

УДК 681.3 : 519.76

П. Ф. Павлов, Ю. Ю. Черепанова, И. Ю. Шубин, канд. техн. наук.

О ВОЗМОЖНОСТИ ПОСТРОЕНИЯ ТЕЗАУРУСА СЕМАНТИЧЕСКИХ ПОЛЕЙ И ЕГО ПРИМЕНЕНИЯ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

Зроблено спробу формалізувати критерії побудови деяких парадигматичних полів. Досліджуються питання автоматизованої побудови тезауруса семантичних полів та пропонується його структура. Проаналізована можливість використання тезауруса у інформаційних системах при опрацюванні природно-мовних текстів.

Важным направлением исследований в лингвистике является попытка построения стройной системы языка. Один из методов построения такой системы – выявление семантических полей слов. При этом все множество слов U разбивается на множество семантических полей $U_1, U_2, \dots, U_N$. Семантическое поле U_i – некоторое множество слов $\{v_{1i}, v_{2i}, \dots, v_{mi}\}$, связанных между собой какими-либо семантическими отношениями. Обычно это парадигматические и синтагматические отношения, следовательно, разделяются парадигматические и синтагматические семантические поля. К парадигматическим полям относятся различные классы лексических единиц, имеющие общие признаки в своих значениях. То есть если каждому слову v_i поставить в соответствие множество $W_i = \{a_{i1}, a_{i2}, \dots, a_{ik}\}$ (семантический объем), где a_{iL} – семантический признак слова v_i , то слова v_i, v_j объединяются в одно парадигматическое поле U_z при условии, что $W_i \cap W_j \neq \emptyset$. Обычно в парадигматические поля включают синонимы, антонимы, тематические ряды, лексико-семантические группы слов, совокупности связанных между собой значений полисемантического слова (семантемы) и так далее [1].

Дадим формальное определение некоторых видов парадигматических полей и отношений. Пусть слову v_i соответствует множество семантических признаков $W_i = \{w_{1i}, w_{2i}, \dots, w_{mi}\}$, слову v_j множество $W_j = \{w_{1j}, w_{2j}, \dots, w_{kj}\}$. Тогда для определения вхождения слов v_i, v_j в поле U необходимо проанализировать следующее множество:

$$W = W_i \cap W_j = \{w_1, w_2, \dots, w_z\}.$$

Лексико-семантическая группа определяется как группа слов, в семантических объемах которых совпадает хотя бы один семантический множитель [1], тогда:

$$W \neq \emptyset \Rightarrow v_i, v_j \in U.$$

При построении тематического ряда фиксируются определенные семантические признаки (путем выбора одного конкретного слова и его семантического объема или отдельных семантических множителей). Значит:

$$W_T = \{w_{T1}, w_{T2}, \dots, w_{TL}\}, L \geq 1, W_T \subseteq W_i \Rightarrow v_i \in U.$$

В синонимические ряды Ю. Д. Апресян [2] предлагает включать слова, семантические объемы которых совпадают более чем на 50%. Следовательно, формальное условие вхождения слов v_i, v_j в один синонимический ряд будет выглядеть так:

$$Z \geq \frac{1}{2} \min(m, k).$$

При абсолютных синонимах семантические объемы должны быть одинаковы:

$$W_i = W_j.$$

Под гипонимией понимается отношение, «когда некий класс сущностей включает в себя другой класс сущностей» [3, стр. 40]. Чаще такие отношения называют родовидовыми. Необходимым (но не всегда достаточным) условием того, что v_i и v_j находятся в родовидовых отношениях (v_i является видом v_j), является следующее:

$$W_i \subseteq W_j.$$

Это объясняется тем, что значение родового понятия (более центрального в поле), не может быть уже значения видового понятия [4], следовательно, его семантический объем будет меньше и включается в семантический объем видового понятия.

Синтагматическими называются отношения между сочетаемыми единицами языка. Таким образом, синтагматические семантические поля слова представляют собой связи слов либо их абстрактных категорий в текстах. То есть это либо группы типа «субъект + атрибут», «субъект + предикат + объект», либо типа «существительное + прилагательное», «глагол + наречие» или «белокурый + волосы» [1, стр. 112]. Объединение парадигматического и синтагматического полей дает комплексное семантическое поле. Примером такого поля может служить парадигматическое поле, в котором для каждого значения полисемантического слова приведена его синтагматическая сочетаемость.

Словари терминологии с отражением связей между терминами называются тезаурусами. Данные словари имеют не только общетеоретический интерес, но и практический. Нахождение методов автоматизированного построения тезауруса семантических полей позволит провести объективный анализ обширных лексических пластов и эффективно использовать такой словарь в информационных системах при обработке естественно-языковых текстов.

Для возможности такого использования тезаурус, кроме информации, входящей в обычные терминологические и толковые словари, должен включать дополнительную информацию. Исходя из целей построения тезауруса, имеющего более универсальное применение, разработана следующая структура его организации.

Во-первых, для каждого значения слова предполагается составить и явно указать семантический объем – набор семантических множителей, входящих в данное значение. Наиболее перспективным представляется нахождение компонентов значения по естественно-языковым текстам определений значений слов [5-7]. Семантическими множителями могут быть фрагменты текста дефиниций с номинативным значением и релятивным значением (словосочетания, или просто слова) [5], либо сокращенные основы и квазиосновы слов, входящих в определение [6,7]. Нахождение семантических составляющих значения слова предполагается полностью автоматизировать с помощью метода компонентного анализа [8]. При необходимости может быть получена различная глубина анализа: если на первом шаге на множители раскладывается определение значения слова, то на следующих анализируются семантические составляющие определений семантических множителей, полученных прежде.

Во-вторых, для каждого значения слова предполагается найти связи в текстах выбранной предметной области. При этом для каждой связи будут отражаться тип отношения, категория связанного слова и сами связанные слова. Таким образом, из тезауруса может быть получена синтагматическая сочетаемость различного уровня абстрактности (по типу отношения, по категории связанных слов). Для нахождения синтагматических связей предлагается использовать дистрибутивный и статистический анализы текстов выбранной тематики. Так как не всегда можно автоматически определить используемое в тексте значение полисемантического слова, то для этого необходимо применить различные формальные и неформальные методы. Одним из возможных методов определения актуализированного значения полисемантического слова, который в принципе можно автоматизировать является анализ семантических объемов значений слов контекста. Характерным для правильных значений слов,

дающих оптимальное осмысление предложения, является наличие ряда общих семантических множителей [2]. Это значит, что используемое значение полисемантического слова определяется по максимальной повторяемости семантических множителей в пределах предложения.

Парадигматическое поле слова предполагается в явном виде в тезаурусе не сохранять, а при необходимости получать на основе сравнения семантических объемов по наличию у слов общих множителей. Поле слова может быть получено как по всему его семантическому объему, так и по отдельным семантическим множителям, и по их совокупностям. Для повышения значимости семантической связи в поле предлагается использовать критерий частотности семантического множителя. Так как более значимые множители будут иметь более низкую частоту появления, чем множители широкой семантики, следовательно, может быть установлена наибольшая частота множителя, используемого для включения слова в поле.

Тезаурус такого вида может использоваться для различных целей. Возможность эффективного построения семантических полей может дать богатый материал для лингвистов, исследующих семантическую структуру языка. В системах обработки естественно-языковых текстов проблема разрешения многозначности слова может решаться как методом, описанным выше (с помощью семантических объемов слов), так и с помощью комплексного семантического поля – то есть по связям слова в предложении определяется актуализированное значение слова. Кроме этого, тезаурус может быть применен для решения проблемы обнаружения противоречивости тех предложений, слова которых будут содержать несовместимые семантические множители.

Еще одним из возможных применений тезауруса семантических полей является его использование в качестве компонента интеллектуальных систем. Например, в автоматизированных обучающих системах тезаурус может использоваться при тестировании обучаемого для построения концептуальной модели ответа на естественном языке. Результат тестирования при этом определяется на основе сравнения полученной модели с концептуальными моделями предметной области или эталона ответа. Эталонная модель также может быть получена по неформализованному естественно-языковому тексту всего курса либо по тексту правильного ответа на конкретный вопрос. Концептуальная модель может строиться различными способами. Самым простым путем является нахождение для каждого значащего слова в тексте

множества соответствующих ему семантических множителей и построение семантического объема всего анализируемого текста согласно схеме на рисунке 1.

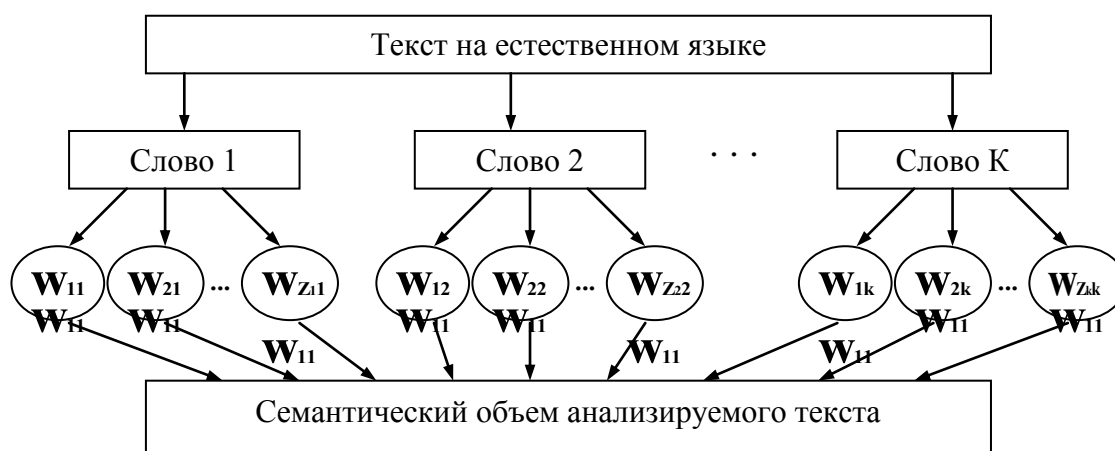


Рисунок 1 – Схема построения семантического объема текста

При этом происходит либо просто объединение всех семантических множителей слов текста, либо действуют некоторые ограничения на их включение в общий семантический объем. При использовании такой модели для получения результата тестирования необходимо проанализировать пересечение семантических объемов ответа обучаемого и эталона ответа. Данный подход является модификацией метода анализа по ключевым словам, но предоставляет некоторые новые возможности, например, позволяет учесть синонимичность естественного языка, сохраняя простоту и эффективность обработки. Повысить качество анализа может разработка и применение результативных механизмов построения семантических объемов и их сравнения. Однако данная модель не позволяет учитывать связи между словами в тексте, поэтому упрощенно представляет его смысл. Поэтому вторая модель является расширением первой и строится с учетом синтагматических связей слов в тексте. В результате получается некоторая сеть, в узлах которой отражаются семантические объемы слов, а дуги показывают их реализованные синтагматические отношения в тексте. Данная модель более точно отражает семантику текста, но потребует больших затрат времени на построение и анализ результата тестирования.

Аналогично тезаурус семантических полей может быть использован в документографических информационно-поисковых системах. При этом концептуальная модель первого вида может быть использована для предварительной

проверки релевантности документа запросу, а модель второго вида – для более точной оценки.

Таким образом, тезаурус семантических полей представляет интерес в связи с возможностью его применения в информационных системах для разрешения некоторых проблем обработки естественно-языковых текстов. Построение тезауруса можно автоматизировать на основе анализа дефиниций слов в терминологических словарях и текстов выбранной тематики. Для более эффективного применения тезауруса необходимо дальнейшее развитие методов выявления комплексных семантических полей.

Литература:

1. Васильев Л. М. Теория семантических полей // Вопросы языкознания. – 1971. - №5. - С. 105-113.
2. Апресян Ю. Д. Избранные труды. Том 1. Лексическая семантика. – М.: Восточная литература, 1995. - 472 с.
3. Шемакин Ю. И. Начала компьютерной лингвистики. – М.: Изд-во МГОУ, 1999.- 114с.
4. Кузнецов А. М. Структурно-семантические параметры в лексике. – М.: Наука, 1980. - 160с.
5. Скороходько Э. Ф. Семантические сети и автоматическая обработка текста. - К.: Наук. думка, 1983. – 284с.
6. Караулов Ю. Н. Частотный словарь семантических множителей русского языка. – М.: Наука, 1980. - 207с.
7. Захаров В. П., Мордовченко П. Г., Сахарный Л. В. Совершенствование лингвистического обеспечения в ИПС «бестезаурусного» типа // НТИ. - Сер. 2. – 1980. - №6. – С.14-17.
8. Павлов П.Ф., Бабина О.И., Черепанова Ю.Ю. О проблеме автоматизированного выявления семантических полей. // Вісник ХДПУ. Збірка наукових праць. – Вип. 42. – Харьков: ХДПУ, 1999. – С. 81-85.