0.1. *Федотов А.М., Самбетбаева М.А.* **Алгоритм морфологического анализатора для казахского языка**

В данной работе рассматриваются модели и существующие алгоритмы нормализации слов естественных языков. Описаны алгоритмы автоматического выделения основ для ряда естественных языков и возможные пути синтеза нормальной формы слова для казахского языка.

Необходимость приведения слов к нормальной форме (построения морфологического анализатора) возникла при работе с информационно-поисковыми тезаурусами с учетом морфологии казахского языка в полнотекстовых базах данных по информационным технологиям. Приведение слов в анализируемом тексте к нормальной форме сильно упрощает работу с ним: индексацию, последующий поиск информации по построенному индексу, а также решение задач классификации (кластеризации) и автоматического реферирования документов научно-технической тематики [1].

Выбор решения. В данной работе рассмотрен два наиболее популярных алгоритма лемматизации, основывающиеся на различных принципах – это алгоритм Портера [2] и алгоритм Яндекса [3]. Исходя из анализа существующих решений были выработаны модель и правила алгоритма получения нормальной формы слова для казахского языка, соединяющая по описаниям приведенные выше алгоритмы.

Морфологическая модель казахского языка. В казахском языке словоформы образуются путем конкатенации корня и аффиксов (суффиксов и окончаний). В казахском языке окончания делятся на четыре вида окончаний. Описанные внизу окончания непосредственно будут использоваться в разрабатываемом алгоритме определения основы слова.

Обозначим через P_i — следующие множества окончаний (аффиксов), для $i = 1, 2, 3, 4$.

P_1 — множество трехбуквенных окончаний (окончание множественного числа);

P_2 — множество окончаний (притяжательные окончания);

P_3 — множество окончаний (личные окончания);

P_4 — множество окончаний (падежные окончания).

Для удобства реализации было исследована систематизация окончаний и порядок правил.

Пусть:

Q — произвольный одноместный предикат;

W — множество нормальной формы слова;

Каждое слово z представим в виде $z = y \wedge x$ как конкатенация двух (или более) слов y и x;

Если слово $x \in P_i$, то обозначим как $P_i(x)$ для всех $i = 1, ..., 4$;

Если слово $x \in W$, то обозначим как $W(x)$;

Если слово $x \in Q$, то обозначим как $Q(x)$.

Исследованы существующие алгоритмы выделения основ и приведение в нормальную форму слов для казахского языка. Разработан алгоритм для морфологического анализатора казахского языка и создан прототип системы нормализации, подтверждающий работоспособность данного алгорима. Отличительными особенностями построеннго алгоритма является его понятность и достаточно легкая воспроизводимость, что позволяет, в частности, без особых трудозатрат применить его в семантических поисковых системах.

Список литературы

[1] Шокин Ю.И., Федотов А.М., Барахнин В.Б. Проблемы поиска информации. / Новосибирск: Наука, 2010. — 196 с.

[2] Porter M.F. An algorithm for suffix stripping // Program. — 1980. — V. 14, No 3, P. 130—137

[3] Segalovich I.V. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. 2003, p. 219-223