OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

СИНТАКСИЧЕСКИЙ АНАЛИЗАТОР КАЗАХСКОГО ЯЗЫКА НА ОСНОВЕ ГРАММАТИКИ СВЯЗЕЙ - LINKGRAMMARPARSER

Бегимтай У.Х.

Евразийский Национальный Университет имени Л.Н.Гумилева,
г. Астана, Республика Казахстан

ulugbek_begimtai@mail.ru

В работе описан принцип работы синтаксического анализатора на основе грамматики связей – linkgrammarparser для казахского языка. Парсер грамматики связей позволяет размечать слова тегами частей речи и определять тип связей между ними. Также в данной работе показан принцип проверки сходства двух предложений и принцип вычисления числа на семантическом графе соответствующих и несоответствующих дуг.

Ключевыеслова:linkgrammarparser, синтаксический анализатор, парсер грамматики связей, семантический анализ казахского языка.

Введение

Информатизация общества приводит к изменению ситуации на трудовом рынке, требуя все новые квалификации и компетентности. Это заставляет большинство людей в корне менять свою специализацию или обучаться новым знаниям. Меняются традиционные формы обучения и виды услуг в области образования. В мире наступает эра электронного образования, в котором количество студентов стало более 150 млн. человек. Что касается Казахстана, то сектор электронного образования почти отсутствует.

Если сравнивать традиционное образование с электронным образованием, то доказано, что в последнем скорость образования на 30-60% быстрее, но качество низкое. Причиной низости качества электронного образования является пассивность и статичность электронных образовательных ресурсов (как правило, простой текстовый или графический материал) и отсутствие диалога с обучающимися в реальном масштабе времени. Ведь при изучении этих ресурсов у обучающегося может появиться масса вопросов к ним, а у них просто нет возможности ответить на эти вопросы, которые с продолжением изучения могут только возрастать. Кроме того, современные методы контроля и оценки знаний не всегда выдают объективную оценку знаний. Эти проблемы можно решить, если статические электронные образовательные ресурсы заменить интеллектуальными анализаторами текстов для проверки качества знания.

В этой статье написана часть технологий проверки знания путем интеллектуального сравнения двух предложений - "предложение запрос" и "предложение претендент". Данная технология позволит автоматизировать процесс проверки знания обучающегося путем интеллектуального анализа текста и интеллектуальной оценки знания на основе грамматики связей - linkgrammarparser.

Общий принцип работы

1.1. Сравнение семантических графов двух предложений

Предположим, даны два предложения $\bar{x}_1$ и $\bar{x}_2$, и второе предложение необходимо сравнить с первым. Будем называть предложение $\bar{x}_1$ запросом, а предложение $\bar{x}_2$ – претендентом. Предположим есть семантический граф для $\bar{x}_1$

$$G_1 = \langle V_1, E_1, s_1, r_1 \rangle \quad (1)$$

и соответствующий семантический граф для $\bar{x}_2$

$$G_2 = \langle V_2, E_2, s_2, r_2 \rangle \quad (2)$$

Каждой дуге семантического графа запроса сопоставляется «равная» ей дуга в семантическом графе претендента.

- V – множество вершин графов

- E – множество дуг графов*(таблица 1)*
- s:V→L, L –подмножество графов
- r:E→M, M – множество семантико-синтаксических отношений.

После этого задается отображение двух графов:

$$F : G_1 \to G_2 \quad (3)$$

Чтобы определить сходство двух предложений $\bar{x}_1$ и $\bar{x}_2$ нужно вычислить число на семантическом графе как соответствующих, так и несоответствующих дуг. Вес каждой дуги соответствует семантико-синтаксическому отношению.

Таблица 1 – Пометки дуг семантического графа

Обозн-е	**Семантико-синтак-ое отношение**	**Пример**	**Вес**
Пример: Кеше(вчера) өте(очень) ашулы(злая) ит(собака) адамға(на человека) қатты(сильно) үрді(лаяла) де(и) қолындағы затын(вещь которую держал в руке) ұрлап кетті(своровала)			
SUB (PRED_ACT_SUB)	Соответствует связи между действием и действующим субъектом	үрді $\xrightarrow{SUB}$ ит ұрлады $\xrightarrow{SUB}$ ит	1
dirOBJ (PRED_ACT_DIR_OBJ)	Соответствует связи между действием и «прямым» объектом действия	ұрлады $\xrightarrow{dirOBJ}$ затты	0,2
indirOBJ	Соответствует связи между действием и «косвенным» объектом действия	үрді $\xrightarrow{indirOBJ}$ адамға	0,2
ATTR (N_ATTR)	Соответствует связи между объектом и его признаком бъект пен оның нышаны арасындағы байланыс	ит $\xrightarrow{ATTR}$ ашулы	0,1
mannCIR (ADV_MAN_ADV)	Соответствует связи между действием и признаком образа действия или признаком и его степенью	үрді $\xrightarrow{mannCIR}$ қатты ашулы $\xrightarrow{mannCIR}$ өте	0,1
timeCIR	Соответствует связи между действием и временной характеристикой действия	үрді $\xrightarrow{timeCIR}$ кеше ұрлады $\xrightarrow{timeCIR}$ кеше	0,1

1.2. Принцип вычисления совпадения двух предложений

Ниже представлена формула вычисления степени совпадения предложений, подходящая для ранжирования претендентов и отвечающая вышеизложенным принципам:

$$y = \frac{\sum_{i=1}^{N} p_i - \left(\frac{\sum_{i=1}^{M} q_i}{\sum_{i=1}^{\tilde{M}} t_i} \right)}{\sum_{i=1}^{K} r_i}, \quad (4)$$

где:

y – коэффициент совпадения претендента предложения с запросом;

K, N– число дуг в семантическом графе запроса и подграфе претендента, состоящем из совпадающих дуг;

$\tilde{M}$ – общее число дуг в семантическом графе претендента;

M – число несовпадающих дуг в семантическом графе претендента,

$$M = \tilde{M} - \sum_{i=1}^{N} N_i \quad (5)$$

r_i, t_i – веса дуг семантических графов запроса и претендента соответственно;

p_i– вес совпадающей дуги в семантическом графе претендента;

q_i – вес несовпадающей дуги в семантическом графе претендента.

Таким образом, чем больше в семантическом графе претендента имеется совпадающих дуг, и чем больше их веса, тем большую оценку он должен получить. Кроме того, в формуле присутствует корректирующее слагаемое.

$$-\frac{\left(\sum_{i=1}^{M} q_i \right)}{\left(\sum_{i=1}^{\tilde{M}} t_i \right)} \quad (6)$$

Оно по абсолютной величине не превышает единицу и служит для ранжирования претендентов, имеющих одинаковое число равнозначных совпадений с запросом. При этом чем легче несовпадающие дуги, тем меньше снижается оценка претендента.

Пример работы парсера

1.3. Обработка предложения предендента

Для примера возьмем предложение на казахском языке:

X_1 - **"Қызыл(красная) түлкі(лиса) ақ(белого)**

қоянды(зайца) өте(очень) тез(быстро) жеді(съела)".

После обработки получаем следующий синтаксический вид:

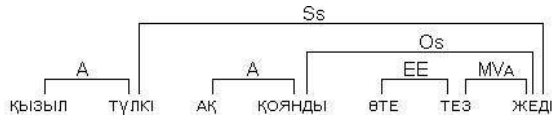

Рисунок 1 - Связь коннекторов

V - {қызыл, түлкі, ақ, қоянды, өте, тез, жеді}

E - {A, Ss, MVa, Os, A, EE}

Каждое слова приводим в "нормализованный вид" и указываем к какой части речи принадлежит каждое из них. Вершины этих графов являются "базовыми метасловами".

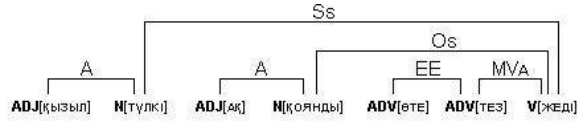

Рисунок 2 - Определение базовых метаслов

Затем происходит построение производных метаслов и формируется конечный набор метаслов.

В нашем случае "базовое метаслово" V[жеді](был съеден) превращается в "производное метаслово" O_SH_3[же](есть), а "базовое метаслово" N[қоянды](зайца) превращается в "производное метаслово" TBS_3[қоян](заяц). Таким образом, граф, состоящий из конечного набора метаслов, будет иметь следующий вид:

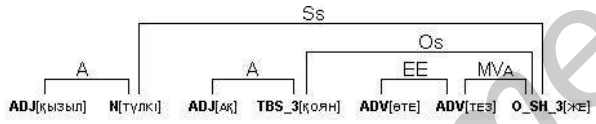

Рисунок 3 - Определение производных метаслов

После выбора главной связи и главного метаслова предложения, которые в данном случае равны Ss O_SH_3[же] соответственно, получится дерево метаслов:

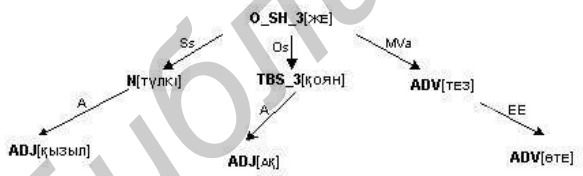

Рисунок 4 - Дерево метаслов

Теперь каждый лист дерева будет втягиваться в своего родителя для построения метасвязей.

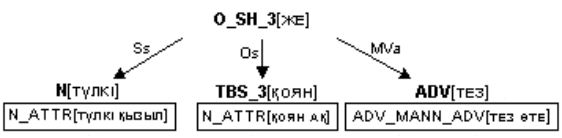

Рисунок 5 - Построение метасвязей

Таким образом, получается конечный набор метасвязей:

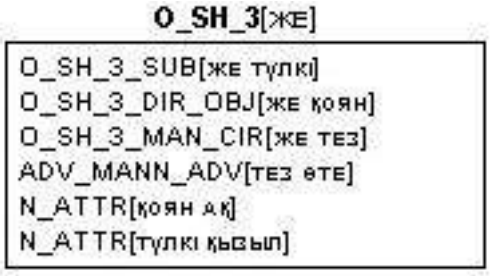

Рисунок 6 - Конечный набор метасвязей

На основе этого набора будет построен следующий семантический граф:

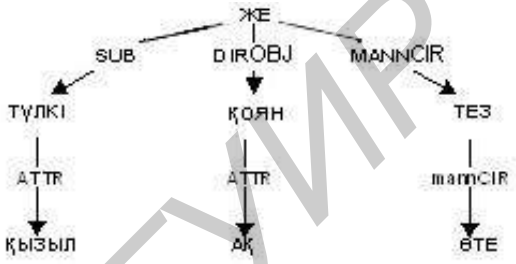

Рисунок 7 - Семантический граф

1.4. Обработка предложения запроса

Пусть теперь имеется второе предложение: X_2 - **"Ақ(белый) қоянды(заяц) жеген(съеденный) түлкі(лисой)".** Синтаксическая структура, порожденная системой LinkGrammarParser, будет иметь следующий вид:

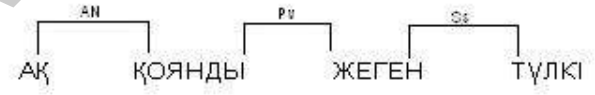

Рисунок 8 - Связь коннекторов

Этот граф обрабатывается аналогично первому. После построения финального набора метаслов и выбора корня получится следующее дерево метаслов:

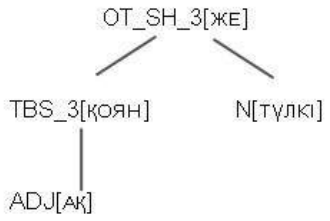

Рисунок 9 - Дерево метаслов

После стягивания графа получится набор метасвязей, указанный ниже:

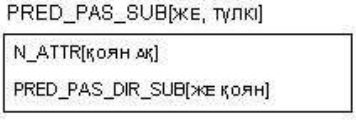

Рисунок 10 - Набор метасвязей

По данным метасвязям будет построен следующий семантический граф:

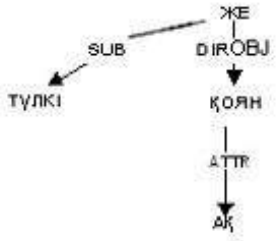

Рисунок 11 - Семантический граф

Сопоставление двух графов и оценка степени совпадения предложений

Далее производится сопоставление семантических графов. Предположим, что первое предложение – это запрос, а второе – претендент.

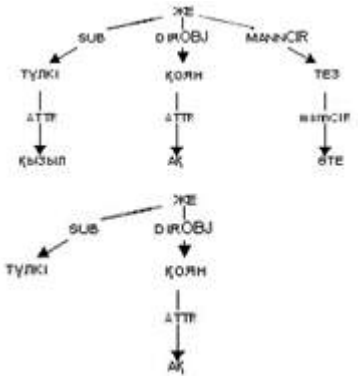

Рисунок 12 - Сопоставление двух графов

Таким образом, получается, что в подграфе второго графа, состоящем из «совпадающих» дуг, имеется одна компонента связности, состоящая из трех дуг. Дуг, не помеченных как «совпадающие», во втором графе нет. Поэтому получается следующая оценка степени совпадения второго предложения с первым

$$y = \frac{3\cdot(p_{SUB} + p_{ATTR} + p_{dirOBJ})}{6\cdot(p_{SUB} + p_{dirOBJ} + 2\cdot(p_{ATTR} + p_{mannCIR}))} = \frac{3\cdot(100+5+25)}{6\cdot(100+25+2\cdot(5+5))} = 0{,}45$$

Заключение

Предложенный же метод основан на предположении, что на вход ему подается правильная диаграмма, в которой все связи расставлены так, как их расставил бы человек, поэтому если на вход будет подана некорректная диаграмма, то и семантическое дерево будет отображать связи между понятиями неверно, то есть так, как они отражены в разборе.

На данный момент проработано относительно небольшое количество синтаксических конструкций, поэтому только это небольшое число конструкций может анализироваться. Дальнейшее расширение словарей повлечет расширение множества анализируемых конструкций.

Библиографический список

[Бениаминов, 2008]Бениаминов, Е.М. О построении Web-сервера в стиле SemanticWiki с открытым контекстным языком представления и запросов/Е. М. Бениаминов// КИИ-2008. Трудыконференции. Т 2, С. 15-21

[Temperley D., Sleator D., Lafferty J.]Temperley D., Sleator D., Lafferty J.Link Grammar Documentation. – 1998./Electronic resource//: http://www.link.cs.cmu.edu/link/dict/index.html

[Murzin F., Perfliev A., Shmanina T.]Murzin F., Perfliev A., Shmanina T.Methods of syntactic analysis and comparison of constructions of a natural language oriented to use in search systems/Murzin F., Perfliev A., Shmanina T.//Bull. Nov. Comp. Center, Comp. Science. – 2010. – Iss. 31. – pp. 91-109.

[БатураТ.В., МурзинФ.А.]Батура Т.В., Мурзин Ф.А. Машинно-ориентированные логические методы отображения семантики текста на естественном языке: моногр. Институт систем информатики им. А.П. Ершова СО РАН./Батура Т.В., Мурзин Ф.А.// – Новосибирск: Изд. НГТУ, 2008. – 248 с.

SYNTACTICAL ANALYZER OF KAZAKH LANGUAGE BASED BY "LINK GRAMMAR PARSER"

Begimtay U.H.

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

ulugbek_begimtai@mail.ru

In work the main concepts of semantic model for parsing sentences in kazakh language by link grammar parser. And in main part this work show one example of parsing kazakh proposal.

Introduction

In this part describes the general condition of the intellectualization of the world and particulary in Kazakhstan.

Main Part

In main part it covers the general principle of operation link grammar parser for kazakh language. Step by step visually demonstrate how parses sentences

Conclusion

In conclusion shows the result of processing offers a specific example.