УДК 004.934

Борисова Н.В, Канищева О.В.

ПОДХОД К ФОРМИРОВАНИЮ СЛОВАРЯ СОЧЕТАЕМОСТИ ТЕРМИНОВ ПРЕДМЕТНОЙ ОБЛАСТИ

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.

Семантические ограничения на сочетаемость указывают, что слово может быть связано синтаксической связью некоторого типа только со словами, относящимися к определенным семантическим классам.

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.

Авторами предлагается подход к автоматическому построению словаря семантической сочетаемости на основе существующих электронных словарей. Данный подход, основанный на использовании математического аппарата алгебры конечных предикатов и предикатных операций, а также метода компараторной идентификации, позволяет получить математические модели семантической сочетаемости слов в словосочетаниях определенной предметной области. Эти модели могут быть использованы лингвистами-аналитиками, лексикографами, а также другими экспертами и специалистами в системах автоматизированной обработки естественного языка.