рено з використанням алгоритму Quick Boruvka з системи Concorde [2]. За незначний обчислювальний час (тривалість обчислень 176 – 352 с) досягнуто покращення на 0.01717 – 0.39125 %. Максимально допустиму різницю між довжиною пар ребер до і після оптимізації вибрано з розрахунком на включення 10 000 – 12 000 точок у "кільце".

Висновки

Застосування розробленого алгоритму дає змогу покращити якість розв'язку для задачі комівояжера. Обміном двома парами ребер чи більшою кількістю можливо зменшити довжину шляху, де оптимізація в локальній області не може вплинути на заміну фрагментів, які розміщені в різних частинах робочої зони і не входять в зону локальної оптимізації.

1. Bazylevych R., Kutelmakh R., Prasad B., Bazylevych L. Decomposition and Scanning Optimization algorithms for TSP // Proceedings of the International Conference on Theoretical and Mathematical Foundations of Computer Science, Orlando, 2008, pp. 110-116. 2. Concorde: http://www.tsp.gatech.edu/concorde.html. 3. Delaunay B. Sur la sphère vide // Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk, No 7, 1934, pp. 793–800. 4. DIMACS TSP Challenge Results: http://www.akira.ruc.dk/~keld/research/LKH/DIMACS_results.html. 5. Helsgaun K. An Effective Implementation of k-Opt Moves for the Lin–Kernighan TSP Heuristic // Datalogiske Skrifter, Writings on Computer Science, No. 109, Roskilde University, 2006. 6. TSP Art Instances: http://www.tsp.gatech.edu/data/art/index.html. 7. Базилевич Р. П., Кутельмах Р. К., Кузь Б. О. Алгоритм розв'язання задачі комівояжера великої розмірності методом "тора" // Вісник Нац. ун-ту "Львівська політехніка". – 2010. – № 686. – С. 179–182.

УДК 811.161.2

О. Дмитраш, А. Романюк, П. Тимощук
Національний університет "Львівська політехніка",
кафедра систем автоматизованого проектування

АНОТУВАННЯ ТЕКСТІВ ДЛЯ ЗДІЙСНЕННЯ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ ІМЕНОВАНИХ СУТНОСТЕЙ

Описано створення анотованого корпусу іменованих сутностей для української мови.

Ключові слова: видобування інформації, розпізнавання іменованих сутностей, корпус.

This paper describes the process of creating an annotated corpus of named entities for Ukrainian language.

Key words: information extraction, named entity recognition, corpus.

Постановка проблеми

Сьогодні, у період постійного зростання обсягів інформації, все гострішою стає необхідність створення ефективних підходів і систем опрацювання неструктурованої інформації та перетворення їх на структуровані дані. Створення та функціонування таких систем безпосередньо залежить від систем видобування інформації.

Видобування інформації – це одне із завдань обробки природної мови, яке полягає в автоматичному видобуванні структурованих даних із неструктурованих або напівструктурованих інформаційних джерел. В останні роки системи видобування інформації набули широкої

популярності у різних сферах, зокрема у медицині, політології, бізнесі, соціології, маркетингу тощо. Процес видобування інформації складається із кількох завдань (розпізнавання іменованих сутностей, виявлення та класифікація зв'язків, виявлення та класифікація подій тощо) [4].

Розпізнавання іменованих сутностей – це одне із найважливіших завдань у процесі видобування інформації. Воно полягає у виявленні слів (сутностей) та зарахуванні їх до однієї із класифікаційних категорій (organization – організація, person – особа, location – місце, date – дата, percentage – відсоток, address – адреса тощо)

Важливо зауважити, що системи розпізнавання іменованих сутностей застосовуються для розв'язання великої кількості завдань обробки природної мови, зокрема для створення питально-відповідальних систем (Question Answering Systems), машинного перекладу тощо.

Поки що не існує систем розпізнавання іменованих сутностей, а відповідно і систем видобування інформації для української мови, хоча аналогічні системи для опрацювання інших мов широко застосовуються у світі. Першим кроком на шляху до розроблення вищезгаданих систем є створення анотованого корпусу іменованих сутностей для української мови. У цій статті йдеться про нашу спробу розроблення такого корпусу на основі опрацювання україномовних відгуків про заклади харчування та медичних статей.

Аналіз останніх досліджень

Як уже згадувалося, протягом останніх років системи розпізнавання іменованих сутностей набули широкої популярності. Про це свідчать численні дослідження та розробки у цій сфері, які застосовують у таких галузях, як медицина, політологія, соціологія, маркетинг тощо.

Усі вищезгадані дослідження проведено для досить обмеженої кількості мов, зокрема для англійської, німецької, іспанської, італійської, російської та кількох інших. Проте жодних напрацювань для української мови не існує, що, своєю чергою, привертає увагу до актуальності такої роботи.

Цілі статті

Мета цього дослідження полягає у створені анотованого корпусу іменованих сутностей для української мови на основі опрацювання відгуків та медичних статей українською мовою.

Цілі статті:

- дослідити завдання розпізнавання іменованих сутностей та довести необхідність створення україномовного корпусу іменованих сутностей;
- проаналізувати наявні програмні засоби для анотування текстових повідомлень;
- вибрати найпридатніший засіб анотування;
- розробити схему анотування текстових повідомлень;
- проаналізувати отримані результати.

Основний матеріал

Анотований корпус іменованих сутностей

Анотований корпус іменованих сутностей – це корпус, що складається із різних текстових повідомлень (тип повідомлень залежить від жанру (тематики) корпусу). Словам у цих повідомленнях присвоюється одна із таких категорій, як організація, особа, місце, дата, адреса тощо.

Процес створення такого корпусу можна розділити на такі етапи:

- визначення жанру (тематики) корпусу (політологія, економіка, соціологія, медицина тощо);
- підбір текстових повідомлень вищезгаданого жанру, які будуть анотованими з метою створення корпусу;
- вибір програмних засобів для анотування текстових повідомлень;
- визначення типів іменованих сутностей та розроблення схеми анотування текстових повідомлень;
- анотування текстових повідомлень.

Підбір текстів для анотації

Для створення повного та функціонального корпусу потрібно передусім визначити жанр (тематику) текстових повідомлень, які згодом будуть анотованими. Створити тематико-незалежний корпус дуже важко, адже в процесі анотації різногалузевих текстів ми постійно стикатимемося із проблемою неоднозначностей іменованих сутностей у різних контекстах.

У цій статті детально описано процес створення корпусу іменованих сутностей на основі опрацювання україномовних відгуків про заклади харчування, які взято з форуму "Посиденьки" (http://posydenky.lvivport.com/) та із сайта "Відпочинок у Львові" (http://v.lviv.ua/). Вищезгадані сайти було вибрано джерелом збору матеріалів, оскільки всі відгуки на цих сайтах подано українською мовою, а кількість відгуків вибраної тематики дає змогу створити корпус належного обсягу. Загалом у ході дослідження анотовано 1200 відгуків.

У ході дослідження ми з'ясували, що найактуальнішими у наш час є розробки систем видобування інформації для медичної галузі. Причиною цього є той факт, що протягом останніх років працівники медицини зіткнулись з проблемою неймовірно швидкого зростання обсягів наукових публікацій у вищезгаданій галузі. Як наслідок, медики стверджують, що практично неможливо ознайомитися з усіма новинами в галузі медицини без допоміжних систем [3].

Саме тому ми вирішили створити анотований корпус іменованих сутностей із галузі медицини. Проте ми також вирішили спершу анотувати певну кількість україномовних відгуків про заклади харчування з метою подальших досліджень та порівнянь двох корпусів.

Програмні засоби та схема анотування текстових повідомлень

У наш час існує дуже багато засобів для анотування текстів. Серед найвідоміших GATE, LbjNerTagger, CRFClassifier [6].

Проаналізувавши вищезгадані засоби, для анотації текстів ми вибрали систему GATE.

GATE (General Architecture for Text Engineering) – це система для опрацювання природної мови, яка дає можливість створити нарізноманітніші схеми анотування [8].

У процесі дослідження розробили схему для анотації текстових повідомлень. Для створення такої схеми насамперед необхідно скласти перелік типів іменованих сутностей, який згодом стане основою схеми для анотації. Такий перелік іменованих сутностей наведено у табл. 1.

Таблиця 1

Типи іменованих сутностей

TYPE	FEATURES	EXAMPLES
address	street	вул. Личаківська, пл. Ринок 5, вул. Ст.Бандери 28
	postcode	79000, K1A 0A6, 10003
	phone	(0322)725569, +38(044)358053, 0936788743
	email	sekine@cs.nyu.edu, rector@lp.edu.ua
	url	http://lp.edu.ua, www.bbc.com
time	date	1.02.2012, 14 квітня, восени, День Незалежності, 2013 р., липень, понеділок, вихідні дні, 4 роки тому
	time	12:10, 6 год., північ
	period	30 хвилин, 2 години, півдня, один тиждень, 3 роки
location	continental_region	Європа, Азія, Північна Америка
	country	Україна, Японія, США, Канада
	region	Західна Україна, Київська обл.,
	city	Лондон, Харків, Торонто
money		50 грн., $40
organization		Microsoft, Світоч, Мапа
percent		100%, 22,5%
person	name	Андрій Худо, А.В.Іванченко, Зірко В.К.
	position	директор, офіціант, професор, лікар
	nickname	Nezabudka, ola-lyola

Продовження табл. 1

1	2	3
food		яблучний пиріг, сік, суші, картопля, м'ясні страви, пиво, чай, солодощі
token	pos	N (Noun - Іменник)
		V (Verb - Дієслово)
		A (Adjective - Прикметник)
		P (Pronoun - Займенник)
		R (Adverb - Прислівник)
		S (Adposition - Прийменник)
		C (Conjunction - Сполучник)
		M (Numeral - Числівник)
		Q (Particle - Частка)
		I (Interjection - Вигук)
		Y (Abbreviation - Скорочення)
		X (Residual - Залишок)
	lemma	
	morphology	Тег згідно з http://nl.ijs.si/ME/V4/msd/html/msd-uk.html

Вигляд розмітки у середовищі GATE відображено на рис. 1–2.

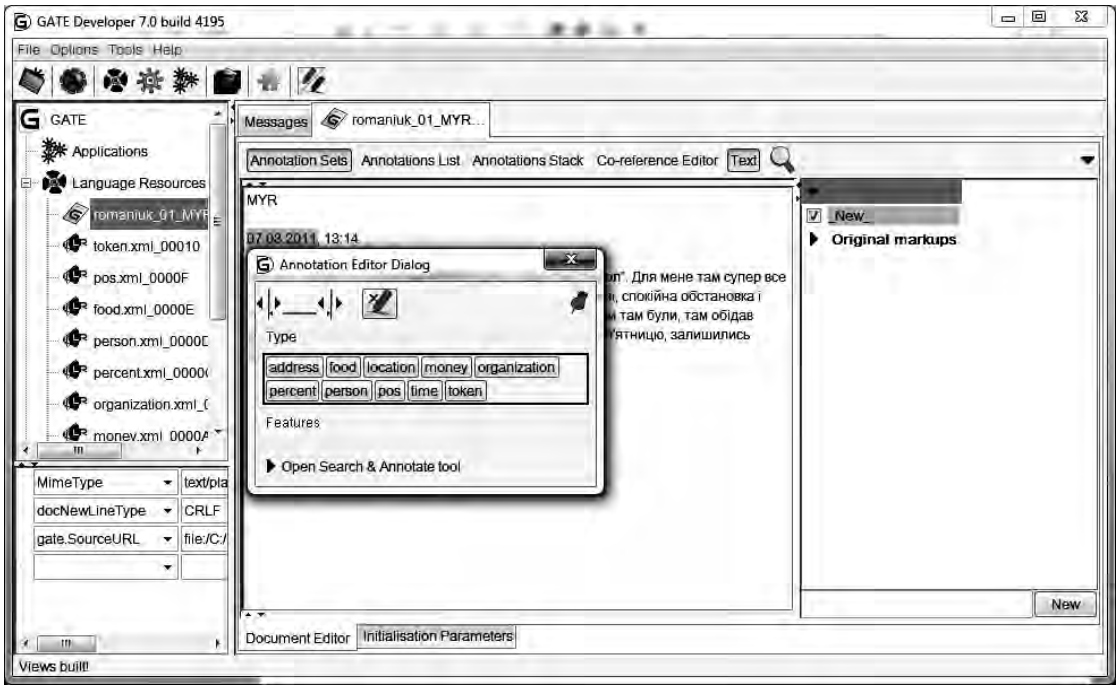

Рис. 1. Схема розмітки у середовищі GATE

Схему анотування для корпусу іменованих сутностей розроблено за допомогою пакета CREOLE (Collection of Reusable Objects for Language Engineering), що містить в собі клас AnnotationSchema [6].

Інформацію про кожну розроблену мітку схеми збережено в окремому xml файлі. У табл. 2 представлено структури різних міток.

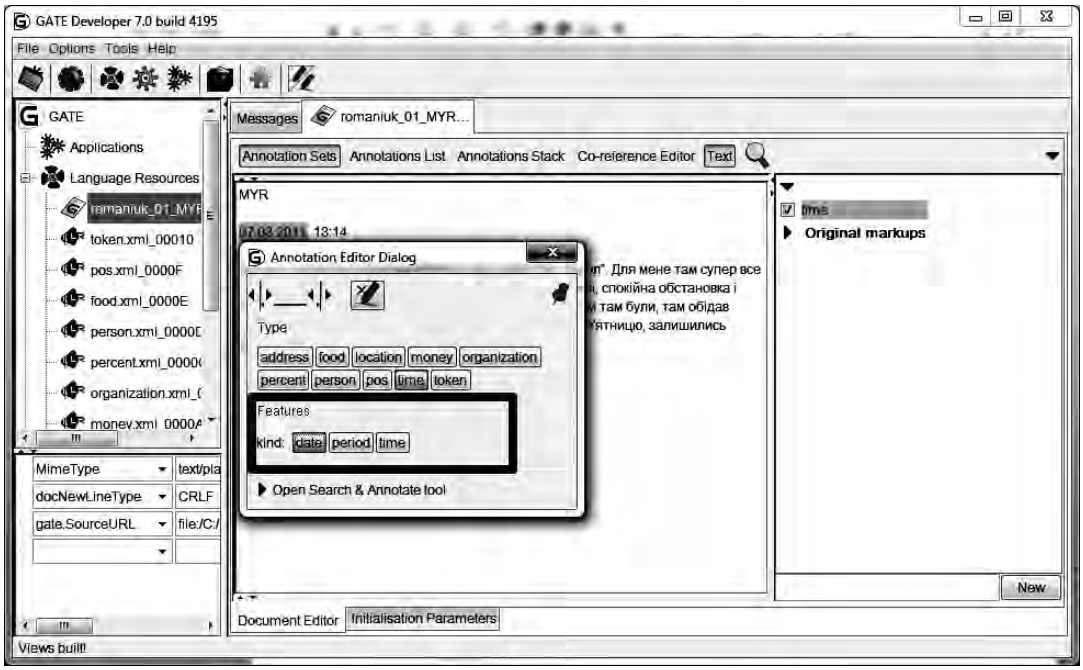

Рис. 2. Схема розмітки у середовищі GATE із відображенням атрибутів мітки time

Таблиця 2

Структура міток схеми анотування

Назва	Схема
1	2
address	<?xml version="1.0"?> <schema> <element name="address"> <complexType> <attribute name="kind" use="optional"> <simpleType> <restriction base="string"> <enumeration value="street"/> <enumeration value="postcode"/> <enumeration value="phone"/> <enumeration value="email"/> <enumeration value="url"/> </restriction> </simpleType> </attribute> </complexType> </element> </schema>
food	<?xml version="1.0"?> <schema> <element name="food" type="string" /> </schema>
location	<?xml version="1.0"?> <schema> <element name="location"> <complexType> <attribute name="kind" use="optional"> <simpleType> <restriction base="string">

Продовження табл. 2

1	2
	<enumeration value="continental_region"/> <enumeration value="country"/> <enumeration value="region"/> <enumeration value="city"/> </restriction> </simpleType> </attribute> </complexType> </element> </schema>
money	<?xml version="1.0"?> <schema> <element name="money" type="string" /> </schema>
organization	<?xml version="1.0"?> <schema> <element name="organization" type="string" /> </schema>
percent	<?xml version="1.0"?> <schema> <element name="percent" type="string" /> </schema>
person	<?xml version="1.0"?> <schema> <element name="person"> <complexType> <attribute name="kind" use="optional"> <simpleType> <restriction base="string"> <enumeration value="name"/> <enumeration value="nickname"/> <enumeration value="position"/> </restriction> </simpleType> </attribute> </complexType> </element> </schema>
time	<?xml version="1.0"?> <schema> <element name="time"> <complexType> <attribute name="kind" use="optional"> <simpleType> <restriction base="string"> <enumeration value="date"/> <enumeration value="time"/> <enumeration value="period"/> </restriction> </simpleType> </attribute> </complexType> </element> </schema>

Закінчення табл. 2

1	2
token	<?xml version="1.0"?> <schema> <element name="token"> <complexType> <attribute name="pos" default="N"> <simpleType> <restriction base="string"> <enumeration value="N"/> <enumeration value="V"/> <enumeration value="A"/> <enumeration value="P"/> <enumeration value="R"/> <enumeration value="S"/> <enumeration value="C"/> <enumeration value="M"/> <enumeration value="Q"/> <enumeration value="I"/> <enumeration value="Y"/> <enumeration value="X"/> </restriction> </simpleType> </attribute> <attribute name="morphology"/> <attribute name="lemma"/> </complexType> </element> </schema>

Описану в цьому розділі схему розмітки можна використовувати для анотування не лише відгуків, а й будь-яких інших текстів, зокрема і медичних статей. Проте для створення анотованого корпусу медичних статей українською мовою ми використовуватимемо ширшу схему розмітки із більшою кількістю типів іменованих сутностей, адже для таких текстів характерна специфічна медична лексика, яка потребує ширшого переліку міток.

Приклад анотованого відгуку

Наведемо конкретний приклад відгуку:

Darcy
16.02.2011, 17:39
Кав'ярня "Золотий Дукат". Смачнюща кава і солодке, дуууууууууже швидке обслуговування, я аж не сподівалася.

Кожен відгук складається із імені або ніку автора, дати, часу і власне тексту самого відгуку.

Після опрацювання у середовищі GATE відгук зберігається у форматі xml. Окрім лише тегування іменованих сутностей, ми застосували до кожного із текстових документів наявну у системі GATE функцію токенізації (ANNIE Tokenizer) та розділення речень (ANNIE SentenceSplitter). Внаслідок цього текст xml файла досить об'ємний. Тому наведемо приклади окремих міток в xml форматі із різних відгуків.

Приклад мітки 'date' у форматі xml:

```
<?xml version="1.0"?>
<schema>
 <element name="date" type="string" />
</schema>
```

Приклад мітки ‘time’ у форматі xml:

<Annotation Id=“4” Type=“time” StartNode=“7” EndNode=“17”><Feature>
<Name className=“java.lang.String”>kind</Name>
<Value className=“java.lang.String”>date</Value></Feature></Annotation>

Приклад мітки ‘person’ у форматі xml:

<Annotation Id=“7” Type=“person” StartNode=“269” EndNode=“276”><Feature>
<Name className=“java.lang.String”>kind</Name>
<Value className=“java.lang.String”>position</Value></Feature></Annotation>

Переглянувши наведені вище приклади різних міток, бачимо, що кожна анотована одиниця має зазначену початкову і кінцеву позицію в тексті та значення своїх ознак (атрибутів). Це, своєю чергою, дає нам можливість подальшого аналізу та опрацювання отриманих файлів у ході подальших досліджень.

Висновки

Процес створення анотованого корпусу іменованих сутностей для української мови можна розділити на такі етапи: визначення жанру (тематики) корпусу (політологія, економіка, соціологія, медицина тощо); підбір текстових повідомлень вищезгаданого жанру, які будуть анотованими з метою створення корпусу; вибір програмних засобів для анотування текстових повідомлень; визначення типів іменованих сутностей та розроблення схеми анотування текстових повідомлень та власне анотування текстових повідомлень.

У цій статті описано результати нашого дослідження, зокрема створення анотованого корпусу іменованих сутностей для української мови на основі опрацювання україномовних відгуків про заклади харчування. Вищезгаданий корпус анотований у середовищі GATE. У статті також описана схема анотування та наведена структура анотованих відгуків, які зберігаються у xml форматі.

Створення такого корпусу має дуже велику практичну цінність та важливе значення для подальших розробок. Зокрема, для створення анотованого корпусу іменованих сутностей на основі україномовних медичних статей та для подальших досліджень у сфері розпізнавання іменованих сутностей та видобування інформації.

1. Azzam S. Using a Language Independent Domain Model for Multilingual Information Extraction / Saliha Azzam, Kevin Humphreys, Robert Gaizauskas and YorickWilks. – Applied Artificial Intelligence: An International Journal. Volume 13, Issue 7, 1999. 2. Benajiba Y., Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition/ Y. Benajiba, M. Diab, P. Rosso. – The International Arab Journal of Information Technology, Vol. 6, No. 5, November 2009. – pp. 464-473. 3. Jurafsky D. Speech and language processing / Daniel Jurafsky, James H. Martin. – [2nd ed.]. – Upper Saddle River, NJ: Pearson Education, Inc., 2009. – pp. 725-764. 4. Moens M. Information Extraction: Algorithms and Prospects in a Retrieval Context / Marie-Francine Moens. – Netherlands: Springer, 2006. – pp.65-85, 199-218. 5. Schäfer U. OntoNERdIE – Mapping and Linking Ontologies to Named Entity Recognition and Information Extraction Resources / Ulrich Schäfer. – Proceedings of the 5th International Conference on Language Resources and Evaluation LREC-2006, Genoa, Italy, ELRA, 5/2006. – pp. 1756–1761. 6. Shapiro S. Natural Language Tools for Information Extraction for Soft Target Exploitation and Fusion / Stuart C. Shapiro, Shane Axtell. – NY, 2007. – pp. 36-37. – Режим доступу: http://www.cse.buffalo.edu/~shapiro/Papers/shaaxt07.pdf. 7. Nédellec C. Ontologies and Information Extraction/ C. Nédellec, A. Nazarenko. - LIPN Internal Report, 2005. 8. Using GATE Developer. – Режим доступу: http://gate.ac.uk/sale/tao/splitch3.html#chap:developer. 9. Wimalasuriya D. Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches / Daya C. Wimalasuriya, Dejing Dou. - Journal of Information Science, XX (X) 2009, pp. 1–20.