляповат*, *аляповата*) от склонения прилагательного *бывалый* (формы *бывал* и *бывала* являются формами глагола *бывать*, а не краткими формами прилагательного *бывалый*).

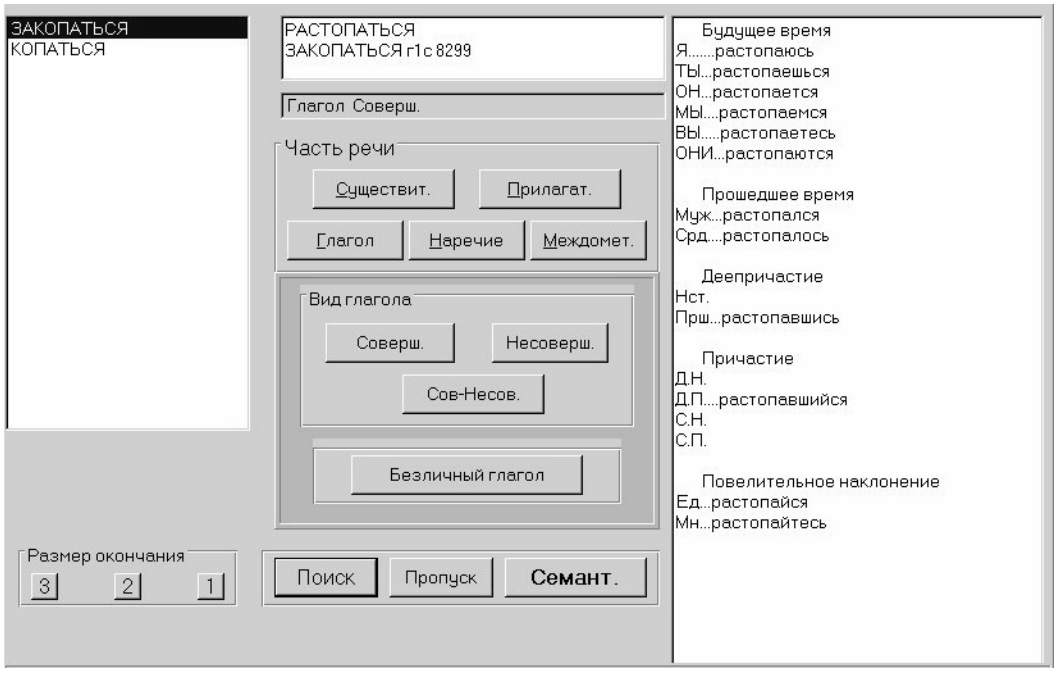

Рис. 2. Подбор морфологического аналога для глагола *растопаться*

Для глаголов (рис. 2) выводится спряжение по лицам для настоящего или будущего времени (в зависимости от вида глагола), мужской и женский род для прошедшего времени, деепричастия настоящего и прошедшего времени, причастия действительного и страдательного залогов и повелительное наклонение для единственного и множественного чисел. Здесь особую важность приобретает вид глагола, а также наличие соответствующих форм причастия, деепричастия и повелительного наклонения. Только учет всех этих параметров позволяет подобрать правильное морфологическое описание лексемы.

Если пользователя не устраивает предлагаемый ему вариант изменения введенного слова, то он может выбрать для образца какое-нибудь другое слово. Набор таких слов предлагается в левом окне. При необходимости возможна процедура отката к совпадению по двум или даже одной букве. После осуществления выбора введенному слову приписываются морфологический описатель и адрес падежных окончаний отобранной лексемы.

В ряде случаев (например, при архаическом или нарочито неправильном написании слова) НС отличается от лексемы, уже имеющейся в словаре одной или двумя формами. Так, например, И.А. Гончаров в романе «Обломов» использует имя *Артемью* вместо *Артемию*, деепричастие *завидя* вместо *завидев* и др. В этих случаях вместо пополнения словаря можно занести подобную словоформу в специальный файл исключений с тем, чтобы перед началом работы морфологического анализатора произвести необходимую замену (*Артемью* на *Артемию).*

Определение семантики НС

После установления морфологических характеристик НС нужно задать его семантику (синтактико-семантические параметры). Для этого, прежде всего, следует установить принадлежность этого слова к определенному классу. Затем необходимо задать возможные связи обрабатываемого НС с другими словами, по возможности описать смысл данного слова с помощью лексических функций и т.п.

Вначале определяется класс НС по классификатору. Используемый **нами** классификатор в настоящее время представляет собой иерархическую структуру из 1600 классов, являющихся основой описания формальной семантики понятий русского языка и отражающих родовидовые отношения между лексемами ([2], с. 101–128). Для облегчения работы предусмотрено несколько видов поиска:

- поиск классов, содержащих имена, отчества или фамилии;
- просмотр классов, содержащих различные названия (географических объектов, фирм, документов и др.);
- поиск классов, содержащих в своих названиях заданное слово;
- поиск класса, содержащего синоним НС.

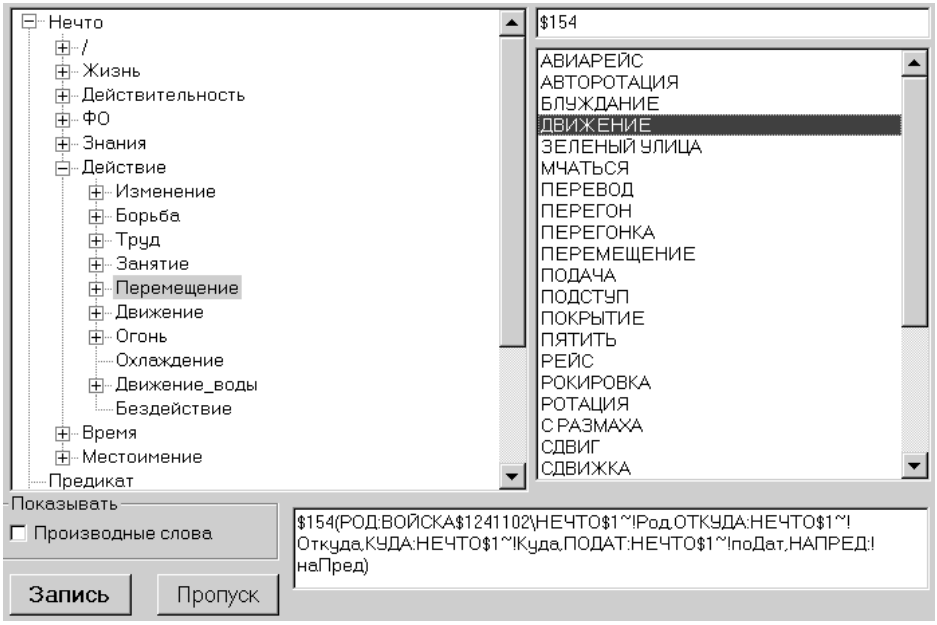

Рис. 3. Подбор семантического аналога

Например, для нового слова *движенье* достаточно в качестве синонима задать слово *движение* и сразу получить, что НС относится к классу $154 «Действие/Перемещение» (рис. 3). При этом показывается дерево классов и полный список слов, принадлежащих данному классу. Из них выбирается наиболее близкое по значению, и его семантика приписывается к НС. В данном случае выбрана лексема, означающая дорожное движение. К ней могут быть подсоединены, например, слова в родительном падеже (движение колонны), откуда (из города), куда (в деревню), в дательном падеже с предлогом «по» (по дороге), и в предложном падеже с предлогом «на» (на машинах).

При необходимости семантика, полученная в нижнем окне, может быть уточнена вручную. Теперь НС с правильной морфологией и семантикой готово для занесения в словарь.

Заключение

Предлагаемая система Adviser апробирована на массивах более 1000 слов различной семантики из произведений И.А. Гончарова и около 5000 фамилий и географических названий и показала прекрасные результаты. Она показала себя достаточно удобной и адекватной задаче определения семантики и морфологии НС. Наибольшие трудности возникли в процессе описания прилагательных, начинающихся с приставки *не-* и образованных от причастий, например, *невысказанный*, *недочитанный*, *незаработанный*. Эти трудности связаны с тем, что причастия являются производной формой от глагола (а в словаре помещены только глаголы). Описание семантики этих прилагательных требует определенного ручного труда, связанного с преобразованием глагол–прилагательное.

Таким образом, описанная система позволяет достаточно быстро и просто пополнять семантический словарь НС, причем работать с ней может даже человек, не знакомый детально с языком описания компьютерного словаря. В отличие от традиционного ручного пополнения словаря, предлагаемые подходы обеспечивают гораздо более высокий уровень точности. Впервые появилась возможность указания точной семантики НС с учетом не только классов, но и аргументов, обеспечивающих связь с подсоединяемыми словами.

Очевидно, что подобные принципы организации системы пополнения семантического словаря с успехом могут быть использованы и при других типах семантических описаний, достаточно только наличия классов или аналогичного принципа построения словаря.

Литература

1. Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы. – М.: Академия, 2006.
2. Тузов В.А. Компьютерная семантика русского языка. – СПб: Изд-во СПбГУ, 2004.

3. Зализняк А.А. Грамматический словарь русского языка. – М.: Русский язык, 1980.
4. Кобзарева Т.Ю. Морфанализ in vivo // Труды Международной конференции Диалог'2004. – М.: Наука, 2004. – С. 286–291.
5. Захаров В.П., Каневский Е.А. Язык И.А. Гончарова через призму современной грамматики // «Прикладна лінгвістика та лінгвістичні технології: MegaLing-2007». – Киев: Довіра, 2008. – С. 131–140.
6. Кожунова О. Опыт применения ДСМ-метода к пополнению семантического словаря // Прикладна лінгвістика та лінгвістичні технології: MegaLing-2006. – Киев: Довіра, 2007. – С. 149–161.
7. Каневский Е.А. Некоторые вопросы пополнения морфологического словаря терминами предметной области // Труды Международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. – М.: РосНИИ искусственного интеллекта, 2001. – Т. 2. – С. 156–160.

Боярский Кирилл Кириллович – Санкт-Петербургский государственный университет информационных технологий, механики и оптики, кандидат физ.-мат. наук, доцент, boyarin9@yandex.ru

Каневский Евгений Александрович – Санкт-Петербургский экономико-математический институт РАН, кандидат технических наук, ведущий научный сотрудник, kanev@emi.nw.ru