5. The Kazcorpus Project, 2013. – URL: http://kazcorpus.kz/klcweb/en/ (accessed: 18 July 2019).

6. Батура, Т.В., Бакиева, А.М., Еримбетова, А.С., Мурзин, Ф.А., Сагнаева, С.К. Грамматика связей, релевантность и определение тем текстов // Институт систем информатики им. А.П. Ершова СО РАН. – Новосибирск: Издательство СО РАН, 2018. – 91 стр.

7. Tussupova, M.J., Murzin, F.A., Yerimbetova, A.S. Filling up Link Grammar Parser dictionaries by using Word2vec techniques // Joint issue of the International Conference "Computational and Information Technologies in Science, Engineering and Education" (CITech-2018), September 25-28. Vestnik EKSTU after D. Serikbayev, Computational Technologies. Vol 1, Part III /EKSTU after D. Serikbayev. – Ust-Kamenogorsk – Novosibirsk, 2018. – P. 169–176.

АВТОМАТИЧЕСКАЯ ГЕНЕРАЦИЯ СТРУКТУРИРОВАННОЙ МАШИННО-ЧИТАЕМОЙ ИНФОРМАЦИИ ИЗ МУЛЬТИЯЗЫЧНЫХ ТЕКСТОВ

Хайрова Н.Ф.[1], Мамырбаев О.Ж.[2], Мухсина К.Ж.[3], Колесник А.С. [4]
nina_khajrova@yahoo.com, morkenj@mail.ru, kuka_ai@mail.ru, kolesniknastya20@gmail.com
[1,4] Национальный технический университет «Харьковский политехнический институт», Украина
[2]Институт информационных и вычислительных технологий КН МОН РК, Казахстан
[3]Казахский Национальный университет им. Аль-Фараби, Казахстан

Open Information Extraction представляет современную стратегию извлечения фактов из коллекций веб-документов. Однако, большая часть современных подходов по извлечению фактов основана на таких, доступных не для всех естественных языков, техниках NLP, как POS-tagging, анализ зависимостей, Named Entity Recognition, Coreference Resolution и др. В этой работе для генерации фактов из текста произвольного веб-контента мы предлагаем использование уравнений алгебры конечных предикатов, выражающих семантические роли участников триплета факта через отношения грамматических и семантических характеристик слов предложения. Модель позволяет извлекать неограниченное количество доменно-независимых фактов из предложений разных языков. В работе показана имплементация модели для английского, казахского и русского языков.

Ключевые слова: *Open Information Extraction; извлечение доменно-независимых фактов из неструктурированных текстов; логико-лингвистическая модель; алгебра конечных предикатов; казахский, русский и английский языки*

Введение

В последние годы непрерывно растет интерес к исследованиям, связанным с такими задачами искусственного интеллекта, как извлечение информации (Information Extraction (IE)), Open Information Extraction (OIE) и извлечение фактов из неструктурированных и полуструктурированных текстов. Такие технологии генерации информации из текстов естественного языка в структурированный формат, выделяющий целевые факты, объекты, отношения, позволяют автоматически просматривать и перерабатывать большие объемы документов, содержащих относительно небольшое количество полезной информации. Результаты таких исследований могут быть использованы для улучшения машиночитаемости информационных ресурсов путем создания баз знаний в формате Resource Description Framework (RDF) или онтологии.

Иногда IE рассматривается как специфический вид информационного поиска (Information Retrieval (IR)). При этом, несмотря на то, что в обоих случаях запросы известны заранее, в результате IE создается структура данных, описывающая соответствующие факты из набора документов, в то время как в результате IR выдается набор ссылок на документ.

Как правило, системы IE включают несколько задач: (1) идентификацию сущностей (Name Entity Recognition (NER)); (2) задачу снятия ко-референтности (coreference resolution) — поиск всех лингвистических выражений, которые ссылаются на один и тот же объект внеязыковой действительности; (3) распознавание семантических ролей участников действия в предложении; (4) определение отношений между сущностями [1].

Обычно системы IE базируются на наборе правил, идентифицирующих информацию, которая должна быть извлечена из текста и представляют результат в виде кортежей двух объектов, с предопределенным типом отношения между ними. В конкретной заранее выбранной области существует несколько предопределенных типов таких отношений [2]. Такой подход не масштабирует корпуса, где число целевых отношений очень велико или где целевые отношения не могут быть определены заранее [3].

Несколько лет назад OIE стало новой парадигмой извлечения информации, работающей с неограниченным числом отношений, игнорирующей специфичные для домена обучающие данные и линейно их масштабирует [4]. В настоящее время предложено множество различных систем Open IE, обычно базирующихся на методах POS-tagging и синтаксическом разборе зависимостей [5, 6], дополнительно использующие лексические ограничения [3], семантические аннотации [7], или другие подходы [8] для минимизации узкоспециализированных отношений.

Однако, к сожалению, на сегодняшний день не для всех естественных языков разработаны подобные инструментарии и ресурсы. Не смотря на наличие разработанных методов Open IE не только для английского, но и для некоторых других языков [9], проблема разработки методов автоматической генерации структурированной машинно-читаемой информации из мультиязычных текстов остается по-прежнему актуальной.

В нашем исследовании мы предлагаем логико-лингвистическую модель, позволяющую извлекать факты из текстов Web-контента казахского, русского и

английского языков. Согласно парадигме Open IE мы рассматриваем факт в виде триплета: *Субъект* - *Предикат* - *Объект*, где *Предикат* выражает семантическое действие, *Субъект* определяет инициатора действия, а *Объект* определяет участника действия, на которого действие направлено. В предложении естественного языка предикат выражается глаголом, а объект и субъект действия представлены существительными или именной группой. Кроме того, факт может содержать несколько атрибутов, таких как время, место, способ действия, принадлежность участников действия и другие, которые могут быть представлены так же именными группами.

Математические средства используемой модели

Основными математическими средствами нашей модели являются логико-алгебраические уравнения алгебры конечных предикатов. Вводим универсум *U,* представляющий совокупность различных элементов языковой системы того или иного естественного языка: предложения, фразы, слова, грамматические и семантические характеристики, морфемы и т. д. На основании того факта, что элементы универсума конечны, определены и детерминированы, мы можем сказать, что введенный универсам конечен и детерминирован [10].

Множество M = {m_1,…, m_n } представляет собой подмножество грамматических и семантических характеристик слов предложения естественного языка, где n – количество характеристик определенных в системе.

Переменная x_i^a называется предметной переменной и описывает наличие признака *a* у слова *i*. Согласно алгебре предикатов:

$$x_i^a = \begin{cases} 1, if\ x_i = a \\ 0, if\ x_i \neq a \end{cases} (1 \le i \le n), \quad (1)$$

Например, уравнение $x_i^{gen} = 1$ означает, слово i стоит в родительном падеже или обладает признаком родительного падежа, а дизъюнкция $x_i^{gen} \vee x_i^{nom} = 1$ означает, что слово *i* имеет родительный или именительный падеж.

На следующем шаге нашей модели мы вводим систему предикатов S. В нашей модели предикат $P_i(x_i) \in S$ равен 1, если грамматические и семантические признаки принадлежат слову, которое может быть частью триплета, в противном случае $P_i(x) = 0$. Многомерный предикат $P(x_1,..., x_n)$ определяет семантическую роль существительного через предметные переменные, описывающие грамматические и семантические характеристики слов в предложении:

$$P(x_1,..., x_n) \to P(x_1) \wedge ... \wedge P(x_n) \quad (2)$$

Предикат, $P(x_1,..., x_n) = 1$ если признаки существительных предложения имеют определенные значения. Это означает, что слово, конъюнкция грамматических и семантических признаков которого описано предикатом (2), представляет участника

(*Субъект* или *Объект*) или *атрибут* действия. Очевидно, что отношения признаков существительного не зависят от конкретного токена.

На практике подмножество согласованных морфологических и синтаксических признаков участников действия не совпадает с декартовым произведением по совокупности всех признаков. Мы можем определить предикат на декартовом произведении $S \times \mathrm{S}$:

$$P(x_1,..., x_n) = \gamma_k(x_1,..., x_n) \times P_1(x_1) \times .... \times P_n(x_n). \quad (3)$$

$k \in [1,h]$, где h – число рассматриваемых в модели участников и атрибутов фактов (*Subject, Object, Predicate, Location, Time* и др.). Предикат $\gamma_k(x_1,..., x_n) = 1$, если заданные морфологические и синтаксические характеристики слова предложения выражают определенное семантическое значение участника или атрибута действия, и $\gamma_k(x_1,..., x_n) = 0$, если конъюнкция грамматических признаков слова не определяет никакой семантической роли. В таком случае, если отношения между морфологическими и синтаксическими характеристиками слов предложения не представляют никаких элементов факта, они исключаются из формулы (3) предикатом $\gamma_k(x_1,..., x_n)$.

Мы имплементируем нашу модель для идентификации и генерации фактов из текстов казахского, английского и русского языков. Во всех этих языках семантические роли участников действия явно выражены в поверхностной структуре языка. Однако в связи с существующими различиями в синтаксисе и морфологии английского, русского и казахского языков, для каждого языка модель реализована по-разному.

Реализация логико-лингвистической модели OIE для русского и английского языков

В качестве грамматических и синтаксических характеристик слов английского предложения, формализующих возможные составляющие триплета факта или его атрибуты, определены такие свойства как: наличие предлога после глагола, притяжательный падеж существительного или местоимения, положение существительного во фразе, наличие любой формы глагола «*to be*», форма основного глагола во фразе [11]. В результате введено конечное множество предметных переменных, описывающих грамматические характеристики слов английского предложения $\{x, z, m, f, n, p\}$.

Предикат γ_{1E} определяет отношения грамматических и семантических характеристик слов, обозначающих *Субъект* триплета факта в английской фразе:

$$\begin{aligned}\gamma_{1E}(z, y, x, m, p, f, n) = y^{out} ((f^{can} \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{could} \vee f^{need} \vee \\ f^{might} \vee f^{would} \vee \\ \vee f^{out}) (n^{not} \vee n^{out}) (p^{I} \vee p^{ed} \vee p^{III}) x^{f} m^{out} \vee (x^{l} (m^{is} \vee m^{are} \vee m^{havb} \vee \\ \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were} \vee m^{be} \vee m^{out}) z^{by}).\end{aligned} \quad (4)$$

Мы так же можем явным образом выделить *Объект* или участника действия, на которого это действие направленно, через предикат γ_{2E}:

$$\gamma_{2E}(z, y, x, m, p, f, n) = y^{out}\,(n^{not} \vee n^{out})\,(f^{can} \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{could} \vee f^{need} \vee f^{might} \vee \vee f^{would} \vee f^{out})\,(z^{out}\,x^{l}\,m^{out}\,(p^{I} \vee p^{ed} \vee p^{III}) \vee x^{f}\,(z^{out} \vee z^{by})\,(m^{is} \vee \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were} \vee m^{be} \vee m^{out})\,(p^{ed} \vee p^{III}). \quad (5)$$

Адаптируя разработанную модель к русскому языку, вводим множество грамматических и семантических характеристик слов русскоязычных предложений $M = \{z, y, x\}$, где z — конечное подмножество грамматических падежей русских существительных, y — конечное подмножество возможных семантических характеристик существительного, а x — подмножество характеристик одушевленности [12].

Тогда, предикат P_z (z) определяет шесть грамматических падежей русского языка (*nom* — именительный, *gen* — родительный, *dat* — дательный, *acc* — винительный, *ins* — творительный и *loc* — предложный):

$$P_z(z) = z^{nom} \vee z^{gen} \vee z^{dat} \vee z^{acc} \vee z^{ins} \vee z^{loc} \quad (6)$$

Предикат $P_x(x)$ определяет семантическую характеристику одушевленности существительного:

$$P_x(x) = x^{anim} \vee x^{inan}, \quad (7)$$

где значение предметной переменной *anim* определяет признак одушевленности существительного, а *inan* — соответственно, признак неодушевленности.

Предикат $P_y(y)$ идентифицирует специальные семантические характеристики существительного:

$$P_y(y) = y^{\text{device}} \vee y^{\text{hum}} \vee y^{\text{tool}} \vee y^{\text{pc:hue}} \vee y^{\text{spacee}} \vee y^{\text{time:moment}} \vee y^{\text{time:period}} \vee y^{\text{s:loc}}, \quad (8)$$

здесь значения *device* и *tool* показывают, что существительное называет концепт, являющийся устройством или инструментом, *hum* обозначает, что концепт, который называет существительное, принадлежит к семантическому классу "*person*", *pc:hum* обозначает семантический класс "часть тела", *time:moment* — семантическая характеристика “определенное время” и *time:period* — наличие семантического признака периода времени, *space* — семантическая характеристика месторасположения. При формировании множества значений предикатных переменных мы использовали таксономические отношения существительных Русского Национального корпуса[4].

[4] http://www.ruscorpora.ru/old/en/corpora-sem.html

Согласно нашей модели, на следующем шаге определяем семантические роли *Субъекта* и *Объекта* триплета факта в русскоязычной фразе через предикаты γ_{1R} и γ_{2R}, соответственно:

$$\gamma_{1R}(x,y,z) = x^{anim} z^{\text{nom}} \vee x^{inan} z^{\text{nom}} (y^{\text{device}} \vee y^{\text{tool}} \vee y^{\text{pc:hue}}) \quad (9)$$

$$\gamma_{2R}(x,y,z) = z^{\text{acc}} (x^{inan} \vee x^{anim}) \quad (10)$$

Реализация логико-лингвистической модели OIE для казахского языка

Казахский язык, в отличие от русского или английского, является агглютинативным языком. Это означает, что слово строится из морфем, каждая из которых имеет определенное морфологическое или семантическое значение. Что противоположно флективному языку, где каждая морфема имеет несколько неразделимых значений одновременно (например, падеж, род, число и т. д.) и аналитическому языку, в котором почти нет флексий. Когда мы корректируем нашу модель для казахского языка, мы вводим множество *M* из десяти грамматических признаков слов казахского языка, большинство из которых представляют собой те или иные типы суффиксов, имеющих особое семантическое или морфологическое значение. К характеристикам слов казахского языка, влияющим на их семантическую роль, мы отнесли следующие: положение анализируемого слова во фразе; существование вспомогательного глагола; наличие или отсутствие суффиксов множественного числа; грамматический падеж анализируемого существительного; семантическое значение, представленное конкретными суффиксами; наличие отрицания в анализируемой фразе; аффиксы предопределённого действия; условное наклонение и некоторые другие.

Тот факт, что множество *M* грамматических особенностей казахского языка гораздо больше, чем сопоставимый набор грамматических признаков русского или английского языков, обусловлен двумя основными причинами. Прежде всего, сложностью казахского языка, в котором существует множество морфологических и семантических характеристик, каждая из которых обычно выражается определенным аффиксом. Вторая причина использования большого количества грамматических особенностей заключается в том, что в случае казахского языка мы рассматриваем и анализируем не только участников действия, но и различные типы действий, т.е.

Предикат $P_x(x)$ определяет местоположение анализируемого слова во фразе

$$P_x(x) = x^1 \vee x^2 \vee x^3 \vee x^{-1} \vee x^{-2} \vee x^{-3} \vee x^0, \quad (11)$$

где 1, 2, 3, -1, -2, -3 показывают смещение слова во фразе, «минус» обозначает начало отсчета с конца фразы; 0 показывает любую другую позицию слова, кроме первых трех и последних трех слов в предложении.

Предикат $P_f(f)$ определяет, есть ли во фразе вспомогательный глагол:

$$P_f(f) = f^{aux} \vee f^0, \quad (12)$$

где *aux* показывает признак существования любого глагола из списка 35 вспомогательных глаголов казахского языка в анализируемой фразе.

Предикат $P_z(z)$ идентифицирует семь грамматических падежей казахского языка: именительный, родительный, дательно-направленный, винительный, местный, творительный и исходный:

$$P_z(z) = z^{Nom} \vee z^{Gen} \vee z^{Dat} \vee z^{Acc} \vee z^{Ela} \vee z^{Ins} \vee z^{Abl} \quad (13)$$

Предикат $P_a(a)$ определяет два возможных типа склонения казахских существительных (простой и притяжательный):

$$P_a(a) = a^{\mathrm{NSim}} \vee a^{\mathrm{NPos}}, \quad (14)$$

где *NSim* – признак простого склонения существительного, *NPos* – признак притяжательного склонения существительного. Предикат $P_n(n)$ идентифицирует особенность отрицательного предложения:

$$P_n(n) = n^{me} \vee n^{\mathrm{emes}} \vee n^{\mathrm{joq}} \vee n^{0}, \quad (15)$$

где *me* - признак отрицательного предложения, который представлен существованием частицы из списка [*ma, me, ba, be, pa, pe*], *emes* и *joq* – признак отрицательного предложения, который представлен наличием слов «*emes*» и «*joq*», соответственно, в предложении; 0 показывает отсутствие какого-либо признака отрицания в предложении.

Предикат P_c(c) определяет наличие или отсутствие множественных суффиксов:

$$P_c(c) = c^{\mathrm{tar}} \vee c^{\mathrm{ter}} \vee c^{\mathrm{dar}} \vee c^{\mathrm{der}} \vee c^{\mathrm{lar}} \vee c^{\mathrm{ler}} \vee c^{0},$$

где *tar, ter, dar, der, lar, ler* показывают наличие множественного суффикса с тем же именем в анализируемом слове.

Предикат $P_b(b)$ определяет наличие некоторой дополнительной семантики или значения анализируемого глагола:

$$P_b(b) = b^{\mathrm{se}} \vee b^{\mathrm{mic}} \vee b^{0}, \quad (16)$$

где *mic* обозначает предполагаемость действия, *se* обозначает условное наклонение, а 0 обозначает отсутствие некоторой дополнительной семантики анализируемого глагола.

Следующие несколько функций связаны с семантическим значением, представленным определенными суффиксами. Предикат $P_y(y)$ идентифицирует

словообразовательные суффиксы, которые образовывают глаголы, существительные, причастия, наречия:

$$P_y(y) = y^{\mathrm{ParP}} \vee y^{\mathrm{Vpas}} \vee y^{VaP} \vee y^{UnFu} \vee y^{FuCo} \vee y^{VAd} \vee y^{OAd} \vee \mathrm{y}^{\mathrm{Psuf}} \vee y^{Usuf} \vee \quad (17)$$
$$\vee\, y^{Part} \vee y^{NoV} \vee y^{NoN} \vee y^{NCom} \vee y^{NDer} \vee y^{\mathrm{y}} \vee y^{0},$$

где:

– *UnFu, FuCo* - это свойство включения суффикса неопределенного будущего времени и будущего предположительного времени, соответственно, в анализируемое слово;

– Psuf и *Usuf* обозначают включение одного из 189 производительных или одного из 65 непродуктивных суффиксов, соответственно, из определенных списков в анализируемом глаголе;

– *NoN, NoV* - особенности образования существительного (NoN - от существительного, NoV - от глагола);

– *Ncom* - особенность включения сложного суффикса образования существительного в анализируемое слово;

– *Nder* - это особенность существования некоторой экспрессии уменьшительные и уничижительные оттенки);

– *Part, ParP* - это особенности генерации причастия с помощью двух разных списков суффиксов;

– *VaP, Oad, Vad* - это особенности формирования деепричастий с помощью трех разных списков суффиксов;

– *Vpas* - это особенность включения одного из 20 суффиксов глагола в анализируемое слово:

– *y* - признак существования суффикса формы инфинитивного глагола,

– а *0* - знак основы глагола (форма второго лица единственного числа, будущего императивного времени).

Предикат P_d(d) определяет существование сослагательного наклонения у анализируемого глагола:

$$P_d(d) = d^{\mathrm{shi}} \vee d^{0}, \quad (18)$$

где *shi* показывает включение суффикса сослагательного наклонения в анализируемый глагол, а 0 показывает отсутствие таких суффиксов.

Предикат $P_m(m)$ определяет наличие личного предикативного или притяжательного окончания у анализируемого слова:

$$P_m(m) = m^{\mathrm{PrFl}} \vee m^{\mathrm{PoFl}} \vee m^{0}, \quad (19)$$

где *PrFl* показывает существование личного предикативного окончания причастия, глагольного наречия, основного или вспомогательного глагола, а *PoFl* показывает наличие личного притяжательного окончания причастия, глагольного наречия, основного или вспомогательного глагола.

Следуя уравнениям (11) - (19) мы можем преобразовать предикат согласования грамматических и семантических характеристик слов, являющихся элементами факта (3) для казахского языка в уравнение:

$$P() = \gamma_k \times P_x(x) \times P_y(y) \times P_z(z) \times P_f(f) \times P_m(m) \times P_n(n) \times P_a(a) \times$$
$$\times P_b(b) \times P_c(c) \times P_d(d).$$

Мы можем определить Объект триплета факта казахской фразы с помощью следующего предиката γ_{2K}:

$$\gamma_{2K} = (x^0 \vee x^2 \vee x^3)(z^{Gen} \vee z^{Acc})(y^{NoV} \vee y^{NoN} \vee y^{NCom} \vee y^{NDer} \vee y^0) \wedge$$
$$\wedge c^{tar} \vee c^{ter} \vee c^{dar} \vee c^{der} \vee c^{lar} \vee c^{ler} \vee c^0 \, a^{NSim}, \quad (20)$$

При определении логико-лингвистического уравнения формального действия в казахской фразе, мы основываемся на гипотезе, что факт – это реальное событие, действие, которое действительно произошло или произойдет. Исходя из этого, мы определяем изъявительное наклонение глаголов и не учитываем повелительное, желательное и условное наклонения, существующие в казахском языке. Предикат γ_{VK} обозначает семантические и грамматические особенности ключевой части триплета факта, а именно Действие или Предикат факта:

$$\gamma_{VK} = (x^{-1} \vee x^{-2} \vee x^{-3})((f^{tur} \vee f^{otur} \vee f^{jatyr} \vee f^{júr}) m^{PrF} z^{Vad} \vee (y^{Oad} \vee y^{FuCo}) m^{PrFl} \vee$$
$$y^{FuCo}(m^{PrFl} \vee (m^{PrFl} f^{edi})) \vee y^{y}(f^{edi} \vee f^{eken}) \vee (y^{Vad} m^{PrFl}(p^{mic} \vee p^0)) \vee$$
$$\vee m^{PoFl}((y^{Vart} \vee y^{Vpa} \vee y^{Vpas}) \vee f^{edi}(n^{joq} \vee n^{emes} \vee n^{me} \vee n^0) \wedge$$
$$\wedge (y^{Part} \vee y^{Vad} \vee f^{otur} \vee f^{tur} \vee f^{jatyr} \vee f^{júr} \vee f^{ParP} \vee f^{UnFu}))) \quad (21)$$

На рисунке 1 показан пример реализации модели для казахского языка. В казахской фразе «*Операторлар үйде мылтық тапты*», согласно формуле (21) глагол «*тапты*» представляет действие (давно прошедшее время). Затем мы можем идентифицировать существительное «*Операторлар*» как субъект действия или субъект факта. Предикат γ_{2K} (20) идентифицирует существительное «*мылтық*», который представляет Объект ди предоставляет объект факта, называемого данной фразой.

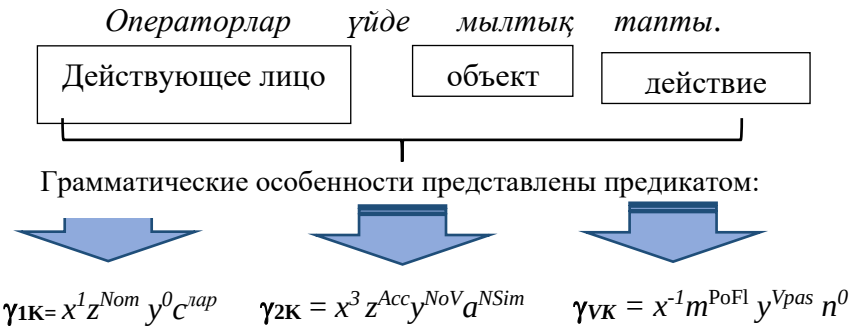

Рисунок 1. Пример идентификации факта во фразе казахского языка. Предикат γ_1 определяет грамматические особенности Субъекта действия, предикат γ_2 определяет Объект действия и γ_{VK} - это Предикат факта.

Выводы

Основным результатом этого исследования является разработка логико-лингвистической модели автоматической генерации структурированной машиночитаемой информации из мультиязычных текстов. Разработанная модель позволяет автоматизировать извлечение фактов в виде триплета "*Субъект - Предикат - Объект*" из веб-контента казахского, русского и английского языков. Наш подход базируется на гипотезе о том, что семантические роли участников действия могут явным образом определяться в поверхностной структуре предложения и, следовательно, могут быть представлены логическими отношениями грамматических и семантических характеристик слов во фразе. При этом, морфологические, синтаксические и семантические характеристики слов, влияющие на возможность слова, выражать элемент факта, зависят от конкретного языка. По этой причине мы получили различные реализации модели для английского, русского и казахского языков.

Проведенный эксперимент показал, что точность нашей модели Open IE достигает более 87% для английского корпуса, более 82% для русского корпуса и 71% для казахского корпуса.

На следующем этапе нашего исследования мы предполагаем, проанализировать особенности фактов, получаемых из корпусов текстов с помощью нашей модели, и дополнить список генерируемых атрибутов фактов.

Финансирование

Работа выполнена при поддержке исследовательского проекта, предоставленного Комитетом науки Министерства образования и науки Республики Казахстан (проект № AP05131073 - Методы, модели поиска и анализа криминальной информации, содержащейся в полуструктурированных и неструктурированных текстовых массивах.).

Литература

1. Starostin A. S., Bocharov V. V., Alexeeva S. V. et al. FactRuEval 2016: evaluation of named entity recognition and fact extraction systems for Russian. In:

Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”, pp. 702-720. 2016.

2. Duc-Thuan Vo, Bagheri, E. Open information extraction. Encyclopedia with Semantic Computing and Robotic intelligence: 2016, Vol. 1, No. 1 (pp. 1630003). World Scientific Publishing Company.

3. Fader, A., Soderland, S., Etzioni, O. Identifying relations for open information extraction. Proceedings of the conference on empirical methods in natural language processing. Edinburgh, Scotland, UK , 2011, pp. 1535-1545

4. Etzioni, O., Banko, M., Soderland, S., Weld, D. Open information extraction from the web. Communications of the ACM, 2008. Vol. 51 No. 12 (pp. 68-74). New York, NY, USA.

5. Gamallo, P., Garcia, M., Fernandez-Lanza, S. Dependency-based open information extraction. Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP Avignon, France. 2012, (pp. 10–18).

6. Akbik, A., Loser, A. KrakeN: N-ary facts in open information extraction. Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction Montreal, Canada. 2012, (pp. 52–56).

7. Gashteovski, K., Gemulla, R., Del Corro, L. MinIE: Minimizing Facts in Open Information Extraction. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) Copenhagen, Denmark, 2017, (pp. 2630-2640).

8. Angeli, G., Premkumar, M. J., D Manning.C. D. Leveraging linguistic structure for open domain information extraction. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics Beijing, China. 2015, (pp. 344-354).

9. Gamallo, P., Garcia, M. Multilingual Open Information Extraction. In Portuguese Conference on Artificial Intelligence. Coimbra, Portugal. 2015, (pp. 711-722).

10. МФ Бондаренко, ЮП Шабанов-Кушнаренко. Мозгоподобные структуры: Справочное пособие. Том первый. Под редакцией акад. НАН Украины ИВ Сергиенко. К.: Наукова думкаа, 2011. – 460 с.

11. Khairova, N.F.,Petrasova, S.,Gautam, A.P. The logical-linguistic model of fact extraction from English texts. Information and Software Technologies. Volume 639 of the series Communications in Computer and Information Science, Springer, ISBN: 978-3-319-46253-0, 2016, pp. 625-635. doi> 10.1007/978-3-319-46254-7_51

12. Khairova, N., Lewoniewski, W., Wecel, K. Estimating the Quality of Articles in Russian Wikipedia Using the Logical-Linguistic Model of Fact Extraction. Conference proceedings. BIS 2017. Part of the Lecture Notes in Business Information Processing book series (LNBIP, volume 288). (pp. 28-40). Poland: Poznan.

ИЗВЛЕЧЕНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ ИЗ НОВОСТНЫХ ИСТОЧНИКОВ НА ОСНОВЕ BI-LSTM

Чикибаева Д.Ю., Мансурова М.Е., Нугуманова А.Б., Кыргызбаева М.Е.