Применение концептуальных сетей для выявления структуры семантической парадигмы прилагательных

Concept lattice implementation in semantic structuring of adjectives

Потемкин С. Б. (potemkin@philol.msu.ru)

МГУ им. М. В. Ломоносова, Россия

Рассматриваются методы формального анализа понятий (FCA) в применении к построению онтологических отношений в классе прилагательных русского языка, характеризующих внешность человека с привлечением WordNet. Выполнен анализ их семантической парадигмы на основе формального контекста, построенного с применением двуязычного словаря.

1. Введение

В настоящее время ведутся активные работы по созданию компьютерного тезауруса русского языка [1, 7, 8], аналогичного по своим структурным и функциональным возможностям распространенному тезаурусу WordNet [16]. Словари такого типа дают широкие возможности для отображения семантических отношений между значениями, получившими лексикализованное выражение в рамках некоторого языка. К сожалению, покрытие лексики подобными тезаурусами для языков, отличных от английского, остается ограниченным, несмотря на значительные усилия по расширению набора синсетов (основная семантическая единица WordNet; набор английских слов, которые в совокупности кодируют некоторое семантическое значение) и их взаимосвязей. Отсюда возникает необходимость автоматизированного получения лексико-семантических отношений из существующих источников, таких как тестовые корпуса или толковые словари. Для решения этой задачи привлекаются, в частности, методы формального анализа понятий (FCA) [11, 13, 14].

Нами разрабатываются методы использования двуязычных (англо-русских) словарей в качестве источника формального контекста и дальнейшего построения концептуальной сети для представления онтологических отношений в классе русских прилагательных.

2. Лексические источники

При определении структуры семантической парадигмы определенной группы слов необходимо опираться на возможно более полные лексические источники. Нами использованы:

- Общие и специальные англо-русские словари, лексическая база данных (ЛБД) [5].

В состав ЛБД включены англо-русские эквиваленты более чем из 30 общих и специальных словарей, включая 3-х томный словарь под редакцией Апресяна, словарь Мюллера, электронные словари Лингво, Полиглоссум, Промт и многие другие. Переводные словари подвергаются своего рода естественному отбору, поскольку они ежедневно используются переводчиками для практических целей, плохой словарь будет отвергнут и не выдержит переиздания;

- Оценка внешности человека (словарь) [2];
- WordNet [16];
- Толковые словари Ожегова, Евгеньевой, частотный словарь Шарова [9];

В настоящем докладе рассматривается семантическая парадигма на материале прилагательных, характеризующих внешность человека. Частотность рассматриваемой лексики весьма значительна: *большой* — 1631 ipm (число словоупотреблений на миллион), *хороший* — 854 ipm, *старый* — 528 ipm, *белый* — 493 ipm, [9] и т. д. Эта группа

выбрана также ввиду ее важности для уточнения системных отношений русской оценочной лексики, представлений о типах лексических значений, особенностей реализации коннотации, нормативных лексических ассоциациях[3], понимания структуры художественного произведения [6]. Важна она и для лингводидактики, как основа для создания различных пособий по развитию речи, обучения русскому языку как русских, так и иностранцев, а также для перевода художественной, юридической, психологической и др. литературы.

Представление значений имен прилагательных строится аналогично представлению других частей речи. Применяется компонентный анализ имен прилагательных с привлечением толковых словарей; корпусные исследования привлекаются для анализа сочетаемости в синтагмах типа прилагательное — существительное, что позволяет кластеризовать прилагательные как атрибуты, приписанные определенным существительным, для которых уже построена некоторая классификация[12]. Используются методы непосредственного полевого тестирования для выявления коннотаций, т. е. сужения круга возможных синтагматических партнеров (прилагательных) данной лексемы (существительного) [4]. Системные отношения в лексике описываются в тезаурусах, где прилагательному приписывается лексическое значение, часто в одном гнезде со сходным по семантике глаголом или существительным.

Представляется многообещающим для выявления семантики прилагательного использовать двуязычные словари и уже построенный иноязычный тезаурус типа тезауруса Роже или получившего в последнее время широкое признание WordNet. Отношения синонимии и антонимии в классе прилагательных достаточно хорошо разработаны, однако и в этой области привлечение двуязычных словарей существенно обогащает списки синонимов и особенно антонимов [5], ручной подбор которых требует значительных усилий. Другие типы отношений: гипонимия, меронимия, метонимия и пр. исследованы значительно меньше. Выявление указанных отношений между прилагательными представляет теоретический и практический интерес, особенно в применении к автоматической обработке текста. В данном случае непосредственная опора на структуру WordNet малопродуктивна. Достаточно сказать, что семантическая организация качественных прилагательных в WordNet полностью отличается от семантической организации существительных или глаголов. Прилагательные организованы в кластеры, каждый из которых привязан к «фокальному» прилагательному, имеющему антоним (к которому привязан свой кластер). Т.е. антонимия оказывается базовым семантическим отношением для кодировки значения прилагательных. Подобный подход объясняется тем фактом, что прилагательные выполняет атрибутивную функцию и что значительное число атрибутов являются биполярными. В классе прилагательных WordNet не построены иерархические отношения, подобные отношению гипонимии для существительных или тропонимии для глаголов и, как правило, не указывается прямой гипероним, вместо него дается ссылка «Pertains to noun …», то есть в классе прилагательных гиперонимом часто является имя существительное, например для прилагательных, обозначающих величину (*большой, малый, узкий, просторный*) родовым гиперонимом является существительное «размер». В настоящем исследовании, однако, мы будем стремиться отыскивать иерархические и др. связи, не выходя за рамки класса прилагательных.

3. Формальный анализ понятий (FCA, Formal Concept Analysis)

Формальный анализ понятий основан на интуитивном представлении о том, что понятие или концепт имеет две стороны: *экстент*, который содержит некоторые объекты, и *интент*, в который входят все атрибуты, свойственные этим объектам [16]. Для проведения формального анализа понятий необходимо, прежде всего, определить *формальный контекст* К:= (G,M,I), где G = множество объектов; M = множество атрибутов; и I = бинарное отношение между элементами G и M, показывающее, какие атрибуты m приписаны каждому из объектов g. Формальный контекст легко представить в виде таблицы. В Таблице 1 в качестве объектов фигурируют некоторые прилагательные русского языка, в качестве атрибутов — набор переводов этих прилагательных; определенное русское слово, напр. *алчный* имеет переводной эквивалент rapacious, на пересечении соответствующей строки и столбца поставлен крест.

Для определения *концепта* в формальном контексте вводится операция деривации → :

$X \subseteq G$: $X \rightarrow X'$: $\{m \in M \mid gIm$ для всех $g \in X\}$

$Y \subseteq M$: $Y \rightarrow Y'$: $\{g \in G \mid gIm$ для всех $m \in X\}$

В нашем примере пусть *X := {ХИЩНЫЙ, прожорливый}* и пусть *Y:={ravening, wolfish},* тогда *X' = {ravening, rapacious, ravenous}, Y' = {ХИЩНЫЙ, жадный}*, далее *X"= {ХИЩНЫЙ, жадный, прожорливый}*, и т. д. Можно показать, что в общем случае $X \subseteq X''$ и $X' = X'''$ а также $Y \subseteq Y''$ и $Y' = Y'''$

Формальным концептом в рамках данного формального контекста является пара (A, B), где A = B', B=A', т. е. A = множество объектов, каждый из которых имеет все атрибуты из множества B; B = множество атрибутов, каждый из которых приписан всем объектам множества A. Все формальные концепты для заданного формального контекста можно сформировать как (X", X') или (Y', Y"), перебирая все подмножества

$X \subseteq G$ или $Y \subseteq M$. Существуют алгоритмы для быстрого построения множества формальных концептов [15].

В нашей таблице выделены ячейки, представляющие формальный концепт (A, B); A= {*алчный, грабительский*}; B= {*rapacious, ravenous*} Формальные концепты данного контекста образуют частично упорядоченное множество $B(K)$, задаваемое отношением $\leq : (A_1, B_1) \leq . (A_2, B_2). \leftrightarrow A_1 \subseteq A_2 \ (\leftrightarrow B_2 \subseteq B_1)$. Это отношение называется отношением *субконцепта-суперконцепта* и $\leq$ определяет полную решетку $\underline{B}(K)$ на $B(K)$, которую можно изобразить в виде помеченного направленного графа (рис. 1). Вершинами графа являются формальные концепты, а ветви отражают отношение субконцепта-суперконцепта.

Для применения методов FCA к выявлению семантической парадигмы прилагательных русского языка предлагается использовать тщательно разработанный семантический тезаурус английского языка WordNet. Основной семантической единицей этого тезауруса является *синсет* — набор английских слов, которые в совокупности кодируют некоторое семантическое значение. Элементом синсета является WM — значение, выраженное отдельным словом (словосочетанием), входящим в синсет. Отдельное слово может входить в различные синсеты, что отражает полисемию, а также и омонимию, присущую данному слову. Между синсетами установлены отношения гипо-гиперонимии (для существительных), тропонимии (для глаголов), антонимии, меронимии и пр. Синсеты, содержащие прилагательные, как правило, не охвачены отношениями гипонимии, установление иерархических отношений между прилагательными затруднительно с теоретической и практической точек зрения [1,12]. Тем не менее, использование синсетов для выявления структуры семантической парадигмы прилагательных представляется возможным и многообещающим. Отметим, прежде всего, что двуязычный англо-русский словарь может эффективно применяться для расширения списка синонимов, а также определения семантической близости между синонимами русского языка [5]. Можно предположить, что взяв набор английских слов, входящих в синсет, $\{e_i\}$, т. е. синонимов с определенным значением, и выписав все их переводы на русский язык $L_j(e_i) = r_{ij}$, в пересечении $\cap_{ij} r_{ij}$ получим набор русских слов, кодирующих значение, эквивалентное значению синсета $\{e_i\}$. Вследствие различного членения действительности в английском и русском языке, которое является прямым отражением несовпадения способов категоризации и, следовательно, концептуализации атрибутивов, а также склонности англичан к большей детализации картины мира и номинации различных признаков, такое пересечение, как правило, оказывается пустым, либо содержит одно-два слова с очень широкой семантикой. Предлагается поэтому воспользоваться средствами FCA, которые позволят выявить всю структуру множеств $\{r_i\}_j$ в их взаимосвязи с синсетом $\{e_i\}$. Формальным контекстом K:= (G,M,I) в данном случае состоит из: множества объектов $G = \cup_j \{r_i\}_j$; всех переводов всех английских слов, входящих в синсет; множества атрибутов $M = \{e_i\}$; бинарного отношения I, определенного функцией L, сопоставляющей каждому английскому слову e_i его j-й русский эквивалент (Табл. 1).

4. Экспериментальные результаты и их интерпретация

В качестве массива данных для проведения экспериментальной апробации методики выбран Словарь «Оценка внешности человека» [2], (далее — Словарь), содержащий более 200 доминант и более 1200 членов синонимических рядов прилагательных, относящихся к внешности человека. В Словаре не выделены качественные и относительные прилагательные, грань между которыми во многих случаях весьма условна. В частности, к описанию лица относятся 603 прилагательных, для которых по изложенной ниже методике были построены 1040 концептуальных сетей с числом атрибутов более 2.

Для каждого прилагательного ar_i из Словаря отыскиваются все английские эквиваленты $ae_{ij}=L_j(ar_i)$, содержащиеся в лексической базе данных (ЛБД). Для каждого ae_{ij} определяется набор синсетов $\{s_k\} = WN(ae_{ij})$ содержащих ae_{ij}. Для каждого из синсетов s_k полученного набора выписываются все русские прилагательные, являющиеся переводными эквивалентами элементов данного синсета; дубли исключаются. Таким образом, получен набор объектов G и набор атрибутов M формального контекста K. На этом этапе мы не выполняем семантического разделения противоречивых переводных эквивалентов (которые в действительности встречаются, напр. *large-handed* переводится как *жадный* и как *расточительный*). Также не отбираются только прилагательные, относящиеся к внешности человека, такой отбор выполняется позже, на этапе анализа построенной концептуальной сети. Все возможные пары эквивалентов включаются в таблицу 1.

Табл. 1 Формальный контекст для синсета. Объекты, входящие в Словарь, выделены заглавными буквами

	edacious	esurient	ravening	rapacious	ravenous	voracious	wolfish
ЗВЕРИНЫЙ							×
ЗВЕРСКИЙ							×
СВИРЕПЫЙ							×
ХИЩНЫЙ			×	×	×		×
алчный				×	×		
грабительский				×	×		

	edacious	esurient	ravening	rapacious	ravenous	voracious	wolfish
волчий					×		×
голодный		×			×		
голодный_как_волк					×		
жадный	×	×	×	×	×	×	×
жаждущий					×		
захватнический				×			
изголодавшийся					×		
ненасытный		×		×	×	×	
относящийся_к_волкам							×
очень_голодный					×		
падкий						×	
похожий_на_волка							×
прожорливый	×	×	×	×	×	×	
свинский				×			
характерный_для_волка							×
эгоистичный				×			

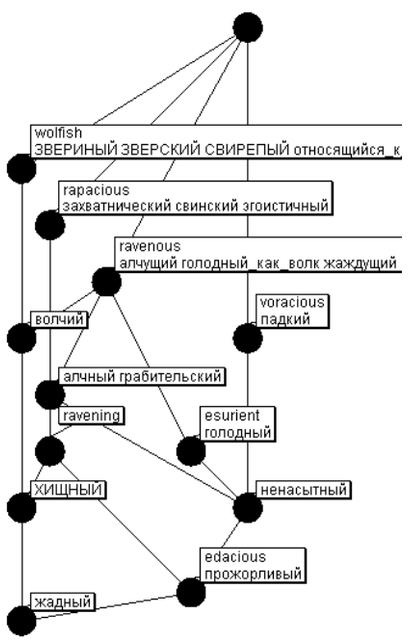

Рис. 1. КС для синсета №00011320 {edacious, esurient, rapacious, ravening, ravenous, voracious, wolfish} к которой относятся объекты {ЗВЕРИНЫЙ, ЗВЕРСКИЙ, СВИРЕПЫЙ, ХИЩНЫЙ}

На рисунке 1 показана концептуальная сеть для формального контекста таблицы 1. В рамках синсета №00011320 объект ХИЩНЫЙ выявляется как гипероним для объектов ЗВЕРИНЫЙ, ЗВЕРСКИЙ, СВИРЕПЫЙ. Такое определение гиперонима в общем случае не представляется корректным (зверь не обязательно хищник, см. Ефремова: *зверь1 = Дикое, обычно хищное животное*), но в качестве характеристики лица звериное, зверское, свирепое лицо скорее всего есть лицо хищное. Из рассмотрения других синсетов выявлены следующие отношения гипонимии:

мертвый ⊆ неподвижный ⊆ вялый
апатичный, оцепенелый ⊆ вялый
изящный ⊆ тонкий
коварный ⊆ хитрый
нахальный, самоуверенный ⊆ дерзкий ⊆ смелый
решительный ⊆ твердый
ястребиный ⊆ хищный
мерзкий, отвратительный, противный, ужасный ⊆ неприятный

Некоторые из этих отношений совпадают с приведенными в Словаре (*изящный ⊆ тонкий, коварный ⊆ хитрый),* остальные выявлены вновь, либо противоречат Словарю, напр. в качестве гипероним к *ястребиный* в Словаре указано прилагательное *беличий* (?).

Кроме выявления отношений гипонимии из концептуальных сетей можно извлечь прилагательные, которые могли бы войти в Словарь: *бесчувственное, будничное, выцветшее, загадочное, заспанное, зловещее, искаженное, легкомысленное, матовое, незамысловатое, нездоровое, неприметное, плоское, подозрительное, полное, полусонное, придурковатое, притворное, разбойничье, смущенное, сухощавое, флегматичное, худое, худощавое,…*

Также выявлены словосочетания, выполняющие атрибутивную роль, которые вообще не включены в словники Словаря: *с буйной растительностью, наводящее скуку, с хитрецой, с хитринкой…* Сопоставление всех полученных иерархических отношений со Словарем не является задачей данного исследования. Предложенный метод позволил выявить дополнительные лексические единицы и установить семантические связи, которые могут использоваться как в лексикографии, так и для решения задач АОТ.

5. Выводы и перспективы исследования

Сложность задачи выявления семантической структуры класса прилагательных подтверждена проведенными ранее исследованиями. Применение методов формального анализа понятий (FCA) для ее решения может оказаться полезным в качестве дополнения к методам корпусных исследований, компонентного анализа и др. Предполагается развить описанные методы с целью формализации выделения иерархических отношений в построенной концептуальной сети. Кроме того, возможно распространение приведенного подхода на иные семантические отношения.

Литература

1. *Азарова И. В., Синопальникова А. А., Яворская М. В.* Принципы построения wordnet-тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2004 М.: 2004. С. 542–547.
2. *Богуславский В. М.* Оценка внешности человека, словарь изд-во «Аст» М. 2004г. 255 стр.
3. *Кедрова Г. Е., Потемкин С. Б.* Семантическое разделение омонимов с использованием двуязычного словаря и словаря синонимов // Труды II Международного конгресса «Русский язык: исторические судьбы и современность», М. 2004 г.
4. *Кобозева И. М.* Лингвистическая семантика изд-во «Эдиториал УРСС», М. 2000г. 350 стр.
5. *Потемкин С. Б.* Лексическая база данных с наложенной семантической метрикой // Труды II Международного конгресса «Русский язык: исторические судьбы и современность», М. 2004 г.
6. *Потемкин С. Б.* Обнаружение события путем анализа антонимов в текстах Н. В. Гоголя и А. П. Чехова, // Слово и словарь — Труды Международной научной конференции «Современные проблемы лексикографии», Гродно 2009 С. 93–95
7. *Портал УИС «Россия»* http://www.cir.ru/.
8. *Сухоногов А. М. Яблонский С. А.* Автоматизация построения англо-русского WordNet. //Труды RDCL 2004. 29 сентября — 1 октября. Пущино, 2004 г.
9. *Шаров С. А.* Частотный словарь http://www.artint.ru/projects/frqlist.asp
10. *Яворская М. В., Азарова И. В.* Структура атрибутивных значений в тезаурусе RussNet (на материале перцептивных прилагательных) // Труды Международной конференции Диалог'2009 С. 542–547.
11. *Cimiano P., Hotho A., Staab S.* Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. // Journal of Artificial Intelligence Research. Volume 24, August 2005 P. 305–339
12. *Mendes S.* Adjectives in WordNet.PT // GWC 2006, Proceedings, P. 225–230.
13. *Priss U.* Linguistic Applications of Formal Concept Analysis. // Ganter; Stumme; Wille (eds.), Formal Concept Analysis, Foundations and Applications. Springer Verlag. LNAI 3626, 2005, P. 149–160.
14. *Stepanova N. A.* Automatic acquisition of lexico-semantic knowledge from corpora. // SENSE'09 Workshop pp. P. 91–100, 2009
15. *Wille R.* Restructuring lattice theory: an approach based on hierarchies of concepts. // Rival, I. (ed.) Ordered Sets. 445–470. Dordrecht-Boston, Reidel, 1982.
16. *WordNet:* An Electronic Lexical Database // Fellbaum Ch. (ed.). MIT Press. 1998.