УДК 004.912

С.В. ПЕТРАСОВА, Н.Ф. ХАЙРОВА

ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИИ ИДЕНТИФИКАЦИИ СЕМАНТИЧЕСКИ СВЯЗНЫХ ЭЛЕМЕНТОВ ТЕКСТА ДЛЯ ОПРЕДЕЛЕНИЯ ЕДИНОГО ИНФОРМАЦИОННОГО ПРОСТРАНСТВА

Аннотация. Предложенная технология позволяет определять единое информационное пространство акторов социальных сетей за счет идентификации семантической эквивалентности коллокаций в текстах. Технология включает модель формального описания семантико-грамматических характеристик коллокатов, идентификацию коллокаций и определение предиката семантической эквивалентности двухсловных коллокаций.

Ключевые слова: семантическая связность, информационное пространство, семантико-грамматические характеристики, предикат семантической эквивалентности, коллокат, коллокация.

ВВЕДЕНИЕ

Важным аспектом формирования информационного пространства становятся социальные сети, форумы, блоги, представляющие базовые объекты современного информационного общества. Установление и развитие социальных связей в информационном обществе является объективным фактором, практически не зависящим от личных характеристик индивида. Разные виды контактов (пространственные, социальные, информационные) являются одновременно и компонентами социальных связей, и этапами их формирования.

Глобальные информационные сети стали средой и инструментом формирования информационных пространств отдельных персоналий и устойчивых социальных групп, образовавшихся на основе взаимных интересов. В общем случае информационное пространство представляет собой продукт интеллектуальной деятельности человека, объединяющей информационные ресурсы, технологии их сопровождения и использования, функционирующие на основе единых принципов, в целях удовлетворения информационной потребности пользователей [1]. При этом основной оценкой информационного социума в настоящее время становится не просто информация, а эффективная коммуникация [2], осуществляемая через установление единых информационных пространств акторов — субъектов (индивидуумов, социальных групп, организаций, институтов), совершающих действия, направленные на другие акторы. Установление таких пространств имеет реальную коммерческую и социальную ценность, например, в виде разработки рекламы для целевой аудитории.

В связи с постоянными изменениями информационного сообщества универсальность и неоднородность информационного пространства пополняется непрерывной динамичностью. Поэтому для адекватного формирования информационных пространств социальных сообществ необходимо повысить уровень автоматизации обработки текстов, в том числе за счет решения задач семантической обработки ресурсов, представляющих определенную информацию индивидуальных акторов [3]. Такой текстовой информацией, например, является персональная информация индивидуума относительно областей интересов, имеющихся контактов, востребованных темах, отмечаемых в блогах и форумах сообщениях. Определение некоторой эквивалентности и тождественности текстовой информации акторов, осуществляемое за счет подходов Natural Language Processing, позволяет выделять единые информационные пространства

определенных социальных групп, основанных на идентичности знаний, образования, возраста, престижности, богатства, расы, пола и т.д.

АНАЛИЗ ЛИТЕРАТУРНЫХ ДАННЫХ

В общем случае для решения задач семантического анализа текста используют лексико-синтаксические шаблоны; N-граммы [4]; терминологические шаблоны; индикаторы связи и профили кластеризуемости [5]; шаблоны пар объектов в сегменте текста [6]; методы опорных векторов, оснащенные языковыми ориентированными ядрами [7]; условные случайные поля [8] и др.

В то же время подходы к решению задачи выделения эквивалентных или близких по смыслу (тождественных) лингвистических элементов в тексте разнятся в зависимости от уровня таких элементов, в частности слов или словосочетаний (коллокаций). При этом если для определения синонимичности слов существует достаточное количество исследований [9–12], то задача выявления смысловой близости коллокаций, включающая идентификацию коллокаций и определение их синонимии, является достаточно нетривиальной и на сегодня не имеет эффективного решения. В данном контексте под коллокацией понимаем комбинацию двух слов, имеющих тенденцию к совместной не случайной появляемости в тексте лексической единицы с признаками синтаксической и семантической целостности.

Большинство разработанных в настоящее время методов идентификации коллокаций в тексте базируется на выявлении синтагматических отношений в естественном языке. В этом направлении существуют два основных подхода: статистический подход (window-based [13], меры ассоциации MI, PMI [14], t-scores, Chi-squared распределение [15]) и подход, основанный на анализе синтаксической структуры коллокаций [16].

На этапе определения смысловой близости словосочетаний также учитываются либо статистические закономерности, либо определяются их синтаксические характеристики. При этом часто семантическая информация (лексическая информация слов) не учитывается или дополнительно привлекаются тезаурусы. Наиболее разработанными методами определения смысловой близости словосочетаний являются выделение синонимических коллокаций в результате сравнения их переводов [17]; выявление перефразирований за счет подобия фрагментов фраз [18]; определение сходства контекста на базе анализа корпусов параллельных переводов [19].

Все перечисленные подходы работают либо на текстах достаточно узких предметных областей, либо (при статистических подходах) имеют достаточно низкую точность определения эквивалентных словосочетаний. Оба недостатка не позволяют использовать данные подходы при выделении единых информационных пространств социальных групп информационных сетей.

ПОСТАНОВКА ЗАДАЧИ ИССЛЕДОВАНИЯ

Целью настоящей статьи является разработка технологии, позволяющей использовать смысловую эквивалентность лингвистических единиц для определения семантической связности данных в текстовом информационном пространстве. При этом анализ показывает, что рассмотрение только синонимии слов недостаточно, необходимо установить формальные признаки семантических связей единиц более высокого уровня лингвистической системы — словосочетаний или коллокаций.

В данной работе предлагается технология определения семантически связных элементов текста, использующая логико-лингвистическую модель идентификации эквивалентных коллокаций [20]. Модель базируется на инструментарии компонентного анализа и аппарате алгебры конечных предикатов. Здесь рассматривают-

ся субстантивные, адъективные и глагольные типы коллокаций украинского языка. Субстантивные коллокации представлены двумя связными существительными. В адъективных коллокациях главным словом выступает существительное, а зависимым словом — прилагательное. Глагольные коллокации представлены глаголом (главный коллокат) и существительным (зависимый коллокат).

ОПИСАНИЕ ТЕХНОЛОГИИ ИДЕНТИФИКАЦИИ СЕМАНТИЧЕСКИ СВЯЗНЫХ ЭЛЕМЕНТОВ ТЕКСТА

Предлагаемая технология автоматической идентификации семантически связных данных включает следующие этапы: 1) выделение семантико-грамматических характеристик коллокатов — слов, которые потенциально могут являться элементами субстантивных, адъективных и глагольных словосочетаний; 2) идентификация коллокаций — словосочетаний, образованных двумя рядом стоящими словоформами; 3) определение синонимичных коллокатов — слов, близких по смыслу, образующих словосочетания; 4) идентификация семантической эквивалентности двухсловных коллокаций — словосочетаний, имеющих общие элементы смысла.

Выделение семантико-грамматических характеристик коллокатов. На первом этапе выразим семантико-грамматические характеристики потенциальных коллокатов в виде парадигматической таблицы (табл. 1), связывающей характеристики слов с возможной их ролью в субстантивных, адъективных и глагольных словосочетаниях. Здесь x определяет главное, а y — зависимое слова словосочетаний, где тип коллокации $x_1 y_1$ — субстантивный, $x_2 y_2$ — адъективный, $x_3 y_3$ — глагольный, c определяет семантический характер, a — грамматический характер.

Для описания семантических и грамматических отношений вводятся предметные переменные a_1, a_2, a_3, c:

- предметная переменная a_1 определяет часть речи (N — существительное, A — прилагательное, V — глагол);
- предметная переменная a_2 определяет падеж существительных N и прилагательных A (*Nom* — именительный падеж, *Gen* — родительный падеж, *Acc* — винительный падеж, *Dat* — дательный падеж, *In* — творительный падеж, *Prt* — предложный падеж);
- предметная переменная a_3 определяет возвратность глагола V (*Ref* — возвратный глагол, *NonRef* — невозвратный глагол);
- предметная переменная c определяет возможные семантические роли словсуществительных N; значения c представлены в первом столбце таблицы (*Ag* — агенс, *Att* — атрибут, *Pac* — пациенс, *Adr* — адресат, *Ins* — инструмент, *M* — место).

Таблица 1

c	a	x_1	y_1	y_2	x_2	x_3		y_3
		N	N	A	N	V		N
						Ref	*NonRef*	
Ag	*Nom*	q^1		q^{12}	q^{18}	q^{24}	q^{25}	
Att	*Gen*	q^2	q^7	q^{13}	q^{19}			q^{26}
Pac	*Acc*	q^3	q^8	q^{14}	q^{20}			q^{27}
Adr	*Dat*	q^4	q^9	q^{15}	q^{21}			q^{28}
Ins	*In*	q^5	q^{10}	q^{16}	q^{22}			q^{29}
M	*Prt*	q^6	q^{11}	q^{17}	q^{23}			q^{30}

Формальными номерами ячеек $q = \overline{1, 30}$ парадигматической таблицы обозначаются возможные согласованные значения грамматических и семантических характеристик слов (потенциальных коллокатов):

$$q^1 = a_1^N a_2^{Nom} c^{Ag};\ q^2 = a_1^N a_2^{Gen} c^{Att};\ q^3 = a_1^N a_2^{Acc} c^{Pac};\ q^4 = a_1^N a_2^{Dat} c^{Adr};$$

$$q^5 = a_1^N a_2^{In} c^{Ins};\ q^6 = a_1^N a_2^{Prt} c^M;\ q^7 = a_1^N a_2^{Gen} c^{Att};\ q^8 = a_1^N a_2^{Acc} c^{Pac};$$

$$q^9 = a_1^N a_2^{Dat} c^{Adr};\ q^{10} = a_1^N a_2^{In} c^{Ins};\ q^{11} = a_1^N a_2^{Prt} c^M;\ q^{12} = a_1^A a_2^{Nom};$$

$$q^{13} = a_1^A a_2^{Gen};\ q^{14} = a_1^A a_2^{Acc};\ q^{15} = a_1^A a_2^{Dat};\ q^{16} = a_1^N a_2^{In};\ q^{17} = a_1^A a_2^{Prt};$$

$$q^{18} = a_1^N a_2^{Nom} c^{Ag};\ q^{19} = a_1^N a_2^{Gen} c^{Att};\ q^{20} = a_1^N a_2^{Acc} c^{Pac}; \quad (1)$$

$$q^{21} = a_1^N a_2^{Dat} c^{Adr};\ q^{22} = a_1^N a_2^{In} c^{Ins};\ q^{23} = a_1^N a_2^{Prt} c^M;\ q^{24} = a_1^V a_3^{Ref};$$

$$q^{25} = a_1^V a_3^{NonRef};\ q^{26} = a_1^N a_2^{Gen} c^{Att};\ q^{27} = a_1^N a_2^{Acc} c^{Pac};$$

$$q^{28} = a_1^N a_2^{Dat} c^{Adr};\ q^{29} = a_1^N a_2^{In} c^{Ins};\ q^{30} = a_1^N a_2^{Prt} c^M.$$

Выполняя операцию почленной конъюнкции, можно выявить повторные элементы множества q:

$$a_1^N a_2^{Nom} c^{Ag} = q^1 \wedge q^{18};\ a_1^N a_2^{Gen} c^{Att} = q^2 \wedge q^7 \wedge q^{19} \wedge q^{26};$$

$$a_1^N a_2^{Acc} c^{Pac} = q^3 \wedge q^8 \wedge q^{20} \wedge q^{27};\ a_1^N a_2^{Dat} c^{Adr} = q^4 \wedge q^9 \wedge q^{21} \wedge q^{28};$$

$$a_1^N a_2^{In} c^{Ins} = q^5 \wedge q^{10} \wedge q^{22} \wedge q^{29};\ a_1^N a_2^{Prt} c^M = q^6 \wedge q^{11} \wedge q^{23} \wedge q^{30};$$

$$a_1^A a_2^{Nom} = q^{12};\ a_1^A a_2^{Gen} = q^{13};\ a_1^A a_2^{Acc} = q^{14};\ a_1^A a_2^{Dat} = q^{15};$$

$$a_1^A a_2^{In} = q^{16};\ a_1^A a_2^{Prt} = q^{17};\ a_1^V a_3^{Ref} = q^{24};\ a_1^V a_3^{NonRef} = q^{25}.$$

Упрощая множество уравнений (1), переопределяем переменную q, задействуя переменную r:

$$r^1 = q^1 \wedge q^{18};\ r^2 = q^2 \wedge q^7 \wedge q^{19} \wedge q^{26};\ r^3 = q^3 \wedge q^8 \wedge q^{20} \wedge q^{27};$$

$$r^4 = q^4 \wedge q^9 \wedge q^{21} \wedge q^{28};\ r^5 = q^5 \wedge q^{10} \wedge q^{22} \wedge q^{29};$$

$$r^6 = q^6 \wedge q^{11} \wedge q^{23} \wedge q^{30};\ r^7 = q^{12};\ r^8 = q^{13};$$

$$r^9 = q^{14};\ r^{10} = q^{15};\ r^{11} = q^{16};\ r^{12} = q^{17};\ r^{13} = q^{24};\ r^{14} = q^{25}.$$

Тогда парадигматическую табл. 1 можно переписать в упрощенном нормализованном виде (табл. 2). Перепишем систему уравнений (1) с учетом зависимости переменной r от предметных переменных, выражающих семантико-грамматические характеристики a_1, a_2, a_3, c:

$$r^1 = a_1^N a_2^{Nom} c^{Ag};\ r^2 = a_1^N a_2^{Gen} c^{Att};\ r^3 = a_1^N a_2^{Acc} c^{Pac};\ r^4 = a_1^N a_2^{Dat} c^{Adr};$$

$$r^5 = a_1^N a_2^{In} c^{Ins};\ r^6 = a_1^N a_2^{Prt} c^M;\ r^7 = a_1^A a_2^{Nom};\ r^8 = a_1^A a_2^{Gen};\ r^9 = a_1^A a_2^{Acc},$$

$$r^{10} = a_1^A a_2^{Dat};\ r^{11} = a_1^A a_2^{In};\ r^{12} = a_1^A a_2^{Prt};\ r^{13} = a_1^V a_3^{Ref};\ r^{14} = a_1^V a_3^{NonRef}.$$

Таблица 2

c	a	x_1	y_1	y_2	x_2	x_3		y_3
		N	N	A	N	V		N
						Ref	*NonRef*	
Ag	*Nom*	r^1		r^7	r^1	r^{13}	r^{14}	
Att	*Gen*	r^2	r^2	r^8	r^2			r^2
Pac	*Acc*	r^3	r^3	r^9	r^3			r^3
Adr	*Dat*	r^4	r^4	r^{10}	r^4			r^4
Ins	*In*	r^5	r^5	r^{11}	r^5			r^5
M	*Prt*	r^6	r^6	r^{12}	r^6			r^6

Вводимое бинарное отношение P позволяет связать переменную r с предметными переменными a_1, a_2, a_3, c.

Бинарный предикат P_1 связывает переменную r с предметной переменной a_1, определяющей грамматическую характеристику (часть речи):

$$P_1(a_1, r) = a_1^N (r^1 \vee r^2 \vee r^3 \vee r^4 \vee r^5 \vee r^6) \vee$$

$$\vee\, a_1^A (r^7 \vee r^8 \vee r^9 \vee r^{10} \vee r^{11} \vee r^{12}) \vee a_1^V (r^{13} \vee r^{14}).$$

Бинарный предикат P_2 связывает переменную r с предметной переменной a_2, определяющей грамматическую характеристику (грамматический падеж):

$$P_2(a_2, r) = a_2^{Nom} (r^1 \vee r^7) \vee a_2^{Gen} (r^2 \vee r^8) \vee a_2^{Acc} (r^3 \vee r^9) \vee$$

$$\vee\, a_2^{Dat} (r^4 \vee r^{10}) \vee a_2^{In} (r^5 \vee r^{11}) \vee a_2^{Prt} (r^6 \vee r^{12}).$$

Бинарный предикат P_3 связывает переменную r с предметной переменной a_3, определяющей грамматическую характеристику (возвратность глагола):

$$P_3(a_3, r) = a_3^{Ref} r^{13} \vee a_3^{NonRef} r^{14}.$$

Бинарный предикат P_4 связывает переменную r с предметной переменной c, определяющей семантическую характеристику (семантическую роль):

$$P_4(c, r) = c^{Ag} r^1 \vee c^{Att} r^2 \vee c^{Pac} r^3 \vee c^{Adr} r^4 \vee c^{Ins} r^5 \vee c^M r^6.$$

Таким образом, можем ввести предикат идентификации потенциальных коллокатов, который характеризуется системой бинарных отношений P_1–P_4:

$$P(a_1, a_2, a_3, c, r) = P_1(a_1, r) \wedge P_2(a_2, r) \wedge P_3(a_3, r) \wedge P_4(c, r) =$$

$$= a_1^N a_2^{Nom} c^{Ag} \vee a_1^N a_2^{Gen} c^{Att} \vee a_1^N a_2^{Acc} c^{Pac} \vee a_1^N a_2^{Dat} c^{Adr} \vee$$

$$\vee\, a_1^N a_2^{In} c^{Ins} \vee a_1^N a_2^{Prt} c^M \vee a_1^A a_2^{Nom} \vee a_1^A a_2^{Gen} \vee a_1^A a_2^{Acc} \vee$$

$$\vee\, a_1^A a_2^{Dat} \vee a_1^A a_2^{In} \vee a_1^A a_2^{Prt} \vee a_1^V a_3^{Ref} \vee a_1^V a_3^{NonRef}.$$

Идентификация коллокаций. Строим парадигматическую таблицу (табл. 3), в которой жирным шрифтом выделяем семантико-грамматические характеристики рядом стоящих слов, образующих коллокации.

Таблица 3

c	a	x_1	y_1	y_2	x_2	x_3		y_3
		N	N	A	N	V		N
						Ref	*NonRef*	
Ag	*Nom*	$\boldsymbol{r_x^1}$		$\boldsymbol{r_y^7}$	$\boldsymbol{r_x^1}$			
Att	*Gen*	$\boldsymbol{r_x^2}$	$\boldsymbol{r_y^2}$	$\boldsymbol{r_y^8}$	$\boldsymbol{r_x^2}$			r_y^2
Pac	*Acc*	$\boldsymbol{r_x^3}$	r_y^3	$\boldsymbol{r_y^9}$	$\boldsymbol{r_x^3}$	r_x^{13}	$\boldsymbol{r_x^{14}}$	$\boldsymbol{r_y^3}$
Adr	*Dat*	$\boldsymbol{r_x^4}$	r_y^4	$\boldsymbol{r_y^{10}}$	$\boldsymbol{r_x^4}$			r_y^4
Ins	*In*	$\boldsymbol{r_x^5}$	r_y^5	$\boldsymbol{r_y^{11}}$	$\boldsymbol{r_x^5}$			r_y^5
M	*Prt*	$\boldsymbol{r_x^6}$	r_y^6	$\boldsymbol{r_y^{12}}$	$\boldsymbol{r_x^6}$			r_y^6

Например, x_1 с набором семантико-грамматических характеристик $\{(a_1^N a_2^{Nom} c^{Ag}),\ (a_1^N a_2^{Gen} c^{Att}),\ (a_1^N a_2^{Acc} c^{Pac}),\ (a_1^N a_2^{Dat} c^{Adr}),\ (a_1^N a_2^{In} c^{Ins}),\ (a_1^N a_2^{Prt} c^M)\}$ образует коллокацию с зависимым словом y_1, обладающим набором семантико-грамматических характеристик $\{a_1^N a_2^{Gen} c^{Att}\}$.

Выражаем зависимость переменной r от предметных переменных x, y, обозначающих главный и зависимый коллокаты:

— для субстантивных коллокаций $(N_x N_y)$

$$(r_x^1 \vee r_x^2 \vee r_x^3 \vee r_x^4 \vee r_x^5 \vee r_x^6) r_y^2 =$$

$$= (x^{NNomAg} \vee x^{NGenAtt} \vee x^{NAccPac} \vee x^{NDatAdr} \vee x^{NInIns} \vee x^{NPrtM})\, y^{NGenAtt};$$

— для адъективных коллокаций $(A_y N_x)$

$$r_y^7 r_x^1 = y^{ANom} x^{NNomAg};\ r_y^8 r_x^2 = y^{AGen} x^{NGenAtt};\ r_y^9 r_x^3 = y^{AAcc} x^{NAccPac};$$

$$r_y^{10} r_x^4 = y^{ADat} x^{NDatAdr};\ r_y^{11} r_x^5 = y^{AIn} x^{NInIns};\ r_y^{12} r_x^6 = y^{APrt} x^{NPrtM};$$

— для глагольных коллокаций $(V_x N_y)$

$$r_x^{14} r_y^3 = x^{VNonRef}\, y^{NAccPac}.$$

Вводимое бинарное отношение P позволяет связать переменную r с предметными переменными x и y:

$$P_5(r_x, r_y) = (r_x^1 \vee r_x^2 \vee r_x^3 \vee r_x^4 \vee r_x^5 \vee r_x^6) r_y^2, \tag{2}$$

$$P_6(r_y, r_x) = r_y^7 r_x^1 \vee r_y^8 r_x^2 \vee r_y^9 r_x^3 \vee r_y^{10} r_x^4 \vee r_y^{11} r_x^5 \vee r_y^{12} r_x^6, \tag{3}$$

$$P_7(r_x, r_y) = r_x^{14} r_y^3, \tag{4}$$

где уравнение (2) определяет субстантивные коллокации, уравнение (3) — адъективные коллокации и уравнение (4) — глагольные коллокации.

Введем предикат идентификации коллокаций $P(x, y)$, который характеризуется системой бинарных отношений P_5–P_7:

$$P(x, y) = P_5(r_x, r_y) \wedge P_6(r_y, r_x) \wedge P_7(r_x, r_y) = (x^{NNomAg} \vee x^{NGenAtt} \vee x^{NAccPac} \vee$$

$$\vee x^{NDatAdr} \vee x^{NInIns} \vee x^{NPrtM})\, y^{NGenAtt} \vee y^{ANom} x^{NNomAg} \vee y^{AGen} x^{NGenAtt} \vee$$

$$\vee y^{AAcc} x^{NAccPac} \vee y^{ADat} x^{NDatAdr} \vee y^{AIn} x^{NInIns} \vee y^{APrt} x^{NPrtM} \vee x^{VNonRef}\, y^{NAccPac}.$$

Таблица 4

x_i \ y_i	${}_a A\ Nom$	${}_a A\ Gen$	${}_a A\ Acc$	${}_a A\ Dat$	${}_a A\ In$	${}_a A\ Prt$	${}_a\ NGen\ {}_c Att$	${}_a\ NAcc\ {}_c Pac$	${}_a\ NDat\ {}_c Adr$	${}_a\ NIn\ {}_c Ins$	${}_a\ NPrt\ {}_c M$	${}_a V\ Ref$	${}_a V\ Non\ Ref$
${}_a NNom_c Ag$	1	0	0	0	0	0	1	0	0	0	0	0	0
${}_a NGen_c Att$	0	1	0	0	0	0	1	0	0	0	0	0	0
${}_a NAcc_c Pac$	0	0	1	0	0	0	1	0	0	0	0	0	0
${}_a NDat_c Adr$	0	0	0	1	0	0	1	0	0	0	0	0	0
${}_a NIn_c Ins$	0	0	0	0	1	0	1	0	0	0	0	0	0
${}_a NPrt_c M$	0	0	0	0	0	1	1	0	0	0	0	0	0
${}_a V\ Ref$	0	0	0	0	0	0	0	0	0	0	0	0	0
${}_a V\ Non Ref$	0	0	0	0	0	0	0	1	0	0	0	0	0

Предикат $P(x, y) = 1$, если семантико-грамматические характеристики двух рядом стоящих словоформ множества $M = \{m_1, \ldots, m_n\}$ позволяют создать словосочетание, и $P(x, y) = 0$ в противном случае (табл. 4).

Определение синонимичных коллокатов. На следующем этапе для установления синонимии между коллокатами используется метод автоматической идентификации семантических корреляций толерантности и эквивалентности, детально описанный в работе [21]. Этот метод для определения семантически связных данных использует меру семантической близости $f(t', t'')$ между двумя языковыми единицами t' и t''. Мера семантической близости выражается отношением теоретико-множественного пересечения и объединения множеств терминов дефиниций глоссария.

Например, чтобы вычислить меру синонимии (или меру семантической близости) f для терминов t_1 = «авторизація», t_2 = «аутентифікація» и t_3 = «ідентифікація» в глоссарии [22], определяется пересечение и объединение множеств слов каждой дефиниции:

$$f(t', t'') = \frac{2 \times N(d_1 \bigcap d_2)}{N(d_1 \bigcup d_2)},$$

где $f(t', t'')$ — величина семантической близости между терминами t' и t''; d_1, d_2 — дефиниции лингвистических единиц толкового словаря t' и t''; $N(x_1 \bigcap x_2)$ — количество общих слов в определениях терминов t' и t''; $N(x_1 \bigcup x_2)$ — количество всех слов в определениях терминов t' и t''.

Результат определения меры семантической эквивалентности между t_1, t_2, t_3: $f(t_1, t_2) = 0{,}4$; $f(t_2, t_3) = 0{,}45$; $f(t_1, t_3) = 0{,}39$.

В работе [21] доказывается, что при значении коэффициента семантической близости больше 0,35 слова t_1 = «*авторизація*», t_2 = «*аутентифікація*» и t_3 = «*ідентифікація*» считаются связанными отношением эквивалентности.

Идентификация семантической эквивалентности двухсловных коллокаций. Синонимичные слова могут образовывать близкие по смыслу словосочетания, например «*зберігати дані*» ≈ «*містити відомості*», и при этом могут формировать несвязные по смыслу словосочетания, например «*зберігання даних*» ≠ «*інформація репозитарію*».

Для выделения семантически связных коллокаций используется логико-лингвистическая модель [20]. Введем предикат семантической эквивалентности двухсловных коллокаций

$$P(x_1, y_1) * P(x_2, y_2) = \gamma_i(x_1, y_1, x_2, y_2) \wedge P(x_1, y_1) \wedge P(x_2, y_2),$$

где символ * обозначает операцию определения смысловой близости, знак $\wedge$ определяет конъюнкцию, предикат $\gamma_{i(x_1}, y_1, x_2, y_2)$ исключает коллокации,

между которыми не может быть установлена смысловая эквивалентность.

Предикат $\gamma_1(x_1, y_1, x_2, y_2) = x_1^{VNonRef}\, y_1^{NAccPac} x_2^{VNonRef}\, y_2^{NAccPac}$ показывает семантическую близость глагольных коллокаций ($V_x N_y$), например

$$\textbf{визначати}\ x_1^{VNonRef}\ \textbf{відомості}\ y_1^{NAccPac} \approx \textbf{встановлювати}\ x_2^{VNonRef}\ \textbf{дані}\ y_2^{NAccPac}.$$

Предикат $\gamma_2(x_1, y_1, x_2, y_2) = x_1^{NNomAg}\, y_1^{NGenAtt} x_2^{NNomAg}\, y_2^{NGenAtt}$ показывает семантическую близость субстантивных коллокаций ($N_x N_y$), таких как

$$\textbf{швидкість}\ x_1^{NNomAg}\ \textbf{передачі}\ y_1^{NGenAtt} \approx \textbf{темп}\ x_2^{NNomAg}\ \textbf{відправлення}\ y_2^{NGenAtt}.$$

Предикат $\gamma_3(x_1, y_1, x_2, y_2) = y_1^{ANom} x_1^{NNomAg}\, y_2^{ANom} x_2^{NNomAg}$ показывает семантическую близость между адъективными коллокациями ($A_y N_x$), например

$$\textbf{булева}\ y_1^{ANom}\ \textbf{операція}\ x_1^{NNomAg} \approx \textbf{логічна}\ y_2^{ANom}\ \textbf{процедура}\ x_2^{NNomAg}.$$

Таким образом, предикат семантической эквивалентности коллокаций, состоящих из выявленных на предыдущих этапах попарно синонимичных коллокатов, имеет вид

$$\begin{aligned}\gamma(x_1, y_1, x_2, y_2) = (\, & y_1^{ANom} x_1^{NNomAg} \vee\ y_1^{AGen} x_1^{NGenAtt}\, y_1^{AAcc} x_1^{NAccPac} \vee\ y_1^{ADat} x_1^{NDatAdr} \vee \\ \vee\ & y_1^{AIn} x_1^{AInIns} \vee\ y_1^{APrt} x_1^{NPrtM}\,)(\, y_2^{ANom} x_2^{NNomAg} \vee\ y_2^{AGen} x_2^{NGenAtt} \vee\ y_2^{AAcc} x_2^{NAccPac} \vee \\ \vee\ & y_2^{ADat} x_2^{NDatAdr} \vee\ y_2^{AIn} x_2^{NInIns} \vee\ y_2^{APrt} x_2^{NPrtM}\,) \vee (x_1^{NNomAg} \vee\ x_1^{NGenAtt} \vee \\ \vee\ & x_1^{NAccPac} \vee\ x_1^{NDatAdr} \vee\ x_1^{NInIns} \vee\ x_1^{NPrtM}\,) y_2^{NGenAtt} (x_2^{NNomAg} \vee\ x_2^{NGenAtt} \vee\ x_2^{NAccPac} \vee \\ \vee\ & x_2^{NDatAdr} \vee\ x_2^{NInIns} \vee\ x_2^{NPrtM}\,) y_2^{NGenAtt} \vee\ x_1^{VNonRef}\, y_1^{NAccPac} x_2^{VNonRef}\, y_2^{NAccPac}. \qquad (5)\end{aligned}$$

Если предикат $\gamma(x_1, y_1, x_2, y_2) = 1$, то слова с соответствующими характеристиками образуют два эквивалентных по смыслу словосочетания. В противном случае рассматриваемые словосочетания не эквивалентны по смыслу.

Следовательно, коллокации могут считаться близкими по смыслу, если:

— главное слово x_1 в первой коллокации определено как синонимичное главному слову x_2 во второй коллокации ($x_1 \approx x_2$), а зависимое слово y_1 в первой коллокации синонимично зависимому слову y_2 во второй коллокации ($y_1 \approx y_2$);

— семантико-грамматические характеристики коллокатов словосочетаний ($x_1 y_1$) и ($x_2 y_2$) удовлетворяют предикату семантической эквивалентности (5).

Например, коллокации $coll_1$ = «*процес аутентифікації*», $coll_2$ = «*процедура ідентифікації*» имеют семантико-грамматические характеристики $coll_1 = x_1^{NNomAg}\, y_1^{NGenAtt}$, $coll_2 = x_2^{NNomAg}\, y_2^{NGenAtt}$ (определены на первом и втором этапах технологии); между главными коллокатами и зависимыми коллокатами словосочетаний установлены отношения семантической эквивалентности $x_1 \approx x_2$, $y_1 \approx y_2$ (определены на третьем этапе).

В результате предикат $\gamma_2(x_1, y_1, x_{2,} y_{2)}$ показывает связные по смыслу коллокации (четвертый этап предложенной технологии):

$$\textbf{процес}\ x_1^{NNomAg}\ \textbf{аутентифікації}\ y_1^{NGenAtt} \approx$$

$$\approx \textbf{процедура}\ x_2^{NNomAg}\ \textbf{ідентифікації}\ y_2^{NGenAtt}.$$

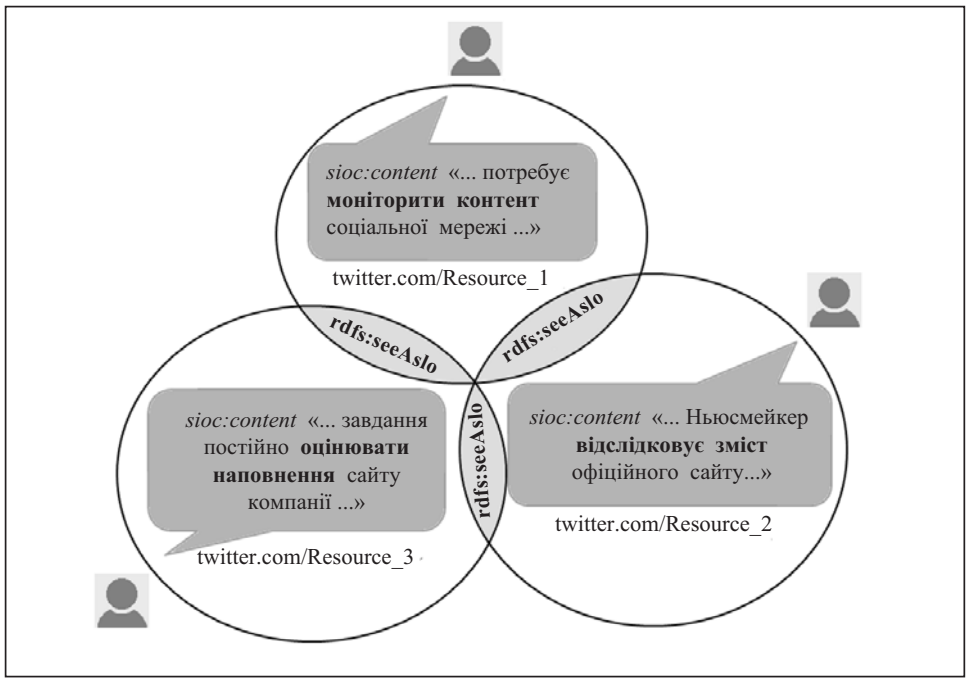

Рис. 1. Пример семантически связного контента в информационном пространстве Twitter

ПЕРСПЕКТИВНЫЕ ВАРИАНТЫ ИСПОЛЬЗОВАНИЯ РАЗРАБОТАННОЙ ТЕХНОЛОГИИ

Разработанная технология идентификации семантически связных элементов текста может быть использована различными инструментариями социальных сетей, блогов, форумов. Например, встраивание разработанной технологии в семантический инструментарий SIOC (Semantically Interlinked Online Communities) [23], описывающий метаданные на платформе RDF, позволяет использовать элементы разметки Twitter для определения единого информационного пространства социальной сети.

В рассматриваемом фрагменте связных твитов (рис. 1) элементы content пространства имен sioc первого сообщения содержат текст *sioc:content* «… *потребує* ***моніторити контент*** *соціальної мережі* …»; второго сообщения — текст *sioc:content* «… *Ньюсмейкер* ***відслідковує зміст*** *офіційного сайту* …»; третьего сообщения — текст *sioc:content* «… *завдання постійно* ***оцінювати наповнення*** *сайту компанії* …».

Использование технологии позволяет определить эквивалентный смысл. Предикат $\gamma_1(x_1, y_1, x_2, y_2)$ показывает семантическую близость ***глагольных*** коллокаций: **моніторити** $x_1^{V\ NonRef}$ **контент** $y_1^{N\ Acc\ Pac} \approx$ **відслідковувати** $x_2^{V\ NonRef}$ **зміст** $y_2^{N\ Acc\ Pac} \approx$ **оцінювати** $x_3^{V\ NonRef}$ **наповнення** $y_3^{N\ Acc\ Pac}$, при этом $x_1 \approx x_2 \approx x_3$, $y_1 \approx y_2 \approx y_3$, принадлежащих различным твитам. Наличие нескольких подобных синонимичных элементов может стать дополнительным условием выделения единого информационного пространства.

ЗАКЛЮЧЕНИЕ

Предложенная технология идентификации семантически связных элементов текста позволяет определить единое информационное пространство акторов социальных сетей. Использование данной технологии во взаимодействии со статистическими методами обработки позволит эффективно определять близ-

кие по смыслу фрагменты текстов в информационно-поисковых, экспертных, аналитических информационных системах широкого назначения.

СПИСОК ЛИТЕРАТУРЫ

1. Додонов А.Г., Ландэ Д.В., Путятин В.Г. Компьютерные сети и аналитические исследования. Киев: ИПРИ НАН Украины, 2014. 486 с.
2. Кастельс М. Информационная эпоха: экономика, общество и культура. Москва: ГУ-ВШЭ, 2000. 606 с.
3. Хайрова Н.Ф., Петрасова С.В. Информационные интеллектуальные системы и семантический веб: учебное пособие. Харьков: НТУ «ХПИ», 2015. 169 с.
4. Arefyev N.V., Panchenko A.I., Lukanin A.V. et al. Evaluating three corpus-based semantic similarity systems for Russian. *Компьютерная лингвистика и интеллектуальные технологии*: по материалам ежегодной международной конференции «Диалог» (Москва, 27–30 мая 2015 г.). № 14(21): В 2 т. Т. 2: Доклады специальных секций. Москва: Изд-во РГГУ, 2015. С. 106–119.
5. Саломатина Н.В., Гусев В.Д., Ильина Л.Ю. О возможностях автоматизации выявления связей между терминами предметной области (на примере катализа). *Компьютерная лингвистика и интеллектуальные технологии*: по материалам ежегодной международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). № 9(16). Москва: Изд-во РГГУ, 2010. С. 430–436.
6. Hasegawa T., Sekine S., Grishman R. Discovering relations among named entities from large corpora. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (ACL ‘04). Stroudsburg, PA, USA, 2004. P. 415–422.
7. Bunescu R., Mooney R. Learning to extract relations from the web using minimal supervision. *Proceedings of the 45th Annual Meeting on Association for Computational Linguistics* (ACL ‘07). Prague, Czech Republic, 2007. P. 576–583.
8. Culotta A., McCallum A., Betz J. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. *Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. New York, 2006. P. 296–303.
9. Nakov S. Automatic acquisition of synonyms using the web as a corpus. *Proceedings of the 3rd Annual South-East European Doctoral Student Conference.* 2008. Vol. 2. P. 216–229.
10. Hua Wu, Ming Zhou. Optimizing synonym extraction using monolingual and bilingual Re-sources. *Proceedings of the Second International Workshop on Para-phrasing* (PARAPHRASE ‘03). Stroudsburg, PA, USA, 2003. Vol. 16. P. 72–79.
11. Мисуно И.С., Рачковский Д.А., Слипченко С.В. Векторные и распределенные представления, отражающие меру семантической связи слов. *Математические машины и системы*. 2005. № 3. С. 50–66.
12. Митрофанова О.А. Семантические расстояния: проблемы и перспективы. Материалы XXXIV междунар. филол. конф. СПбГУ, 2005. С. 59–63.
13. Church K.W., Hanks P. Word association norms, mutual information, and lexicography. *Computational Linguistics*. 1990. Vol. 16, Iss. 1. P. 22–29.
14. Evert S., Krenn B. Methods for the qualitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (ACL ‘01). Stroudsburg, PA, USA, 2001. P. 188–195.
15. Захаров В.П., Хохлова М.В. Выделение терминологических словосочетаний из специальных текстов на основе различных мер ассоциации. Интернет и современное общество «IMS-2014». С.-Петербург: Университет ИТМО, 2014. С. 290–293.
16. Akinina Y.S., Kuznetsov I.O., Toldova S.Y. The impact of syntactic structure on verb-noun collocation extraction. *Компьютерная лингвистика и интеллектуальные технологии*: по материалам ежегодной международной конференции «Диалог» (Бекасово, 29 мая–2 июня 2013 г.). № 12(19): В 2 т. Т. 1: Основная программа конференции. Москва: Изд-во РГГУ, 2013. С. 2–17.
17. Hua Wu, Ming Zhou. Synonymous collocation extraction using translation information. *Proceedings of the 41th Annual Meeting on Association for Computational Linguistics (*ACL ‘03). Stroudsburg, PA, USA, 2003. Vol. 1. P. 120–127.

18. Marius P., Péter D. Aligning needles in a haystack: Paraphrase acquisition across the web. *Proceedings of the Second International Joint Conference: Natural Language Processing* (IJCNLP 2005). Jeju Island, Korea, 2005. P. 119–130.

19. Barzilay R., McKeown Kathleen R. Extracting paraphrases from a parallel corpus. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (ACL'01). Stroudsburg, PA, USA, 2001. P. 50–57.

20. Khairova N., Petrasova S., Gautam A.P.S. The logical and linguistic model for automatic extraction of collocation similarity. *Econtechmod: An International Quarterly Journal on Economics in Technology, New Technologies and Modelling Processes*. Lublin; Rzeszow, 2015. Vol. 4, N 4. P. 43–48.

21. Хайрова Н.Ф., Петрасова С.В., Ленков С.В. Метод автоматической идентификации семантических корреляций терминов глоссария. *Збірник наук. праць Військового ін-ту Київ. нац. ун-ту ім. Тараса Шевченка*. 2014. Вип. 46. С. 222–228.

22. Півняк Г.Г., Бусигін Б.С., Дівізінюк М.М. та ін. Тлумачний словник з інформатики. Донецьк: Нац. гірнич. ун-т, 2010. 600 с.

23. Breslin J.G., Harth A., Bojars U., Decker S. Towards semantically-interlinked online communities. *Proceedings of the Second European Conference on the Semantic Web: Research and Applications*. Berlin; Heidelberg: Springer-Verlag, 2005. P. 500–514.

Надійшла до редакції 15.06.2016

С.В. Петрасова, Н.Ф. Хайрова

ВИКОРИСТАННЯ ТЕХНОЛОГІЇ ІДЕНТИФІКАЦІЇ СЕМАНТИЧНО ЗВ'ЯЗНИХ ЕЛЕМЕНТІВ ТЕКСТУ ДЛЯ ВИЗНАЧЕННЯ ЄДИНОГО ІНФОРМАЦІЙНОГО ПРОСТОРУ

Анотація. Запропонована технологія дозволяє визначати єдиний інформаційний простір акторів соціальних мереж за рахунок ідентифікації семантичної еквівалентності колокацій у текстах. Технологія включає модель формального опису семантико-граматичних характеристик колокатів, ідентифікацію колокацій та визначення предиката семантичної еквівалентності двослівних колокацій.

Ключові слова: семантична зв'язність, інформаційний простір, семантико-граматичні характеристики, предикат семантичної эквівалентності, колокат, колокація.

S.V. Petrasova, N.F. Khairova

USING SEMANTICALLY SIMILAR TEXT ELEMENTS IDENTIFICATION TECHNOLOGY TO DETERMINE A COMMON INFORMATION SPACE

Abstract. The proposed technology allows determining a common information space of social network actors by identifying the semantic equivalence of collocations in texts. The technology includes the model of formal description of the semantic and grammatical characteristics of collocates, identification of collocations, and determination of a semantic equivalence predicate of two-word collocations.

Keywords: semantic similarity, information space, semantic and grammatical characteristics, semantic equivalence predicate, collocate, collocation.

Петрасова Светлана Валентиновна,
аспирантка Национального технического университета «Харьковский политехнический институт», e-mail: svetapetrasova@gmail.com.

Хайрова Нина Феликсовна,
доктор техн. наук, профессор Национального технического университета «Харьковский политехнический институт», e-mail: nina_khajrova@yahoo.com.