УДК004.853

Н. И. Гурин, доц., канд. физ.-мат. наук; Я. А. Жук, асп.
(БГТУ, г. Минск)

# МОРФОЛОГИЧЕСКИЙ АНАЛИЗ ТЕКСТА ПРИ ПОМОЩИ РЕГУЛЯРНЫХ ВЫРАЖЕНИЙ И БАЗ ДАННЫХ

Одним из предварительных этапов автоматического семантического анализа текста, требующегося для построения базы знаний диалоговой информационной системы, является морфологический анализ слов, т.е. определение части речи слова и его формы. Данная задача возникает как при разбиении сложных предложений на простые для различения причастных и деепричастных оборотов речи от перечислений, так и при семантическом анализе простых предложений в связи с тем, что элементом перехода, стоящим между двумя смысловыми единицами, с точки зрения актуального членения предложения является глагол.Наличие большинства морфологических признаков зависит от части речи, к которой принадлежит слово. Глаголы обладают спряжением, числом, временем, родом в прошедшем времени и лицом в настоящем и будущем. Вспомогательные части речи, наречия и обозначенные цифрами числительные являются неизменяемыми частями речи. Остальные основные части речи (существительное, прилагательное, местоимение) изменяются по падежам и числу.

На основе существующих методов распознавания части речи слов по словарям и по правилам разработан гибридный метод на основе баз данных. В базах данных хранятся наборы из последних символов слов в различных формах без привязки к морфемам и соответствующие слову морфологические признаки. Данные наборы символов применяются для распознавания части речи и формы слов в порядке убывания длины, что позволяет в случае неверного распознавания внести в соответствующую базу данных формы слова целиком. Учитывая отличия в наборе морфологических признаков у различных частей речи, в базе данных было создано 3 таблицы: для неизменяемых слов, для глаголов и для остальных частей речи.В таблице глаголов хранятся 12 форм слов (инфинитив, 2 числа по 3 лица, 3 рода и множественное число для прошедшего времени, деепричастие), в таблице несклоняемых частей речи −одна форма слова, в таблице основных частей речи − 12 полей (2 числа по 6 падежей).

Сравнение результатов, извлекаемых при помощи запросов на выборку при помощи регулярных выражений из разных баз данных, между собой позволило исправить ошибки в распознавании части речи некоторых слов, выявленные в ходе тестирования.