Черепанова Ю.Ю.

Харьковский национальный университет радиоэлектроники, Украина

Использование компонентного анализа для разрешения неоднозначности слов естественного языка

Одной из проблем, усложняющих автоматическую обработку информации на естественном языке, является неоднозначность языковых знаков. Она может проявляться в нескольких формах:

– простые омонимы (слова одной части речи, одинаковые по написанию, но разные по лексическому значению);

– слова, обозначающие разные лексические значения многозначного слова (имеющие в своих значениях что-то общее);

– синтаксические омонимы, совпадающие по написанию в косвенных формах;

– омонимия разных словоформ одного слова.

Омонимия третьего и четвертого типа в системах, имеющих морфологические, синтаксические и семантические анализаторы естественного языка, частично снимается на этапе морфологического анализа. Для дальнейшего разрешения многозначности и снятия омонимии применяются алгоритмы, входящие в блоки синтаксического и семантического анализаторов, которые проверяют семантику связи слов естественного языка (например, задаются синтаксические правила связи слов в определенной форме, или анализируются валентности слов). Описанные методы разрешения неоднозначности слов требуют больших затрат времени при подготовке информационного обеспечения системы.

Кроме того, для решения некоторых задач не требуется использование морфологической и синтаксической информации, значит, трудоемкая разработка анализатора, увеличивающего затраты памяти и времени обработки системы, становится неоправданной.

Разработанная структура тезауруса семантических полей [1] позволяет довольно эффективно применить для разрешения многозначности слов метод компонентного анализа, предложенный Апресяном [2].

Каждому слову приписывается семантический объем – набор семантических множителей, входящих в значение слова. В свою очередь, каждый множитель может иметь свой семантический объем, получаемый на следующем шаге компонентного анализа. Построение семантического объема наиболее эффективно по дефинициям толкового словаря, кодированием значащих слов в семантические множители [3].

Для определения значения многозначного слова, используемого в тексте, полученные семантические объемы слов текста сравниваются и выбирается то значение слова, семантический объем которого дает наибольшее число совпадений семантических множителей. При этом может понадобиться от 1 до 6 шагов компонентного анализа. Для более эффективного и быстрого анализа можно задать глубину анализа неоднозначного слова сразу несколько большей, чем для остальных слов (например, 3), и, при необходимости, наращивать глубину для всей фразы.

Пример работы метода приведен на рисунке 1. В примере показано разрешение неоднозначности слова «месяц» в двух фразах: «Ярко сиял месяц» и «За этот месяц он многое успел». Выбрано два значения слова «месяц»: «Единица исчисления времени по солнечному календарю, срок в 30 суток», «Диск луны или его часть». Для каждого слова найдены определения в словаре Ожегова, Шведовой [5]. Пунктиром выделены семантические поля, полученные на различных шагах компонентного анализа.

Для первой фразы глубина анализа составила 2. Актуализированным оказалось второе значение слова месяц, общий семантический множитель – «свт».

Для второй фразы глубина анализа составила 1.

Актуализированным оказалось первое значение слова месяц, общие семантические множители – «врм», «срк».

Данный метод требует небольших трудозатрат разработчика, но является вероятностным. Например, во фразе «Положение месяца указывало на скорый конец ночи» разрешение неоднозначности слова «месяц» данным методом даст неверный результат. Однако употребление значения слова в его «верном» контексте значительно более частое, чем в «неверном» (как в приведенной фразе). Поэтому, несмотря на свой вероятностный характер, данный метод эффективен хотя бы для предварительного разрешения неоднозначности.

Литература:

1. Черепанова, Ю.Ю. Контроль знаний с ответами на естественном языке [Текст] / Ю.Ю. Черепанова //Восточно-европейский журнал передовых технологий. Информационные технологии. – 4\2(40)2009. – С.32-36.

2. Апресян, Ю. Д. Избранные труды. [Текст]. Т. 1. Лексическая семантика / Ю.Д. Апресян. – М.: Восточная литература, 1995. – 472 с.

3. Черепанова, Ю.Ю. Методы и алгоритмы построения семантического объема слова [Текст] / Ю.Ю. Черепанова //Восточно-европейский журнал передовых технологий. Информационные технологии. – 4\2(34)2008. – С.21-25.

4. Черепанова, Ю.Ю. О теоретико-множественном и теоретико-категорном подходах к моделированию семантических полей [Текст] / Ю.Ю. Черепанова // Проблемы бионики: Всеукр. межвед. науч.-техн. сб. – 2001. – Вып 54. – С.75-78.

5. Ожегов С.И. Толковый словарь русского языка: 80000 слов и фразеологических выражений. [Текст] / С.И.Ожегов, Н. Ю. Шведова. - 3-е изд., стереотипное – М.: АЗЪ, 1996 – 928с.

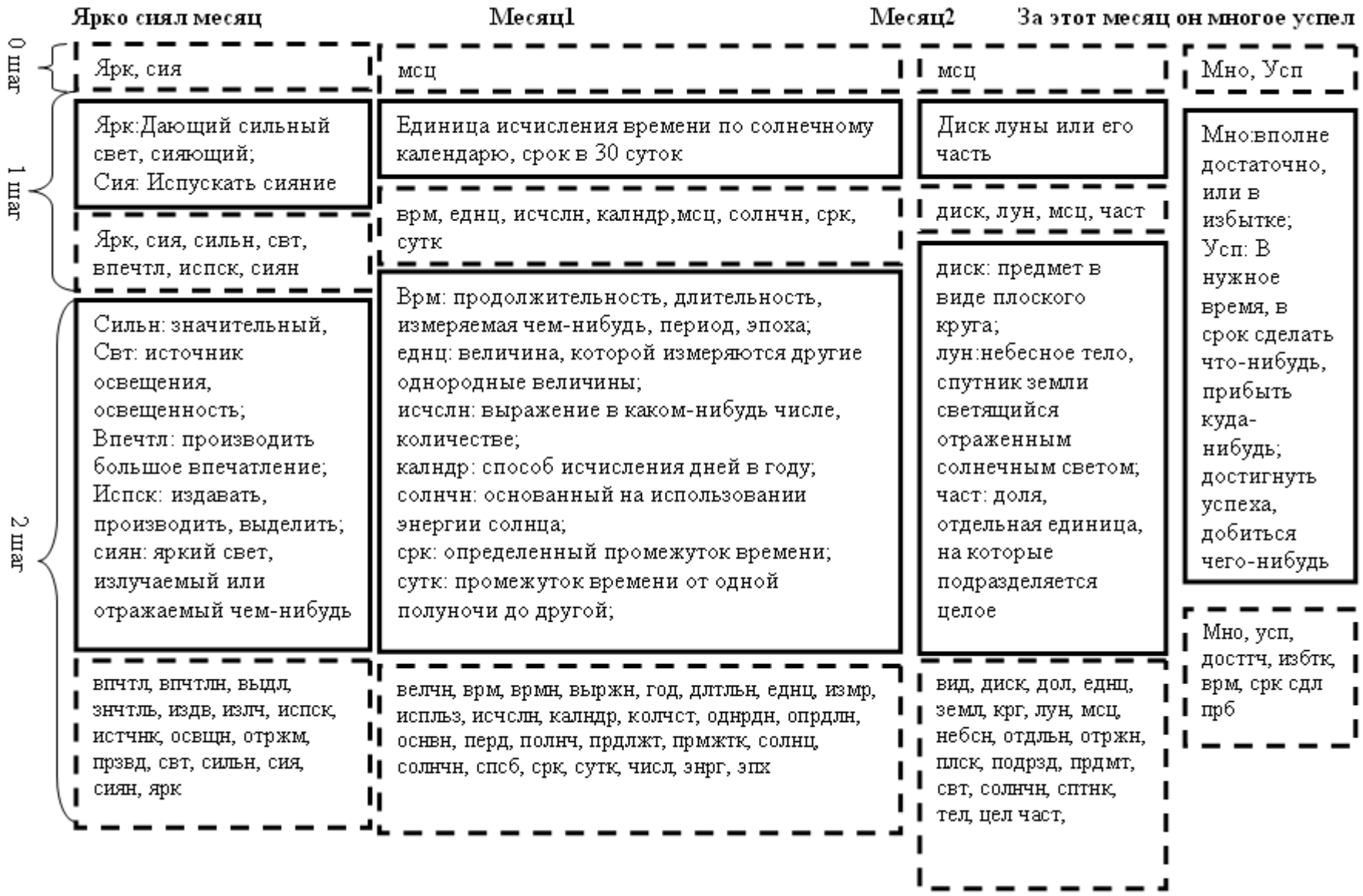

Рис.1 Пример разрешения неоднозначности слова «месяц»