началом использования данных, такого, который бы с заданной вероятностью гарантировал безотказность постоянного поступления пакетов. Данная методика может быть использована в программном обеспечении для проигрывания мультимедиа, такого как просмотр онлайн видео фильмов и музыки, а также онлайн телевидения и радио. Внедрение разработанной технологии не требует особого нового оборудования или переписывания протоколов, а может быть легко добавлено в виде модуля проигрывателя мультимедиа на компьютере клиента.

Литература

1. Таненбаум Э. Компьютерные сети. 5-е издание: пер. с англ. / Э. Таненбаум. – СПб. – ПИТЕР, 2010. – 844с.
2. Столлингс В. Современные компьютерные сети. 2-е изд.: пер. с англ. / В. Столлингс. – СПб.: ПИТЕР, 2003. – 783с.
3. Олифер В. Г. Компьютерные сети. Принципы, технологии, протоколы: Учебник для вузов. 4-е изд. / Олифер В. Г., Олифер Н. А. — СПб.: Питер, 2010. — 944 с.
4. Гмурман В.Е. Теория вероятностей и математическая статистика / В.Е. Гмурман. – М.: Высш. шк., 1972. – 368с.
5. Старков И.И. Статистическая обработка наблюдений / И.И. Старков. – М.: БИНОМ, 2003 – 312с.
6. Крамер Т. Математические методы статистики: пер. с англ. / Крамер Т. – М.: МИР, 1975. – 648с.
7. Хемди А. Таха Введение в исследование операций, 7-е издание.: пер. с англ. / Хемди А. Таха – М.: Издательский дом «Вильямс», 2005. – 912с.
8. Кельберт М. Я. Вероятность и статистика в примерах и задачах. Т. II: Марковские цепи как отправная точка теории случайных процессов и их приложения / М. Я. Кельберт, Ю. М. Сухов – М.: МЦНМО, 2009. — 295 с.
9. Кемени Дж. Конечные цепи Маркова: пер. с англ. / Дж. Кемени, Дж. Снелл – М.: Наука, 1970. – 271с.

УДК 681.3:519.76

РАЗРЕШЕНИЕ НЕОДНОЗНАЧНОСТИ СЛОВ ЕСТЕСТВЕННОГО ЯЗЫКА

Ю.Ю. Черепанова
Ассистент
Кафедра «Программное обеспечение ЭВМ»
Харьковский национальный университет радиоэлектроники
пр. Ленина, 14, г. Харьков, Украина, 61166
Контактный тел.: (057) 402-14-46
E-mail: cher_y@list.ru

Розглянуто проблему вирішення неоднозначності слів природної мови. Запропоновано метод і алгоритм вирішення неоднозначності слів на основі застосування компонентного аналізу із використанням тезаурусу семантичних полів для побудови семантичних обсягів слів

Ключові слова: неоднозначність слів природної мови, компонентний аналіз, тезаурус

Рассмотрена проблема разрешения неоднозначности слов естественного языка. Предложены метод и алгоритм разрешения неоднозначности слов на основе применения компонентного анализа с использованием тезауруса семантических полей для построения семантических объемов слов

Ключевые слова: неоднозначность слов естественного языка, компонентный анализ, тезаурус

The problem of resolution of ambiguity words of natural language is considered. The method and algorithm disambiguate words based on the use of component analysis using a thesaurus of semantic fields for the construction of semantic volumes of words are proposed

Key words: ambiguity words of natural language, component analysis, thesaurus

1. Введение

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.

В таких системах зачастую проблема разрешения многозначности решается следующими способами:

- в небольших системах разрабатываются и используются правила и возможные сценарии использования словоформ различных значений в окружающем контексте. Данный метод, являясь частным случаем метода, описанного выше, повышает трудоемкость наполнения системы.
- полисемия исключается вообще. Это достигается искусственным приписыванием слову одного значения, например подключением и первоочередным использованием специальных отраслевых словарей, ограничивая естественный язык исключением из него слов общей семантики, что ограничивает и возможности таких систем: например, в поисковых системах не все релевантные запросу документы будут выданы.
- неоднозначность игнорируется, все значения слова объединяются. Такой метод также не для всех систем является приемлемым. Например, в поисковых системах ответ на запрос будет содержать большой процент информационного шума.

Цель работы – разработать метод разрешения неоднозначности слов с возможностью наиболее полной автоматизации.

2. Использование компонентного анализа для разрешения неоднозначности слов

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. Ниже приведен алгоритм разрешения неоднозначности слов с использованием тезауруса семантических полей [1].

3. Алгоритм разрешения неоднозначности слов

На входе текст $T=\{v_1,...,v_i\}$, содержащий омонимичное слово $v_j \in T$ ($f_o(v_j)=\{o_{j1},...,o_{jk}\}, k \geq 2$). На выходе m - номер актуализированного значения слова v_j.

Инициализация: $W_{Tкон}=\{\ \}$, $W_{vjz}=f_{код}(v_j), z=\overline{1,k}$. Отображение $f_{код}(v)$ представляет собой кодирование слов в квазиосновы, семантические множители [4].

1-й шаг.

Выделить из множества T множество слов текста контекста омонимичного слова v_j: $T_{кон}=\{v_1,..,v_{j-1},v_{j+1},..,v_i\}$.

2-й шаг.

Кодирование слов $v_l \in T_{кон}$ в семантические множители a_l : $\{a_1,...,a_l\}=f_{код}\{v_1,...,v_{j-1},v_{j+1},...,v_i\}$;

3-й шаг.

Включение полученных множителей в семантический объем текста контекста: $W_{Tкон}=W_{Tкон}\cup\{a_1,...,a_l\}$;

4-й шаг.

Найти множества $W_{Cz}=W_{Tкон}\bigcup W_{vjz}$;

5-й шаг.

Найти m, такое, что $|W_{Cm}|=\max|W_{Cz}|$, для $z=\overline{1,k}$. Если m найдено, выход.

6-й шаг.

Найти множество определений $\{o_1,...,o_i\}$, являющееся сечением отражения $f_{дек1}: A \to O$ по множеству $W_{Tкон}$. $T_{кон}=o_1\cup o_2\cup...\cup o_i$, $W_{vjz}=W_{vjz}\bigcup f_{код}(o_{jz})$. Перейти к шагу 2.

4. Пример работы метода

Пример работы метода приведен на рис. 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».

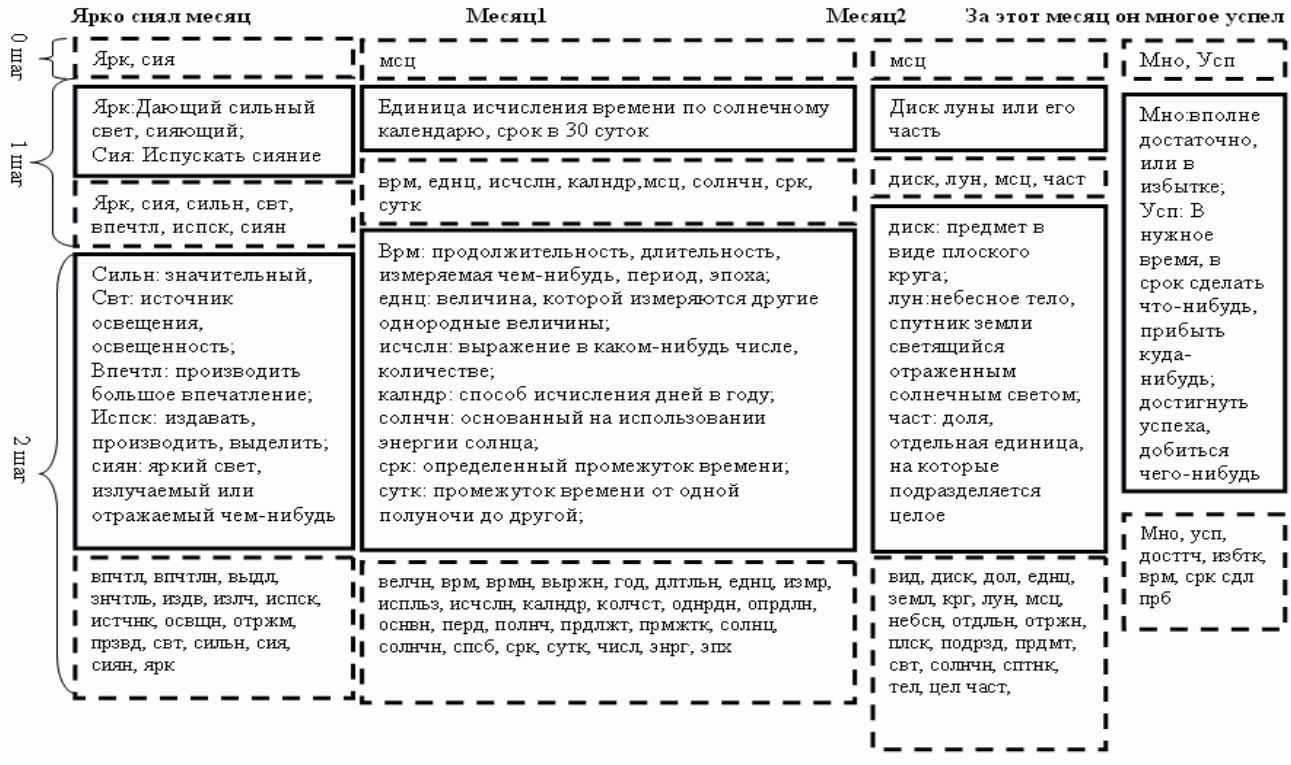

Рис.1. Пример разрешения неоднозначности слова «месяц»

5. Выводы

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.