OSTIS-2012
(Open Semantic Technologies for Intelligent Systems)

УДК 004.432.4

ИНТЕЛЛЕКТУАЛЬНЫЙ МОРФОЛОГИЧЕСКИЙ АНАЛИЗАТОР, ОСНОВАННЫЙ НА СЕМАНТИЧЕСКИХ СЕТЯХ

Шарипбаев А.А., Бекманова Г.Т., Ергеш Б.Ж., Бурибаева А. К., Карабалаева М.Х.

Евразийский национальный университет имени Л. Н. Гумилева,
г. Астана, Казахстан

sharalt@mail.ru
gulmira-r@yandex.ru
banu_kazakh@yahoo.com
buribayeva@mail.ru
mkarabal@mail.ru

Разработан интеллектуальный морфологический анализатор казахского языка на основе формализации морфологических правил с помощью семантических сетей.

Ключевые слова: морфологические правила, морфологический анализатор, семантические признаки, семантическая сеть, синхронизированное линейное дерево.

Введение

Очевидная сложность обработки естественно-языковых процессов вызвана трудностью их формализации. Сложность заключается в невозможности словоизменения слов для какой либо части речи по заданной траектории без предварительной обработки словаря начальных форм, поскольку существует зависимость словоизменения слова от его смысла, то есть от его семантического содержания.

1. Построение семантической базы начальных форм слов

В казахском языке словоформы образуются путем конкатенации корня и аффиксов (суффиксов и окончаний). При этом каждый аффикс связан с наборами семантических признаков и порядок добавления аффиксов строго определен. Например, для имен существительных к основе слова вначале добавляется суффикс и далее окончание множественного числа, затем притяжательное окончание, далее следует падежное окончание и последним окончание формы спряжения (добавляется только к одушевленным существительным) [Казахская грамматика, 2002].

Для формализации правил добавления суффиксов и окончаний используются семантические сети, в которых вершины представляют морфологические единицы, а дуги задают отношения между ними. Для разработки морфологического анализатора используется база данных начальных форм слов, на которой будет формироваться словарь словоформ казахского языка со всеми семантическими признаками.

Новые словоформы образуются с учетом морфологических и семантических признаков начальных форм следующим образом: **с**начала к начальной форме слова добавляются суффиксы. Затем, двигаясь слева направо, определяется категория (глухие, звонкие и т.п.) последней буквы (последнего звука) начальной формы слова для добавления того или иного окончания [Бекманова, 2009].

В качестве семантических признаков начальных форм слов выступают такие категории как часть речи, одушевленность и неодушевленность для имен существительных, образование сравнительных и превосходных степеней прилагательных, образование собирательных и порядковых числительных, для глагола сочетание в сложных формах с вспомогательными глаголами как «отыр», «тұр», «жатыр», «жүр» и др. Всего в базе знаний, по которой осуществляется словоизменение, более 100 семантических признаков. База данных начальных слов с семантическими признаками представлена на рисунке 1.

word	имя существительное	прилагательное	числительное	глагол	местоимение
анализ	1	0	0	0	0
аналитик	1	0	0	0	0
аналитикалық	0	1	0	0	0
аналық	0	1	0	0	0
анар	1	0	0	0	0
анархизм	1	0	0	0	0
анархист	1	0	0	0	0
анау	0	0	0	0	1
анықтама	1	0	0	0	0
аңғал	0	1	0	0	0
ангар	1	0	0	0	0
аңғар	1	0	0	0	0
аңғарғыш	0	1	0	0	0
аңғарлы	0	1	0	0	0
аңғару	0	0	0	1	0
аңғарымпаздық	1	0	0	0	0

Рисунок 1 – Вид база данных начальных форм слов казахского языка

2. Формализация морфологических правил казахского языка.

Для формализации правил добавления суффиксов и окончаний предлагается использовать семантическую нейронную сеть, предложенную в [Шуклин, 2001]. С помощью такой сети генерируются словоформы казахского языка и порождается структура словаря начальных форм в виде синхронизированного линейного дерева.

Для представления словоформы и ее признаков будут использоваться следующие метасимволы:

- разделитель между словами,
(- начало слова,
) - конец слова,
! - начало признака словоформы (падеж и т. д.),
* -конец признака словоформы.

Рассмотрим пример для слова «бала - ребенок» (основа слова) и двух его словоформ «балам - мой ребенок», «балаң - твой ребенок» (в казахском языке одушевленные существительные изменяются по лицам с помощью личных окончаний). Рецептор возбуждается на символ начала слова «(». Далее переходит в состояние «б», при подаче символа «б», далее последовательно «(ба», «(бал», «(бала» , и затем одновременно два субсостояния «(балам)» и «(балаң)» (рисунок 2).

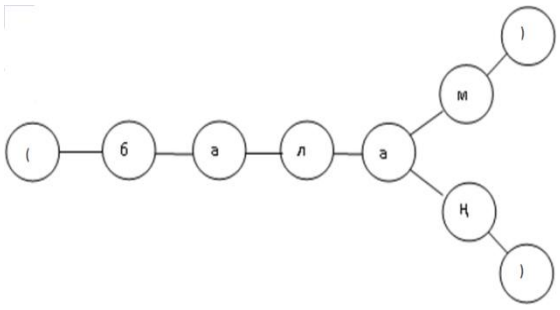

Рисунок 2 – Синхронизированное линейное дерево для словоформ

На рисунке 3 приведен пример структуры связей леммы, определяющей следующие признаки: имя существительное (зат есім) –«!зе*» , одушевленное – «!жа*», притяжательное окончание (тәуелдік жалғау) первого лица – «!11*» (бірінші жақ), притяжательное окончание (тәуелдік жалғау) второго лица – «!22*» (екінші жақ). При подаче на лемму слова «(балам)» она переходит в возбужденные субсостояния: «(балам)», «!зе*», «!жа*», «!11*» а при подаче слова «балаң» в возбужденные субсостояния: «(балаң)», «!зе*», «!жа*», «!22*».

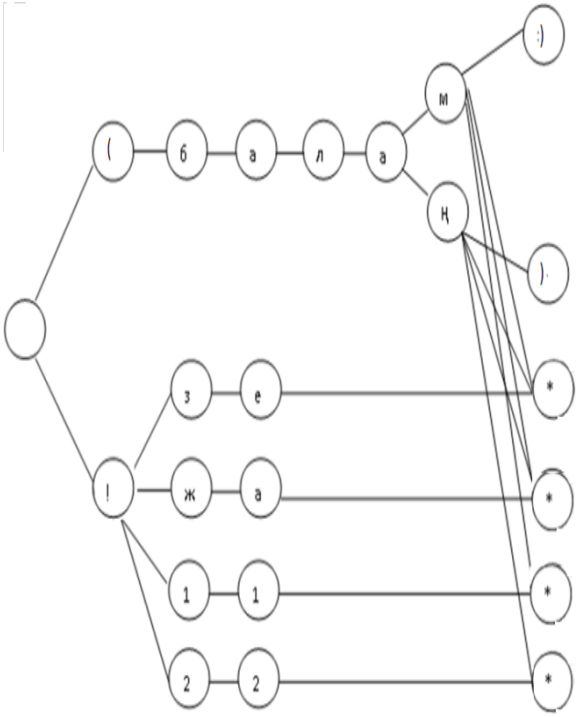

Рисунок 3 – Синхронизированное линейное дерево для словоформ и их морфологической информации

Нейроны рецепторы распознают отдельные символы входной символьной последовательности. На выходе рецептор генерирует сигнал, означающий наличие или отсутствие соответствующего символа в анализируемом тексте. Нейроны-эффекторы выдают результат распознавания отдельных фрагментов входной символьной последовательности. Заменив в синхронизированном линейном дереве сигнал от рецептора сигналом от эффектора того же дерева, получим возможность использовать в качестве входных символов фрагменты символьных последовательностей.

Для обозначения таких фрагментов во входной символьной последовательности будем применять метасимволы скобок: "**(**" и "**)**". Тогда приведенный пример перепишется в виде: ((бала)м), ((бала)ң), (((бала)м)ның), (((бала)ң)да) [Шарипбаев, 2009].

Таким образом формализованы все части речи казахского языка, общее количество формальных правил составило около 20000 записей, например

только для глагола это число 13000 формальных правил.

Ниже приводятся фрагмент формальных правил словоизменения на примере существительного с учетом закона сингармонизма, который обуславливает добавления мягких или твердых окончаний в зависимости от мягкости или твердости основы. Приведенный пример показывает фрагмент правил, где «зе» – зат есім (имя существительное), «жа» - жанды (одушевленность), «01» заканчивается на твердые гласные а, о, ұ, «))» между закрывающими скобками помещены окончания существительных, после «!» морфологическая информация.

((зежа01)мын)!жі11
((зежа01)мыз)!жі11
((зежа01)сың)!жі22
((зежа01)сыңдар)!жі22
((зежа01))!жі33
((зежа01)м)!тә11
((зежа01)мыз)!тә110
((зежа01)ң)!тә22
((зежа01)ңыз)!тә22
((зежа01)сы)!тә33
((зежа01)лар)!кт
(((зежа01)лар)мыз)!ктжі11
(((зежа01)лар)сыңдар)!ктжі22
(((зежа01)лар))!ктжі33
(((зежа01)лар)ым)!кттә11
(((зежа01)лар)ымыз)!кттә110
(((зежа01)лар)ың)!кттә22
(((зежа01)лар)ыңыз)!кттә22
(((зежа01)лар)ы)!кттә33
((зежа01))!ат0
((зежа01)ның)!ілә
((зежа01)ға)!ба
((зежа01)ны)!та
((зежа01)да)!жс
((зежа01)дан)!шы
((зежа01)мен)!кө
((зежа01)менен)!кө
(((зежа01)м)ның)!тә11ілә
(((зежа01)м)а)!тә11ба
(((зежа01)м)ды)!тә11та
(((зежа01)м)да)!тә11жс
(((зежа01)м)нан)!тә11шы
(((зежа01)м)мен)!тә11кө
(((зежа01)м)менен)!тә11кө
(((зежа01)мыз)дың)!тә110ілә
(((зежа01)мыз)ға)!тә110ба
(((зежа01)мыз)ды)!тә110та
(((зежа01)мыз)да)!тә110жс
(((зежа01)мыз)дан)!тә110шы
(((зежа01)мыз)бен)!тә110кө
(((зежа01)мыз)бенен)!тә110кө
(((зежа01)ң)ның)!тә22ілә
(((зежа01)ң)а)!тә22ба
(((зежа01)ң)ды)!тә22та
(((зежа01)ң)да)!тә22жс
(((зежа01)ң)нан)!тә22шы
(((зежа01)ң)мен)!тә22кө
(((зежа01)ң)менен)!тә22кө
(((зежа01)ңыз)дың)!тә22ілә
(((зежа01)ңыз)ға)!тә22ба
(((зежа01)ңыз)ды)!тә22та
(((зежа01)ңыз)да)!тә22жс
(((зежа01)ңыз)дан)!тә22шы
(((зежа01)ңыз)бен)!тә22кө
(((зежа01)ңыз)бенен)!тә22кө
(((зежа01)лар)дың)!ктілә
(((зежа01)лар)ға)!ктба
(((зежа01)лар)ды)!ктта
(((зежа01)лар)да)!ктжс
(((зежа01)лар)дан)!ктшы
(((зежа01)лар)мен)!кткө
(((зежа01)лар)менен)!кткө
((((зежа01)лар)ым)ның)!кттә11ілә
((((зежа01)лар)ым)а)!кттә11ба
((((зежа01)лар)ым)ды)!кттә11та
((((зежа01)лар)ым)да)!кттә11жс
((((зежа01)лар)ым)нан)!кттә11шы
((((зежа01)лар)ым)мен)!кттә11кө
((((зежа01)лар)ым)менен)!кттә11кө
((((зежа01)лар)ымыз)дың)!кттә11ілә
((((зежа01)лар)ымыз)ға)!кттә11ба
((((зежа01)лар)ымыз)ды)!кттә11та
((((зежа01)лар)ымыз)да)!кттә11жс
((((зежа01)лар)ымыз)дан)!кттә11шы
((((зежа01)лар)ымыз)бен)!кттә11кө
((((зежа01)лар)ымыз)бенен)!кттә11кө
((((зежа01)лар)ың)ның)!кттә22ілә
((((зежа01)лар)ың)а)!кттә22ба
((((зежа01)лар)ың)ды)!кттә22та
((((зежа01)лар)ың)да)!кттә22жс
((((зежа01)лар)ың)нан)!кттә22шы
((((зежа01)лар)ың)мен)!кттә22кө
((((зежа01)лар)ың)менен)!кттә22кө
((((зежа01)лар)ыңыз)дың)!кттә22ілә
((((зежа01)лар)ыңыз)ға)!кттә22ба
((((зежа01)лар)ыңыз)ды)!кттә22та
((((зежа01)лар)ыңыз)да)!кттә22жс
((((зежа01)лар)ыңыз)дан)!кттә22шы
((((зежа01)лар)ыңыз)бен)!кттә22кө
((((зежа01)лар)ыңыз)бенен)!кттә22кө
((((зежа01)лар)ым)сыңдар)!кттә11жі22
((((зежа01)лар)ың)быз)!кттә22жі11
((((зежа01)лар)ы)сыңдар)!кттә33жі2

3 Алгоритм морфологического анализа

Предложен следующий алгоритм морфологического анализа слов казахского языка:

1. Слово считывается;
2. Открывается словарь начальных форм и в нем выполняется поиск считанного слова;
3. Если слово найдено, то перейти к шагу 12, иначе шаг 4 ;
4. Слово в цикле посимвольно считывается, начиная с последнего символа, то, что получается, ищем в словаре окончаний;
5. Если окончание найдено, то остаток ищем в словаре начальных форм;
6. Запоминаем морфологическую информацию окончания;

7. Если остаток слова найден в словаре начальных форм, то

8. Если это глагол, то считываем слово слева перед ним и определяем время глагола

9. Если это деепричастие, то считываем слово справа после него и определяем время глагола

10. Если это прилагательное, то считываем слово слева перед ним и проверяем его входит ли оно в список вспомогательных слов, использующихся для образования превосходных степеней, если входит, то это превосходная степень прилагательного.

11. Если такое слово не найдено, то переходим к шагу 4, иначе к шагу 12;

12. Конец.

Это алгоритм реализован на языке программирования Borland Delphi.

С помощью построения морфологических правил и генерации слов казахского языка были получены следующие результаты:

- создана база данных начальных форм слов объёмом 45 000 слов с разметкой частей речи и других признаков, необходимых для генерации словаря словоформ; получена формальная модель словоизменения и словообразования казахского языка с учетом семантики на основе семантической нейронной сети;
- автоматически сгенерирована база данных казахских словоформ объёмом более 2 800 000 словарных статей с полной морфологической информацией;
- разработаны алгоритмы и программы морфологического анализа естественно-языковых текстов с учетом семантики на основе семантической нейронной сети и клеточных автоматов;

ЗАКЛЮЧЕНИЕ

Наибольшее количество словоформ генерируется из начальной формы существительного, прилагательного и начальной формы глагола. Полученные словари могут быть изданы в качестве орфографических словарей. Полученные формализации, методы и алгоритмы могут использоваться в системах обработки естественно-языковых текстов (орфографических корректорах, переводчиках, обучающих системах), системах распознавания и синтеза казахской речи, а также в семантических поисковых системах.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

[Казахская грамматика, 2002] Казахская грамматика. Фонетика, словообразование, морфология, синтаксис. – Астана, 2002.

[Бекманова, 2009] Бекманова Г. Т. Некоторые подходы к проблемам автоматического словоизменения и морфологического анализа в казахском языке // Вестник Восточно-Казахстанского государственного технического университета им. Д. Серикбаева. – Усть-Каменогорск: 2009г. – № 4. – с. 192-197.

[**Шуклин, 2001**] Шуклин Д. Е**.** Структура семантической нейронной сети, реализующей морфологический и синтаксический разбор текста // Кибернетика и системный анализ. Киев. Изд-во Института кибернетики НАН Украины, 2001. - No 5. с. 172-179.

[Шарипбаев, 2009] Шарипбаев А. А., Бекманова Г. Т. Построение логической семантики слов казахского языка // Материалы Всероссийской конференции с международным участием «Знания-Онтологии-Теории (ЗОНТ-09)». – Новосибирск: 2009. – Том 2. – с. 246-249.

THE INTELLECTUAL MORPHOLOGICAL ANALYZER BASED ON SEMANTIC NETWORKS

Sharipbaev A.A., Bekmanova G.T., Ergesh B.J., Buribaeva A.K., Karabalaeva M.H.

Eurasian National University named after L.N. Gumilev, Astana, Kazakhstan

sharalt@mail.ru
gulmira-r@yandex.ru
banu_kazakh@yahoo.com
buribayeva@mail.ru
mkarabal@mail.ru

In this article the order of construction of semantic base of initial words is described, morphological rules of the Kazakh language are formalized. In the Kazakh language of a word form are formed by coupling of a root and affixes (suffixes and the terminations). Thus each affix is connected with sets of semantic signs and the order of addition of affixes is strictly defined. For example, for nouns to a word basis the suffix and further plural termination, then the possessive termination in the beginning is added, the case inflection further follows and the last the termination of the form of conjugation On the basis of formal rules develops algorithm of the morphological analyzer.