Представление устойчивых лексических сочетаний в компьютерном тезаурусе RussNet

Азарова И. В.
azic@bsr.spb.ru

Синопальникова А. А.
anna.sinopalnikova@gmail.com

Смрж П.
smrz@fit.vutbr.cz

В докладе обсуждаются вопросы уточнения структуры компьютерного тезауруса RussNet, ориентированного на «ядерную» структуру современной лексической системы русского языка. При построении wordnet-словарей выявляются частотности значений слов в современных текстах, и определяется набор контекстных маркеров, разграничивающих значения. Описывается методика анализа контекстных маркеров, позволяющая разграничивать свободные и устойчивые словосочетания.

Компьютерный тезаурус RussNet

Компьютерный тезаурус RussNet построен в соответствии с набором основополагающих принципов создания wordnet-словарей[1]. Элементарной единицей тезауруса является набор синонимичных лексем (минимально – одна лексема) знаменательной части речи с определенным значением, так называемый «синсет». Некоторое слово может входить в несколько синсетов в зависимости от числа его значений. В случае лексических лакун допускается включение в синсет устойчивых словосочетаний (*время года, большой палец* и проч.). Элементарные объекты связаны между собой родовидовыми отношениями в набор семантических деревьев (например, *время, совокупность, еда, растения, животные, человек, двигаться, говорить, мыслить* и проч.) Узлы семантических деревьев связаны между собой различными семантическими отношениями, например: часть-целое, антонимия, каузация и проч. Помимо внутренних семантических отношений, имеются внешние, которые связывают элементы структуры национального тезауруса с выделенным набором понятий Межъязыкового лингвистического индекса (ILI).

В методологическом плане стандартная процедура построения wordnet-словарей включает следующие положения.

(1) Словарь опирается на сбалансированный корпус современных текстов; для RussNet он включает тексты 1985-2004 годов общим объемом около 21 млн. словоупотреблений, основу (60%) которого составляют газетные и журнальные статьи на темы повседневной жизни, экономики, политики, науки, культуры и спорта в сочетании с фрагментами литературно-художественных текстов (15%), деловыми текстами и законами (10%), фрагментами научных трудов (15%).

(2) Ядерная структура тезауруса задается примерно двумя тысячами наиболее частотных слов (существительных, глаголов, прилагательных, наречий), которые встречаются более 100 раз на миллион словоупотреблений[2] в корпусе современных текстов.

(3) Разные значения некоторого слова, представленные в тезаурусе, упорядочены в соответствии с частотностью их употребления в корпусе текстов.

(4) В wordnet-словарях представлена, как правило, общая, нетерминологическая лексика, хотя расширения базовой структуры будут включать терминологические элементы, которые тесно связаны с определенными тематическими областями.

При разработке RussNet стандартная методика были расширена следующими положениями.

(1) Элементы синонимического ряда различаются стилистической окраской и частотностью употребления в корпусе. При этом один из синонимов является доминантой синсета – наиболее частотным, нейтральным способом выражения лексического значения в данном языке; остальные элементы ряда значительно уступают доминанте в частотности и закреплены обычно за какой-либо функциональной сферой использования языка.

(2) Основным инструментом при разграничении значений слова в RussNet является контекстный анализ. Анализируя структуру контекста слова в корпусе текстов, мы выделяем статистически значимые маркеры, в качест-

[1] Более подробно методика построения тезауруса описана в статьях (*Азарова и др.* 2003; *Azarova et al.* 2002) и на сайте Санкт-Петербургского университета: http://www.phil.pu.ru/depts/12/RN/

[2] Далее этот показатель обозначается ipm (items per million).

ве которых может выступать и определенная грамматическая форма, и принадлежность к некоторому семантическому дереву родовидовой иерархии RussNet, или оба этих показателя вместе. Эти признаки должны проявляться устойчиво: более чем в 33% контекстов для рассматриваемого значения в корпусе; контекстные маркеры задают рамку валентностей для некоторого значения слова.

(3) Значения слов, частотность появления которых в корпусе составляет менее 1% контекстов для слова, считаются окказиональными (неустойчивыми) и не включаются в тезаурусное описание.

Изучение функционирования слов в реальных текстах является необходимой (основной) составляющей процесса построения RussNet. На различных стадиях нашей работы мы используем контекстные данные для:

- **Установления доминанты** синонимического ряда (на основе частотных показателей). Например, доминантной синонимического ряда *{еда, пища, продукт$_3$, харчи}* является лексема *еда*, так как по сравнению с остальными членами синсета, она обладает наиболее высокой частотой встречаемости – 106,8 ipm (ср.: *пища* – 50,1 ipm, *продукт$_3$* – 24,9 ipm, *харчи* – 6,8 ipm).
- **Различения значений слова** на основе контекстных маркеров. Например, различные грамматические и лексические маркеры сопровождают реализацию в тексте разных значений глагола *мешать*: *мешать$_1$ мне работать* => <кому? N3 *{человек$_1$, лицо$_2$}*> <Inf *{делать$_1$}*>; *мешать$_2$ клей палочкой* => <что? N4 *{вещество$_1$, субстанция$_1$}*> < чем? N5 *{предмет$_1$, вещь$_1$}*>.
- **Установления семантических отношений** между единицами (на основе типичных лексико-синтаксических конструкций). Например, конструкция *X и другие Y* обычно является показателем отношения гипонимии: *После этого процедите через марлю и заправляйте им **винегрет, салат, сельдь** и другие **блюда**.*

Рамки валентностей

Использование данных RussNet для решения прикладных задач, связанных с семантической обработкой текстов (Азарова, Иванов, Секликов 2004), показало, что необходимо включить в структуру RussNet не только результаты контекстного анализа, но и самих контекстных данных. Идея расширения лексикона за счет контекстной информации была сформулирована в нескольких работах по вариантам wordnet-представлений и общим словарям (Stranakova-Lopatkova, Zabokrtsky 2002; Agirre, Martinez 2002; Bentivogli, Pianta 2004.) Форма представления контекстной информации может варьироваться, однако чаще всего она представлена в виде **рамок валентностей**.

В RussNet мы различаем два типа валентностей: активную и пассивную. Для признаковых слов (глаголов, прилагательных и их дериватов) задаются активные валентности, например: *взять* => <что? N4 {*предмет$_1$, вещь$_1$*}>; для зависимых слов (наречий и существительных) указываются пассивные валентности, например: <в лицо> {*лицо$_4$*}> <= {*сказать$_1$*}. Активные валентности наиболее информативны

Активная рамка валентностей некоторого признакового слова задает грамматические и семантические параметры его зависимых слов. Формат активной рамки валентностей состоит из перечня позиций, которые по данным анализа контекстов корпуса регулярно сопутствуют реализации данного значения у слова. Каждая позиция имеет характеристику обязательности/факультативности, обязательная позиция рамки реализуется чрезвычайно регулярно (более чем в 66% случаев), причем в тех случаях, когда в непосредственном контексте (данном предложении) этот элемент отсутствует, он задан для некоторого предшествующего слова, является подразумеваемым из более общего текстового контекста. Факультативная валентность реализуется не столь последовательно (более чем в 33% контекстов), но при разграничении значений данная позиция значима. Элементы контекстного окружения, появляющиеся с меньшей частотой[3], считаются окказиональными (незначимыми).

Валентная позиция характеризуется семантически и грамматически. Одна из семантических характеристик задает функцию данного элемента ситуации. Традиционный способ различения валентностей на основе ролевой семантики не был принят из-за того, что ограниченный набор семантических ролей во многих случаях недостаточен для функциональной характеристики, кроме того, в случаях разграничения объекта и результата, пациенса и адресата приписывание определенной метки носит не всегда объективный характер. Мы исходим из предположения, что **семантические функции валентностей** тесно связаны с семантическими деревьями RussNet. Например, для дерева *двигаться* помимо субъекта движения, важными функциями будут: начальная точка, от которой было начато движения, и конечная точка, к которой направлено движение. Естественно, что эти функции важны для глаголов из соответствующего дерева в разной степени, например, для глагола *двигаться* характерна только функция субъекта, а для глагола *направиться* – субъекта и конечной точки движения. Чем более конкретно значение признакового синсета, тем большее число валентностей возможно. Привязка семантических функций к определенному классу слов сближает наш подход с методикой, принятой в проекте FrameNet[4]. Однако, мы предполагаем, что нет необходимости фиксировать все семантические позиции (они могут уточняться) для дерева и вводить

[3] Пороговые значения для разграничения обязательных, факультативных и окказиональных валентностей будут в дальнейшем уточняться. Возможно, необходимо ввести интервал значений, например: 25-35% и 65-75%. В настоящее время накапливаются данные о распределении значений в данных диапазонах.

[4] Baker, Fillmore, Lowe 1998. См. также FrameNet homepage: http://framenet.icsi.berkeley.edu/~framenet

смысловые обозначения. Например, функция конечной точки движения для глагола *направиться* чаще всего реализуется при помощи указания объекта, к которому или внутрь которого происходит движение (*направиться в дом, к дому, в кусты* и т.д.). Однако, локализация может быть дополнена уточнением лица, по направлению к которому происходит движение (*направиться к нам, в комнату к отцу* и т.д.). В качестве функции будет выступать обобщенное название ***объект***$_2$.

Грамматическая характеристика валентности в рамке указывает на **типовые способы** реализации некоторой валентной позиции. Например, в приведенном выше примере с *направиться* два синтаксических варианта будут включены в рамку: сочетание предлога *в* и винительного падежа имени и предлогом *к* в сочетании с дательным имени. Эти способы характеристики направления движения покрывают 71% контекстов употребления глагола, кроме того, встречаются и другие способы грамматической реализации этой позиции, но они встречаются окказионально (1-5% контекстов), и вряд ли возможно составить исчерпывающий перечень. Регулярные замены грамматических форм, например, изменение формы винительного падежа при переходных глаголах под отрицанием, не вносятся в рамки валентностей, потому что замена зависит от контекста употребления признакового слова (находится на уровне синтаксической реализации) и не выполняется обязательно, поэтому эта трансформационная операция выполняется на основании правил на уровне семантического блока отождествления типов.

Еще одна характеристика задает семантический тип слов, которые могут занимать соответствующую валентную позицию. **Семантическая квалификация валентностей** осуществляется посредством отсылок к **семантическим деревьям тезауруса RussNet**. Например, субъектная позиция глагола *направиться* отсылает к дереву "человек". В других случаях, семантическая квалификация может указывать на часть семантического дерева, например для прилагательного *большой* в значении 'обладающий высокой интенсивностью признака' (*большой друг, большой дурак, большой демократ*) семантической отсылкой будет часть дерева "человек", указывающая на именование человека через квалифицирующий признак, при этом другая часть этого дерева: именования людей по возрасту (*дедушка, ребенок*), полу (*юноше, женщина*) и проч. не будут семантическими маркерами данного значения, ср. *большой дядя, большой юноша, большой мальчик*. Семантическая отсылка может указывать на конкретный синсет (синсеты) в RussNet структуре (например, для валентности глагола *кукарекать*). Семантические деревья RussNet определенным образом коррелируют с традиционным способом описания способов наполнения валентностей, например помете "одушевленные" соответствуют деревья с вершинами *человек* и *животные*, "предметы" – *естественный объект, артефакт, еда* и т.д. Наборы деревьев также используются для семантической квалификации.

Выделение устойчивых словосочетаний

В связи с автоматической обработкой большого количества текстовых данных особое значение приобретает вопрос о формальных признаках, позволяющих разграничивать свободные и устойчивые сочетания лексических единиц. На данный момент практическое решение этого вопроса остается одной из актуальнейших проблем современной компьютерной лингвистики (Sag et al. 2002; Calzolari et al. 2002.).

В wordnet-словарях нет единой схемы включения устойчивых словосочетаний. В принстонском WordNet (Fellbaum 1998; Fellbaum (ed.) 1998), наряду с фразеологическими единицами, в синсеты регулярно включаются свободные словосочетания, назначением которых является облегчение использования словаря человеком-пользователем. Такой подход применительно к русскому языку дал бы огромное число текстовых форм. В GermaNet[5] вставлялись многословные определения для более четкой фасетной организации согипонимов, причем эти синсеты помечаются как искусственные единицы, чтобы они не смешивались с реальными языковыми формами выражения значения.

В RussNet проблема представления устойчивых словосочетаний решается с привлечением статистических методов обработки текстовых данных, поскольку традиционные лексикографические источники, так же как и ассоциативные словари, не предоставляют последовательной информации. Общая стратегия состоит в том, чтобы устойчивые словосочетания включались в синсеты, поскольку они эквивалентны по смыслу отдельным лексемам, а свободные словосочетания были представлены в форме валентностных рамок.

Хотя в толковых словарях обычно во внимание принимается несколько критериев (лексическая ограниченность, воспроизводимость некоторой конструкции в неизменном виде и др.), граница между свободными и устойчивыми (фразеологизированными) словосочетаниями устанавливается довольно субъективно. Критерии разграничения свободных и устойчивых сочетаний регулярно нарушаются. Например, фактор лексической ограниченности зачастую вступает в противоречие с регулярными примерами «вариантов» фразеологизмов (таких как: *бросаться, кидаться, лезть в глаза*) заполнения как активной, так и пассивной позиции.

Ассоциативный словарь (Караулов и др. 2002) также регулярно приводит элементы устойчивых словосочетаний среди наиболее устойчивых (частых) реакций (например, реакция *в долгу* для *оставаться* с частотой 19), что дополнительно указывает на то, что такая информация активно используется в языковой модели человека.

[5] Kunze C., Naumann K. GermaNet homepage. http://www.sfs.uni-tuebingen.de/lsd

Однако, установить четкую зависимость между частотой реакции и характеристикой устойчивости сочетания не представляется возможным, что не позволяет нам использовать данные ассоциативного словаря непосредственно.

Таким образом, в работе над RussNet мы вынуждены опираться прежде всего на данные, полученные при статистической обработке корпуса текстов. Для проведения контекстного анализа нами используется корпус-менеджер Бонито[6], встроенные функции которого позволяют нам вычислять различные показатели взаимной встречаемости единиц в текстах. Эффективным способом выявления компонентов устойчивых сочетаний является блок статистик, в котором для указанного диапазона контекстов («окна»), задаваемого количеством слов, вычисляются параметры, абсолютной *freq(x,y)* и относительной *freq(x,y)/N* частоты сочетания слов, T-коэффициент и коэффициент MI (mutual information, коэффициент взаимной зависимости):

$$T(x,y) = \frac{freq(x,y) - \frac{freq(x) \bullet freq(y)}{N}}{\sqrt{freq(x,y)}}$$

$$MI\ (x,y) = \log_2 \frac{freq\ (x,y) \cdot N}{freq\ (x) \cdot freq\ (y)}$$

Наиболее информативным является последний показатель. В отечественной традиции ему соответствует коэффициент «неслучайности», который широко использовал Н.Д. Андреевым (Андреев) для автоматического выделения морфем и морфоподобных сегментов. Он предполагал, что у коэффициента есть интервал значений, которые помогают выявить осмысленные языковые единицы. Значения интервала необходимо подбирать, поскольку на разных уровнях языковой структуры соотношения частот имеют собственную структуру.

Ниже в Таблице 1 приводятся значения статистических показателей, полученные с помощью Бонито, для контекстов глагола *набрать* в окне, состоящем из 1, 2 и 5 слов.

word	MI-score	T-score	Rel. f [%]	Abs. f
номер	10.38/ 10.66/ 10.86	7.478/ 8.241/ 8.827	4.505/ 5.471/ 6.275	56/ 68/ 78
в	1.47/ 1.955/ 2.718	4.041/ 5.553/ 8.265	0.0094/ 0.0131/ 0.0222	40/ 56/ 95
две	6.832/ 6.832/ 6.925	3.839/ 3.839/ 3.967	0.3851/ 0.3851/ 0.4108	15/ 15/ 16
воды	6.465/ 6.687/ 7.339	3.425/ 3.705/ 4.661	0.2985/ 0.3483/ 0.5473	12/ 14/ 22
скорость	8.615/ 9.156/ 9.243	3.308/ 3.993/ 4.116	1.325/ 1.928/ 2.048	11/ 16/ 17
полную	8.959/ 8.959/ 9.111	2.994/ 2.994/ 3.157	1.682/ 1.682/ 1.869	9/ 9/ 10
силу	7.269/ 7.421/ 7.559	2.981/ 3.144/ 3.299	0.5214/ 0.5794/ 0.6373	9/ 10/ 11
побольше	9.255/ 9.255/ 9.618	2.641/ 2.641/ 2.996	2.065/ 2.065/ 2.655	7/ 7/ 9
телефон	8.156/ 8.349/ 8.349	2.636/ 2.82/ 2.82	0.9642/ 1.102/ 1.102	7/ 8/ 8
на	1.285/ 1.639/ 2.67	2.502/ 3.256/ 5.778	0.0083/ 0.0105/ 0.02151	18/ 23/ 47
высоту	9.421/ 9.643/ 9.643	2.446/ 2.642/ 2.642	2.317/ 2.703/ 2.703	6/ 7/ 7
воздуха	7.526/ 7.749/ 9.189	2.436/ 2.633/ 4.351	0.6231/ 0.7269/ 1.973	6/ 7/ 19

Таблица 1. Статистические характеристики контекстов слова *набрать* (размер окна -1,+1; -2,+2; -5,+5 слов)

Как видно из таблицы, высокие показатели MI и T выделяют как слова, являющиеся контекстными маркерами (*воды, воздуха, скорость, высоту, номер, телефон*), так и слова, не ассоциирующиеся со значениями глагола (*две, побольше, на*). Динамика изменения показателей при расширении окна указывает, что быстро меняющиеся значения свидетельствуют о случайном попадании в окно анализа частотных слов, стабильный прирост значений – о контекстных маркерах со слабо фиксированной позицией, практически стабильные значения отмечают устойчивые позиционно-связанные элементы контекста. Последняя характеристика связана еще с одной особенностью устойчивых словосочетаний – они приближаются в структурном плане к составным лексемам, тяготея к непроницаемости, т. е. к контактному расположению компонентов. Действительно, расположение компонентов устойчивых оборотов *большой палец, иметь в виду* показывают практически абсолютную контактность, которую в первом случае может разрывать только ряд (*большой и указательный пальцы*), а во втором – наречия (*имея также в виду*). Более того, если устойчивое словосочетание разрывается, меняет значение составного целого (*большой волосатый палец*).

Проблемой является то, что как в отношении лексической избирательности, так и контактности наблюдается не простая дихотомия (контактно-дисконтактно), а целая шкала возможных соотношений компонентов словосочетания в контексте.

Например, контексты глагола *набрать* в значении 'вздохнуть' составляют только 3% от общего числа контекстов в нашем корпусе. Контекстным маркером данного значения является генитив *воздуха*, который может быть расположен как контактно, так и после указания части тела человека (*в грудь, в легкие, полную грудь, полные легкие*). Если первый маркер является абсолютно обязательным, т.е. реализуется в 100% контекстов, то второй – в 88%. Глагол *вздохнуть* в синонимичном значении имеет несколько иную структуру контекстных маркеров: упо-

[6] Rychly, Smrz 2004. См. также Bonito homepage: http://nlp.muni.cz/projects/bonito

минание воздуха не встречается, зато регулярно появляется наречие, характеризующее силу вздоха (*глубоко, широко, коротко*), или указание части тела (*полной грудью*). Регулярно наречные выражения при глаголе *вздохнуть* указывают на эмоциональное состояние человека и являются контекстными маркерами для наиболее частотного значения глагола: 'сделать вздох и выдох, выражая некоторое чувство' (*вздохнуть облегченно, с облегчением, разочарованно, устало, страдальчески*). Таким образом, именно сочетание *набрать воздуха* входит в синсет *вздохнуть*$_3$.

В некоторых случаях сочетание слов может использоваться и как устойчивое, и как свободное. Например, *большой дом* используется как устойчивое словосочетание в качестве топонима *Большой Дом* (при этом компоненты словосочетания всегда расположены контактно) или как свободное сочетание, но тогда компоненты могут разрываться (*…жил он размеренно и однообразно, недалеко от института, в большом вычурном доме эпохи так называемых архитектурных излишеств*).

Дальнейшие направления работы

Одной из наиболее сложных надстроек Bonito является так называемый Skecht Engine – программа для вычисления «словесных описаний» (Word Sketches) на основе морфологически размеченного корпуса и заранее заданных типов грамматических конструкций (например, V + N3). В случае с RussNet эти грамматические конструкции соответствуют грамматическим характеристикам валентностных рамок. Таким образом, используя наш корпус, мы можем автоматически извлекать лексические наполнения всех реализованных в тексте валентностных рамок для интересующих нас слов, проводить их статистический анализ, на основе полученных данных оценивать значимость (обязательность/факультативность) валентностей, сравнивать описания для различных лексических единиц, и устанавливать семантические отношения между ними.

Литература

Азарова и др. Компьютерный тезаурус русского языка типа WordNet // Труды Международной конференции Диалог-2003. М., 2003. С. 43–50

Azarova et al. RussNet: Building a Lexical Database for the Russian Language. In: Proceedings: of the Workshop on WordNet Structures and Standardisation, and how these affect Wordnet Application and Evaluation. Las Palmas, Spain, 2002. pp. 60–64.

Stranakova-Lopatkova M., Zabokrtsky Z. Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In: Proceedings of LREC-2002. Las Palmas, Spain, 2002. pp. 949-956.

Agirre E., Martinez D. Integrating Selectional Preferences in wordnet. In: Proceedings of the GWC-2002. Mysore, India, 2002.

Bentivogli L., Pianta E. Extending WordNet with Syntagmatic Information. In: Proceeding of the 2nd Global WordNet Conference. Brno, Czech Republic, 2004. pp. 47-53.

Baker C.F., Fillmore C. J., Lowe J. B. The Berkeley FrameNet Project. In: Proceedings of the COLING/ACL 1998. Montreal, 1998. pp. 86–90.

Rychly P., Smrz P. Manatee, Bonito and Word Sketches. In: Proceedings of the 2nd International Conference on Corpus Linguistics (Corpora-2004). St.Petersburg, 2004. pp. 116-121.

Sag I., Baldwin T., Bond F., Copestake A., Flickinger D. Multiword Expressions: A Pain in the Neck of NLP. In: Proceedings of the CICLING 2002. Mexico City, Mexico, 2002.

Calzolari N., Fillmore C., Grishman R., Ide N., Lenci A., McLeod C., Zampolli A. Towards Best Practice for Multiword Expressions in Computational Lexicons. In: Proceedings of LREC 2002. Las Palmas, Spain, 2002.

Fellbaum C. Towards a Representation of Idioms in WordNet. In: Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems, COLING/ACL. Montreal, 1998. pp. 52-57.

Fellbaum C. (ed.) WordNet: WordNet: An electronic lexical database. Cambridge, Mass.: MIT Press, 1998.

Караулов Ю.Н., Черкасова Г.А., Уфимцева Н.В., Сорокин Ю.А., Тарасов Е.Ф. Русский ассоциативный словарь. Т. 1. От стимула к реакции. М., 2002.