У даній статті наведено порівняльний аналіз різних підходів до формування словників семантичної сполучуваності та представлено підхід до формування словника сполучуваності термінів предметної області. Основою запропонованого підходу є апарат алгебри скінченних предикатів і предикатних операцій. Розробки з даної проблематики можна використати в галузі лексикографії, семантичного аналізу текстової інформації та ін.

Ключові слова: автоматизована обробка природної мови, інтелектуальні системи, алгебра скінченних предикатів, лексикографія

В данной статье приводится сравнительный анализ различных подходов к формированию словарей семантической сочетаемости и представлен подход к формированию словаря сочетаемости терминов предметной области. Основой предложенного подхода является аппарат алгебры конечных предикатов и предикатных операций. Разработки по данной проблематике можно использовать в области лексикографии, семантического анализа текстовой информации и др.

Ключевые слова: автоматизированная обработка естественного языка, интеллектуальные системы, алгебра конечных предикатов, лексикография

УДК 004.934

ФОРМИРОВАНИЕ СЛОВАРЯ СОЧЕТАЕМОСТИ ТЕРМИНОВ ПРЕДМЕТНОЙ ОБЛАСТИ

Н. В. Борисова
Ассистент*
E-mail: borisova_nv@mail.ru
О. В. Канищева
Кандидат технических наук, доцент*
E-mail: olya-kanisheva@rambler.ru
*Кафедра интеллектуальных компьютерных систем
Национальный технический университет "Харьковский политехнический институт"
ул. Фрунзе, 21, г. Харьков, Украина, 61002

1. Введение

Несмотря на развитие технологий представления информации в формальном, доступном для компьютерной обработки виде, основной объем информации порождается, хранится и передается в виде текстов на естественном языке (ЕЯ). В связи с лавинообразным ростом количества информации в самых разных сферах человеческой деятельности возникает острая необходимость автоматического решения различных задач, связанных с обработкой текстов на ЕЯ, в том числе перевода с одного языка на другой, поиска информации в текстовых массивах, извлечения информации из текстов, реферирования и др.

Системы автоматической обработки текстов на ЕЯ (АОТ-системы), использующие методы поверхностного анализа (например, основанные на поиске ключевых слов) для решения перечисленных задач, в большинстве случаев не позволяют достичь результата, качество которого достаточно для практического применения.

Причина кроется в необходимости учета не только слов, составляющих текст на ЕЯ, но и взаимосвязей между ними, не представленных в тексте в явном виде. Для выявления таких взаимосвязей требуется глубокий и полный анализ различных языковых явлений, представленных в тексте, и в первую очередь – выделение синтаксических отношений между словами текста (синтаксический анализ) [1].

Для автоматического выделения синтаксических отношений требуется привлекать различную информацию о сочетаемости слов. Простейшим типом такой информации может служить формальное описание синтаксического поведения различных частей речи.

Таким образом, для качественного решения задачи автоматического синтаксического анализа необходимо подробное и полное описание принципов сочетаемости (морфо-синтаксических, семантических и лексических).

2. Литературный обзор исследований и постановка проблемы

Как показал обзор существующих в свободном доступе лингвистических описаний сочетаемости слов русского языка (словарей сочетаемости, комбинаторных словарей), данные источники информации о сочетаемости обладают существенными недостатками. Во-первых, большинство таких источников рассчитано на пользователя-человека, поэтому зачастую авторы вместо того, чтобы приводить формальное и последовательное описание сочетаемости некоторого слова, ограничиваются рядом примеров и ссылок на аналогичные слова, апеллируя к интуиции пользователя словаря. Во-вторых, доступные словари (в особенности те, которые формализованы в достаточной для практического применения степени) покрывают лишь небольшую часть лексики русского языка. В-третьих, в большинстве словарей сведения о семантических ограничениях на сочетаемость либо не приводятся вовсе, либо не формализованы в достаточной мере. Основной причиной перечисленных недостатков является чрезвычайно высокая трудоемкость ручного

© Н. В. Борисова, О. В. Канищева, 2013

формирования описаний сочетаемости, носящих комбинаторный характер (по сути, требуется описать множество пар, или даже n количество слов, способных образовывать допустимые словосочетания) [2].

Альтернативой использованию лингвистических описаний сочетаемости является автоматический сбор статистики совместной встречаемости слов на большой текстовой коллекции и формирование статистического описания сочетаемости. При этом имеет смысл использовать неразмеченные (т.е. не обработанные экспертами) тексты, поскольку создание достаточной по объему размеченной коллекции является очень сложной и трудоемкой задачей. Такой подход позволяет свести к минимуму объем требуемого ручного труда, а также обеспечить довольно полный охват лексики.

Однако простая статистика совместной встречаемости слов не дает всей необходимой информации о сочетаемости. Это связано с проблемой разреженности данных о совместной встречаемости, извлеченных из коллекции текстов на ЕЯ: лишь небольшая часть сочетающихся между собой слов реально встретятся вместе в коллекции.

Свойство разреженности является фундаментальным для текстов на ЕЯ, поэтому решить данную проблему невозможно ни увеличением объема, ни изменением состава текстовой коллекции.

Таким образом, актуальным является создание методов автоматизированного формирования описаний сочетаемости, позволяющих извлекать информацию о сочетаемости из различных текстовых коллекций, обобщать ее и представлять в таком виде, в котором эксперты могут эффективно работать с ней. Другой актуальной проблемой является учет сформированных таким образом, а также содержащихся в существующих словарях, описаний сочетаемости для улучшения качества и повышения эффективности автоматического синтаксического анализа.

3. Цель и задачи исследования

Авторами предлагается подход к автоматическому построению словаря семантической сочетаемости на основе существующих словарей. Данный подход основан на использовании математического аппарата алгебры конечных предикатов и предикатных операций.

4. Семантическая сочетаемость слов

В процессе автоматического синтаксического анализа текстов на русском языке постоянно возникает задача выбора из нескольких синтаксических структур предложения правильной структуры. Во многих случаях правильный выбор можно сделать только при наличии описаний сочетаемости слов, входящих в анализируемое предложение [3].

Семантические ограничения на сочетаемость указывают, что слово может быть связано синтаксической связью некоторого типа только со словами, относящимися к определенным семантическим классам. Например, в рамках экологической предметной области прямым дополнением при глаголе сбрасывать может быть только слово, обозначающее некоторую жидкость (жидкие отходы, сточные воды, стоки).

При описании в словаре и учете в процессе анализа семантических ограничений возникают следующие сложности.

Во-первых, описание семантических классов простым перечислением входящих в них слов на практике оказывается плохим решением: списки слов получаются огромными и заведомо неполными; нет способа оценить степень принадлежности слова семантическому классу. Во-вторых, попытки автоматического извлечения информации о семантических ограничениях на сочетаемость из корпуса текстов наталкиваются на проблему разреженности данных: если слово w сочетается с любыми словами w` из достаточно крупного семантического класса, то в любом сколь угодно большом корпусе встретится лишь часть возможных словосочетаний w w`. Поэтому после извлечения слов, встретившихся с w в корпусе, необходимо на их основе каким-то образом описать все множество сочетающихся с w слов [4].

5. Моделирование семантической сочетаемости слов предметной области

Пусть M – это множество слов, участвующих в образовании словосочетаний: $M=\{m_1,m_2,\ldots,m_n\}$, где n определяется количеством рассматриваемых словосочетаний. На этом множестве введем систему предикатов S таким образом, чтобы любой предикат $P(t)\in S$ обращался в 1 на множестве слов с какой-то определенной семантической ролью, и был равен 0 в противном случае. Понятие семантической роли было введено в работе [5]. Множество предикатов S представляет множество семантических ролей слов из словаря. Каждому элементу m_i из M соответствует некоторый предикат $P_i(t)\in S$, равный 1 при подстановке множества семантических ролей конкретного слова m_i. Следовательно, каждому $m_i\in M$ взаимно однозначно соответствует определенный одноместный подстановочный предикат, который задает множество семантических ролей.

Рассмотрим два множества слов M_1 и M_2, где M_1 – множество слов, стоящих на первом месте в словосочетании. Операция соединения двух слов из M_1 и M_2, множества семантических ролей которых заданы предикатами $P_1(t_1)\in S_1$ и $P_2(t_2)\in S_2$, характеризуются согласованием определенных семантических ролей этих слов. В результате семантического согласования двух рядом стоящих слов получаем множество связей между семантическими ролями, другими словами, – множество пар семантических ролей. Таким образом, между множествами семантическим ролей рядом стоящих слов существует бинарное отношение, которое является подмножеством декартового произведения этих множеств. Наличие или отсутствие согласования между словами определяется с использованием метода компараторной идентификации [6].

Это бинарное отношение можно представить с помощью некоторого двуместного предиката $P(t_1,t_2)$, при этом

$$P(t_1,t_2)\rightarrow P_1(t_1)\cdot P_2(t_2). \qquad (1)$$

Предположим, что существует возможность согласования семантических ролей не зависит от того, к каким словосочетаниям они относятся. Тогда на декартовом произведении множеств $S_1 \times S_1$ можно задать предикат $\alpha(t_1,t_2)$, принимающий значение 1, если семантические роли t_1 и t_2 можно согласовать, и значение 0 в противном случае. Довольно редко подмножество согласуемых семантических ролей совпадает с декартовым произведением всех возможных связей. Некоторые семантические роли рядом стоящих слов в действительности не вступает в согласование, в связи с этим в формулу (1) вводится дополнительный множитель, который стремится исключить нереализованные связи.

Таким образом, бинарное отношение на множествах рядом стоящих слов может быть задано формулой:

$$P_1(t_1) \otimes P_2(t_2) = \alpha(t_1,t_2) \cdot P_1(t_1) \cdot P_2(t_2), \qquad (2)$$

где $\otimes$ обозначена операция соединений морфемных семантических ролей.

Действительно, логическое произведение предикатов $P_1(t_1) \cdot P_2(t_2)$ описывает все возможные связи между словами, а предикат $\alpha(t_1,t_2)$ исключает часть нереализованных связей [7, 8].

Для этого с помощью словарей [9, 10] можно выделить следующие существительные и прилагательные, которые взаимодействуют друг с другом в научных текстах экологической направленности. Существительные: x_1 – воздействие, x_2 – ландшафт, x_3 – нагрузка, x_4 – рельеф, x_5 – среда, x_6 – явление, x_7 – фактор, x_8 – ареал, x_9 – продуктивность, x_{10} – доза, x_{11} – выброс, x_{12} – концентрация, x_{13} – поступление, x_{14} – сброс, x_{15} – излучение. Прилагательные: y_1 – антропогенный, y_2 – первичный, y_3 – летальный, y_4 – предельно допустимый, y_5 – электромагнитный, y_6 – радиоактивный.

Семантическая роль x_7 является общей для прилагательных y_1, y_2, y_3, а для x_{10} такими прилагательными являются y_3 и y_4.

Таким образом, анализ словаря показал, что представленные выше существительные и прилагательные реализуют следующие композиции семантических ролей: x_1y_1 – воздействие антропогенное, x_2y_1 – ландшафт антропогенный, x_3y_1 – нагрузка антропогенная, x_4y_1 – рельеф антропогенный, x_5y_1 – среда антропогенная, x_6y_1 – явление антропогенное, x_7y_1 – фактор антропогенный, x_7y_2 – фактор первичный, x_7y_3 – фактор летальный, x_8y_2 – ареал первичный, x_9y_2 – продуктивность первичная, $x_{10}y_3$ – доза летальная, $x_{10}y_4$ – доза предельно допустимая, $x_{11}y_4$ – выброс предельно допустимый, $x_{12}y_4$ – концентрация предельно допустимая, $x_{13}y_4$ – поступление предельно допустимое, $x_{14}y_4$ – сброс предельно допустимый, $x_{15}y_5$ – излучение электромагнитное, $x_{15}y_6$ – излучение радиоактивное.

Графическое отображение семантической сочетаемости слов предметной области, приведенных выше, представлено на рис. 1.

Для математического описания связей между семантическими ролями слов воспользуемся формулой (2). Для нашего примера $\alpha(t_1,t_2)$ может быть представлено следующим образом:

$$\begin{aligned}\alpha(t_1,t_2) = {} & t_1^{x_1}t_2^{y_1} \vee t_1^{x_2}t_2^{y_1} \vee t_1^{x_3}t_2^{y_1} \vee t_1^{x_4}t_2^{y_1} \vee \\ & \vee t_1^{x_5}t_2^{y_1} \vee t_1^{x_6}t_2^{y_1} \vee t_1^{x_7}t_2^{y_1} \vee t_1^{x_7}t_2^{y_2} \vee t_1^{x_7}t_2^{y_3} \vee \\ & \vee t_1^{x_8}t_2^{y_2} \vee t_1^{x_9}t_2^{y_2} \vee t_1^{x_{10}}t_2^{y_3} \vee t_1^{x_{10}}t_2^{y_4} \vee t_1^{x_{11}}t_2^{y_4} \vee \\ & \vee t_1^{x_{12}}t_2^{y_4} \vee t_1^{x_{13}}t_2^{y_4} \vee t_1^{x_{14}}t_2^{y_4} \vee t_1^{x_{15}}t_2^{y_5} \vee t_1^{x_{15}}t_2^{y_6}.\end{aligned} \qquad (3)$$

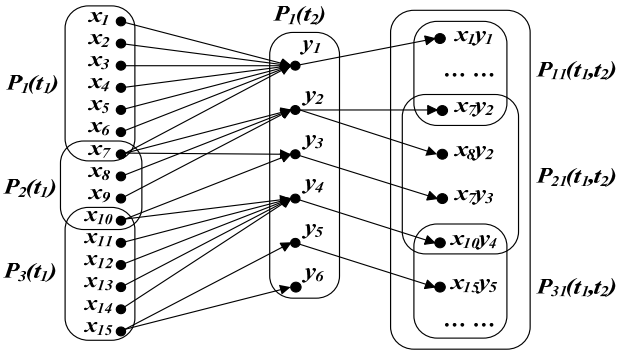

Рис. 1. Семантическая сочетаемость слов предметной области

Множества семантических ролей терминов, стоящих на первом месте в словосочетаниях, задаются предикатами $P_1(t_1)$, $P_2(t_1)$, $P_3(t_1)$, которые могут быть представлены следующим образом:

$$\begin{aligned}P_1(t_1) &= t_1^{x_1} \vee t_1^{x_2} \vee t_1^{x_3} \vee t_1^{x_4} \vee t_1^{x_5} \vee t_1^{x_6} \vee t_1^{x_7} \\ P_2(t_1) &= t_1^{x_7} \vee t_1^{x_8} \vee t_1^{x_9} \vee t_1^{x_{10}} \\ P_3(t_1) &= t_1^{x_{10}} \vee t_1^{x_{11}} \vee t_1^{x_{12}} \vee t_1^{x_{13}} \vee t_1^{x_{14}} \vee t_1^{x_{15}}.\end{aligned} \qquad (4)$$

Множества семантических ролей терминов, стоящих в рассмотренном примере на втором месте, могут выражаться предикатом $P_1(t_2)$:

$$P_1(t_2) = t_2^{y_1} \vee t_2^{y_2} \vee t_2^{y_3} \vee t_2^{y_4} \vee t_2^{y_5} \vee t_2^{y_6}. \qquad (5)$$

В соответствии с формулой (2) опишем множество смысловых значений словосочетаний предметной области, задаваемое с помощью предикатов $P_{11}(t_1,t_2)$, $P_{21}(t_1,t_2)$, $P_{31}(t_1,t_2)$.

$$\begin{aligned}P_{11}(t_1,t_2) = {} & \alpha(t_1,t_2) \cdot P_1(t_1) \cdot P_1(t_2) = (t_1^{x_1}t_2^{y_1} \vee t_1^{x_2}t_2^{y_1} \vee \\ & \vee t_1^{x_3}t_2^{y_1} \vee t_1^{x_4}t_2^{y_1} \vee t_1^{x_5}t_2^{y_1} \vee t_1^{x_6}t_2^{y_1} \vee t_1^{x_7}t_2^{y_1} \vee t_1^{x_7}t_2^{y_2} \vee \\ & \vee t_1^{x_7}t_2^{y_3} \vee t_1^{x_8}t_2^{y_2} \vee t_1^{x_9}t_2^{y_2} \vee t_1^{x_{10}}t_2^{y_3} \vee t_1^{x_{10}}t_2^{y_4} \vee t_1^{x_{11}}t_2^{y_4} \vee \\ & \vee t_1^{x_{12}}t_2^{y_4} \vee t_1^{x_{13}}t_2^{y_4} \vee t_1^{x_{14}}t_2^{y_4} \vee t_1^{x_{15}}t_2^{y_5} \vee t_1^{x_{15}}t_2^{y_6}) \\ & (t_1^{x_1} \vee t_1^{x_2} \vee t_1^{x_3} \vee t_1^{x_4} \vee t_1^{x_5} \vee t_1^{x_6} \vee t_1^{x_7}) \\ & (t_2^{y_1} \vee t_2^{y_2} \vee t_2^{y_3} \vee t_2^{y_4} \vee t_2^{y_5} \vee t_2^{y_6}).\end{aligned} \qquad (6)$$

$$\begin{aligned}P_{21}(t_1,t_2) = {} & \alpha(t_1,t_2) \cdot P_2(t_1) \cdot P_1(t_2) = (t_1^{x_1}t_2^{y_1} \vee \\ & \vee t_1^{x_2}t_2^{y_1} \vee t_1^{x_3}t_2^{y_1} \vee t_1^{x_4}t_2^{y_1} \vee t_1^{x_5}t_2^{y_1} \vee t_1^{x_6}t_2^{y_1} \vee \\ & \vee t_1^{x_7}t_2^{y_1} \vee t_1^{x_7}t_2^{y_2} \vee t_1^{x_7}t_2^{y_3} \vee t_1^{x_8}t_2^{y_2} \vee t_1^{x_9}t_2^{y_2} \vee \\ & \vee t_1^{x_{10}}t_2^{y_3} \vee t_1^{x_{10}}t_2^{y_4} \vee t_1^{x_{11}}t_2^{y_4} \vee t_1^{x_{12}}t_2^{y_4} \vee \\ & \vee t_1^{x_{13}}t_2^{y_4} \vee t_1^{x_{14}}t_2^{y_4} \vee t_1^{x_{15}}t_2^{y_5} \vee t_1^{x_{15}}t_2^{y_6}) \\ & (t_1^{x_7} \vee t_1^{x_8} \vee t_1^{x_9} \vee t_1^{x_{10}}) \\ & (t_2^{y_1} \vee t_2^{y_2} \vee t_2^{y_3} \vee t_2^{y_4} \vee t_2^{y_5} \vee t_2^{y_6}).\end{aligned} \qquad (7)$$

$$P_{31}(t_1,t_2)=\alpha(t_1,t_2)\cdot P_3(t_1)\cdot P_1(t_2)=(t_1^{x_1}t_2^{y_1}\vee$$
$$\vee t_1^{x_2}t_2^{y_1}\vee t_1^{x_3}t_2^{y_1}\vee t_1^{x_4}t_2^{y_1}\vee t_1^{x_5}t_2^{y_1}\vee t_1^{x_6}t_2^{y_1}\vee$$
$$\vee t_1^{x_7}t_2^{y_1}\vee t_1^{x_7}t_2^{y_2}\vee t_1^{x_7}t_2^{y_3}\vee t_1^{x_8}t_2^{y_2}\vee t_1^{x_9}t_2^{y_2}\vee$$
$$\vee t_1^{x_{10}}t_2^{y_3}\vee t_1^{x_{10}}t_2^{y_4}\vee t_1^{x_{11}}t_2^{y_4}\vee t_1^{x_{12}}t_2^{y_4}\vee t_1^{x_{13}}t_2^{y_4}\vee \quad (8)$$
$$\vee t_1^{x_{14}}t_2^{y_4}\vee t_1^{x_{15}}t_2^{y_5}\vee t_1^{x_{15}}t_2^{y_6})$$
$$(t_1^{x_{10}}\vee t_1^{x_{11}}\vee t_1^{x_{12}}\vee t_1^{x_{13}}\vee t_1^{x_{14}}\vee t_1^{x_{15}})$$
$$(t_2^{y_1}\vee t_2^{y_2}\vee t_2^{y_3}\vee t_2^{y_4}\vee t_2^{y_5}\vee t_2^{y_6}).$$

6. Выводы

Полученные математические модели семантической сочетаемости слов в словосочетаниях определенной предметной области наглядно демонстрируют использование математического аппарата алгебры конечных предикатов при решении задач, связанных с естественным языком, а также могут быть использованы в различных лингвистических экспериментах.

Литература

1. Арефьев, Н. В. Методы построения и использования компьютерных словарей сочетаемости для синтаксических анализаторов русскоязычных текстов [Текст] : автореф. дис. ... канд. физико-мат. наук : 05.13.11 / Н. В. Арефьев. – М., 2012. – 22 с.
2. Мальковский, М. Г. Семантические ограничения в словаре сочетаемости: эксперименты по разрешению синтаксической неоднозначности [Электронный ресурс] / М. Г. Мальковский, Н. В. Арефьев. – Режим доступа : http://www.sworld.com.ua/ index.php/uk/ technical-sciences-112/informatics-computer-science-and-automation-112/12730-112-530.
3. Лексическая сочетаемость слов [Электронный ресурс]. – Режим доступа : http://obrazovanie.biniko.com/info_61.php.
4. Автоматическое формирование базы сочетаемости слов на основе очень большого корпуса текстов [Текст] / Э. С. Клышинский, Н. А. Кочеткова, М. И. Литвинов, В. Ю. Максимов // Компьютерная лингвистика и интеллектуальные технологии: материалы международной конференции «Диалог», 26-30 мая 2010 г., Бекасово. – М.: Изд-во РГГУ, 2010. – 9 (16). – С. 181-185.
5. Шаронова, Н. В. Компараторная идентификация лингвистических объектов [Текст] : дис. ... док. тех. наук : 05.25.05 / Н. В. Шаронова. – Харьков, 1994. – 271 с.
6. Бондаренко, М. Ф. Инструментарий компараторной идентификации [Текст] / М. Ф. Бондаренко, Ю. П. Шабанов-Кушнаренко, Н. В. Шаронова // Бионика интеллекта. – 2010. – № 2 (73). – С. 74-86.
7. Шабанов-Кушнаренко, Ю. П. Компараторная идентификация лингвистических объектов [Текст] / Ю.П. Шабанов-Кушнаренко, Н. В. Шаронова. – К.: ИСДО, 1993. – 116 с.
8. Шабанов-Кушнаренко, Ю. П. Теория интеллекта: Проблемы и перспективы [Текст] / Ю. П. Шабанов-Кушнаренко. – Х.: Вища шк., 1987. – 158 с.
9. Реймерс, Н. Ф. Природопользование: Словарь-справочник [Текст] / Н. Ф. Реймерс. – М.: Мысль, 1990. – 637 с.
10. Некос, А. Н. Екологія та неоекологія. Українсько-російський словник-довідник. [Текст] / А. Н. Некос, Н. В. Борисова. – Харків: Вид-во ХНУ ім. В.Н. Каразіна, 2001. – 236 с.