# ПРИМЕНЕНИЕ КОМПОНЕНТНОГО АНАЛИЗА ДЛЯ РАЗРЕШЕНИЯ НЕОДНОЗНАЧНОСТИ ЕСТЕСТВЕННОГО ЯЗЫКА

Черепанова Ю. Ю.
Научные руководители – Шубин И. Ю., Павлов П. Ф.
Харьковский национальный университет радиоэлектроники

Одной из проблем, усложняющих автоматическую обработку информации на естественном языке, является неоднозначность языковых знаков. Она может проявляться в нескольких формах. Во-первых, это могут быть простые омонимы (слова одной части речи, одинаковые по написанию, но разные по лексическому значению), во-вторых, слова, обозначающие разные лексические значения многозначного слова (имеющие в своих значениях что-то общее), в-третьих, синтаксические омонимы, совпадающие по написанию в косвенных формах, в-четвертых, существует 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 (например, в поисковых системах), значит, трудоемкая разработка анализатора 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 адаптирования системы.

– полисемия исключается вообще. Это достигается искусственным приписыванием слову одного значения, например подключением и первоочередным использованием специальных отраслевых словарей, что не только ограничивает естественный язык, исключая из него слова общей семантики, но 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.

В качестве достаточно эффективного метода разрешения неоднозначности предлагается использовать метод компонентного анализа (КА). Каждому слову приписывается семантический объем (СО) – набор семантических множителей (СМ), входящих в значение слова. В свою очередь, каждый множитель может иметь свой СО, получаемый на следующем шаге КА. Построение СО наиболее эффективно по дефинициям толкового словаря кодированием значащих слов в СМ. В настоящее время известно несколько алгоритмов автоматизированного получения кодов слов (квазиоснов, сверток). Однако для данной специфики алгоритм должен удовлетворять следующим требованиям:

– получение кодов должно быть максимально автоматизировано;

– не должно возникать синонимии кодов, а именно, словоформы одной лексемы и близкородственные слова, фактически обозначающие одно и то же, должны иметь один код;

– исключение омонимичности – слова, имеющие существенные отличия в своих значениях (даже однокоренные), должны иметь разные коды;

– размер кода должен быть минимальным с учетом предыдущих требований, то есть сокращение кода с увеличением его омонимичности недопустимо.

Для определения значения многозначного слова, используемого в тексте, полученные СО слов текста сравниваются и выбирается то значение, СО которого дает наибольшее число совпадений СМ. При этом может понадобиться от 1 до 6 шагов КА. Например, во фразе «На раздумывание ему дали месяц» для слова «месяц» глубина анализа составила три шага, а во фразе «Ярко светил месяц» - два шага. 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.

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.