УДК 81'322

Андрій Романюк,
Мар'яна Романишин
Національний університет "Львівська Політехніка"

ТОНАЛЬНИЙ СЛОВНИК УКРАЇНСЬКОЇ МОВИ НА ОСНОВІ СЕНТИМЕНТ-АНОТОВАНОГО КОРПУСУ

Докладно розглянуто всі етапи створення сентимент-анотованого корпусу українськомовних відгуків і тонального словника на його основі.

Ключові слова: емоційно-смисловий аналіз, тональний словник, сентимент-анотований корпус, засоби для анотування текстів.

Протягом останніх десяти років у галузі опрацювання природної мови широкої популярності набув емоційно-смисловий аналіз. Про це свідчить численна кількість досліджень у цій сфері. З'явилися такі цікаві проекти, як дослідження емоційного забарвлення відгуків про готелі [Kasper], банки [Deep sentiment analysis], ресторани, коментарів про фільми [Yessenov], продукти, повідомлень у блогах і соцмережах про політичні події тощо.

На сьогодні емоційно-смисловий аналіз застосовується в різних галузях знань: у соціології (наприклад, збирання даних із соціальних мереж про певні вподобання людей), політології (наприклад, збирання даних про політичні погляди), маркетингу (наприклад, збирання даних про найпопулярніші товари), медицині та психології (наприклад, визначення депресивних настроїв) тощо [Давыдов].

Наразі немає доступного емоційно-смислового аналізатора для української мови, але це питання активно опрацьовується. Створення такого аналізатора передбачає насамперед укладання тонального словника та сентимент-анотованого корпусу. У статті описано створення сентимент-анотованого корпусу відгуків українською мовою, а також генерацію тонального словника на основі такого корпусу.

Поняття тонального словника та методи його створення

Тональний словник, у найпростішому його вигляді, становить список слів і словосполучень зі значенням тональності для

кожного слова. Це може бути як числове значення в межах певної шкали (наприклад, від 1 до 10, де 1 – негативне слово, а 10 – позитивне), так і категорія (наприклад, позитивне чи негативне слово). Кожному слову приписується його частиномовна характеристика та початкова форма у випадку, коли словник містить різні словоформи одного слова.

Часто списки слів у словнику обмежують лише самостійними частинами мови. Наприклад, у роботі над повідомленнями російською мовою [Пазельская] використовувався тональний словник із найчастотніших іменників, прикметників, прислівників та дієслів, зібраних зі статей ЗМІ. Для кожного слововживання визначено частину мови, тональність і силу тональності за шкалою від 1 до 3. Прикметники і прислівники поділені на позитивні, негативні та ті, що підсилюють тональність. Іменники поділені на позитивні, негативні, потенційно позитивні та потенційно негативні (залежно від оточення: позитивні в позитивному оточенні, а негативні – в негативному). Дієслова поділені на 8 класів залежно від оточення і ролі в реченні, окремо подані дієслова-зв'язки. Є також слова-інвертори, що змінюють тональність сусідніх слів на протилежне.

Набір слів у тональному словнику переважно стосується однієї тематики, яку опрацьовуватиме аналізатор. Оскільки складання такого списку слів власноруч потребує дуже багато часу, застосовують автоматичні способи видобування слів певної тематики зі сентимент-анотованих корпусів та онтологій, що стосуються певної тематики. Якщо ж тематика словника не передбачена, використовують загальнотематичні семантичні мережі. Наприклад, у праці [Moilanen] послуговуються тональним словником, створеним на основі WordNet 2.1. Прикладом багатомовного тонального словника є SentiWordNet, який часто використовують для здійснення різноманітних завдань у сфері емоційно-смислового аналізу [Denecke]. Достатньо часто послуговуються також General Inquirer Lexicon [Agrin].

Хоч застосування онтологій є досить поширеним способом генерації тональних словників, воно має певні недоліки, яких мож-

на уникнути за умови залучення анотованих корпусів. По-перше, сентимент-анотований корпус є джерелом контекстів для слів, які потраплять до тонального словника, що дає можливість краще визначити емоційне забарвлення кожного слова, а в онтології переважно містять ізольовані слова. По-друге, корпус подає велику кількість нелітературних слів, які важливі для визначення емоційного забарвлення повідомлення, але відсутні в онтологіях.

Оскільки для української мови немає доступного сентимент-анотованого корпусу, було вирішено створити такий корпус власноруч і на його основі згенерувати основу тонального словника.

Створення сентимент-анотованого корпусу

Сентимент-анотований корпус – це корпус текстових повідомлень, у якому кожному повідомленню присвоєно значення емоції, яку воно передає. Створення сентимент-анотованого корпусу – це невід'ємна частина реалізації системи емоційно-смислового аналізу.

По-перше, такий напівавтоматично розмічений корпус дає розуміння того, як за допомогою текстових повідомлень людина висловлює свої емоції щодо певного об'єкта (емоційно забарвлена лексика, ідеограми на кшталт смайликів, пунктуація). По-друге, такий корпус може стати основою для створення тонального словника. По-третє, готовий сентимент-анотований корпус є достатнім засобом для перевірки роботи розробленої системи емоційно-смислового аналізу.

Процес створення сентимент-анотованого корпусу було поділено на такі етапи:

- добір текстових повідомлень для майбутнього корпусу;
- визначення програмних засобів для анотування текстових повідомлень;
- вироблення схеми анотування текстових повідомлень;
- власне анотування текстових повідомлень.

Формування вибірки текстових повідомлень для майбутнього корпусу

Першим кроком для добору повідомлень для майбутнього сентимент-анотованого корпусу є визначення тематики текстів.

Тематика важлива для емоційно-смислового аналізу, оскільки ті самі слова в межах різних тематик можуть мати протилежні тональності. Тематикою повідомлень для нашого дослідження стали відгуки про заклади харчування. На актуальність цієї тематики вказує широке обговорення її на форумах та в соціальних мережах.

Українськомовні відгуки про заклади харчування, які стали основою для корпусу, було взято з популярного форуму http://posydenky.lvivport.com/ та з сайту http://v.lviv.ua/. Вибір припав саме на ці сайти через значну кількість відгуків, які відповідають обраній тематиці, а також тому, що більшість відгуків на цих сайтах написані українською мовою, що дає можливість частково уникнути проблеми фільтрування повідомлень за мовною ознакою та зосередитись власне на створенні корпусу. Структура відгуків на обох сайтах однакова, оскільки кожен відгук містить ідентифікатор автора, час написання відгуку та власне текстове повідомлення.

Визначення програмних засобів для анотування текстів

На сьогодні є багато зручних засобів для анотування текстів. З-поміж найпоширеніших можна назвати: Callisto, WordFreak, GATE, BRAT, DOMEO, CLaRK, Ellogon, UAM та ін [Shapiro]. Коротко про деякі з цих засобів:

Callisto (http://callisto.mitre.org) – це простий засіб анотування, розроблений для підтримки лінгвістичного анотування текстів для будь-якої мови, з підтримкою Unicode. Анотовані тексти зберігаються у форматі ATLAS, який можна імпортувати в xml.

WordFreak (http://wordfreak.sourceforge.net) – це засіб, який підтримує ручне та автоматичне анотування лінгвістичних даних, а також дає можливість застосовувати автоматичне навчання на виправленні анотацій, зроблених людиною. Цей засіб здебільшого використовується для перевірки готового анотованого тексту.

GATE (http://gate.ac.uk/) – це ціла система для опрацювання природної мови, яка, крім інших засобів, містить також і засіб для ручного та автоматичного анотування текстів. GATE дає можливість створювати найрізноманітніші схеми анотування.

DOMEO (http://annotationframework.org/) – це онлайнове середовище, яке допомагає анотувати текст на основі вбудованої онтології. Цей засіб підтримує ручне, напівавтоматичне та автоматичне анотування.

CLaRK (http://www.bultreebank.org/clark/index.html) – це система для розробляння корпусів, основною метою якої є мінімізувати ручну роботу у процесі створення лінгвістичних ресурсів.

Ellogon (http://www.ellogon.org/) – це багатомовне міжплатформне середовище для опрацювання природної мови з відкритим кодом, яке застосовують як окремі науковці, так і компанії, які займаються створенням систем опрацювання природної мови.

Дослідивши описані вище засоби анотування текстів, ми визначили, що для нашого завдання найбільше підходять системи GATE, CLaRK та Ellogon. З-поміж них було вибрано GATE, оскільки ця система проста у використанні, надає зручні засоби редагування анотованого тексту, можливість створення складних схем анотування, можливість зберігати анотовані тексти в форматі xml, працювати з кількома текстовими файлами і кількома схемами анотування одночасно, аналізувати саме українську мову, а також забезпечує підсвічування анотованого тексту різними кольорами, що зручно під час накладання анотацій одна на одну.

Вироблення схеми анотування текстових повідомлень

Схему анотування для сентимент-анотованого корпусу було розроблено за допомогою пакета CREOLE (Collection of REusable Objects for Language Engineering), який має клас AnnotationSchema. Цей пакет дає можливість створювати схеми анотування та діалогові вікна для роботи з ними. Файл конфігурацій creole.xml містить інформацію про ресурси, які використовуються. У нашому випадку – це назви файлів із відповідними мітками [Using GATE Developer].

Розроблена схема анотування українськомовних відгуків про заклади харчування має такі структурні одиниці:

- автор;
- дата;
- відгук;

- цитування попереднього повідомлення;
- речення;
- частина складного речення;
- назва закладу харчування, про який іде мова;
- слово;
- url-адреса.

Автор відгуку позначається міткою nickname. У вхідних даних автор вказаний у першому рядку.

Дату написання відгуку (мітка date) подано після автора.

Власне відгук автора (мітка review) виділено без урахування цитування попереднього повідомлення, якщо таке є. Це потрібно для того, щоб визначати суб'єктивну оцінку автора поточного відгуку, а не автора цитованого повідомлення.

Цитування попереднього повідомлення визначаємо за характерними ознаками (повідомлення починається на "Цитата:") і позначаємо міткою citing.

Окремо виділяється кожне речення відгуку (мітка sentence).

Кожна частина складного речення позначається міткою clause. У випадку простих речень ці дві мітки збігаються. Для кожного такого підречення визначається тональність (атрибут під назвою sentiment): позитивне (positive), негативне (negative) чи нейтральне (neutral). Для цілого складного речення тональність не визначено, оскільки одне складне речення може містити інформацію і з негативним, і з позитивним забарвленням. У відгуку окремо анотуємо також і назву закладу чи закладів харчування, про які йде мова (мітка target).

У кожному емоційно забарвленому підреченні (clause) виділяємо слова чи ідіоматичні словосполучення (позначаються міткою word).

Для кожного виділеного слова було визначено набір атрибутів:

- початкова форма слова (атрибут lemma): значення вписується в автоматизованому режимі;
- частина мови (атрибут part_of_speech) має такі значення: n – іменник, v – дієслово, adj – прикметник, adv – прислівник, pro – займенник, con – сполучник, pre – прийменник, par – частка, exc – вигук, num – числівник, und – інше (напр., смайлики);

- емоційне забарвлення (атрибут sentiment) має такі значення: positive – слова чи словосполучення, що самостійно виражають позитивне значення; negative – слова чи словосполучення, що самостійно виражають негативне значення; neutral – слова чи словосполучення, що самостійно не виражають ні позитивного, ні негативного значення; intensifier – слова-підсилювачі, що не мають самостійного емоційного забарвлення, але підсилюють емоційне забарвлення наступного слова чи цілого підречення (наприклад, такі слова, як *дуже*, *надзвичайно*, *безмежно*, *вкрай*, *досить*); invertor – слова-інвертори, які не мають самостійного емоційного забарвлення, але змінюють емоційне забарвлення наступного слова чи цілого підречення на протилежне (наприклад, такі слова, як *не*, *нема*, *немає*, *неможливо*, *нереально*);
- емоція (атрибут emotion) має такі значення: joy – радість, sadness – сум, anger – злість, fear – страх, disgust – огида, surprise – здивування, none – якщо слово чи словосполучення самостійно не передає емоції.

Атрибути sentiment та emotion матимуть значення neutral і none відповідно для всіх службових частин мови. Лише самостійні частини мови можуть мати інші значення, оскільки лише самостійні частини мови можуть нести певне емоційне забарвлення ізольовано від контексту.

Для того щоб визначити базові емоції для схеми анотування відгуків, було проаналізовано набори базових емоцій за теоріями різних психологів. Шість вищезгаданих емоцій визнані базовими емоціями людини відповідно до концепції відомого психолога Пола Екмана. Базові емоції Екмана у психології – це культурно незалежні емоції, які з'являються в людини впродовж перших шести місяців її життя. Це також такий набір емоцій, які легко передати як мімікою, так і вербально [Ekman]. Звичайно, існують й інші концепції, але базові емоції Екмана вважають у сучасній науковій думці фундаментальними.

Міткою url позначаємо url-адреси, якщо такі є у відгуку.

Інформація про кожну мітку розробленої схеми анотування занесена в окремий xml-файл. На рис. 1 представлено структуру файлу для мітки clause.

```
<?xml version="1.0"?>
<schema>
  <element name="clause" type="string">
    <complexType>
    <attribute name="sentiment" default="neutral"
      <simpleType>
        <restriction base="string">
        <enumeration value="positive"/>
        <enumeration value="negative"/>
        <enumeration value="neutral"/>
        </restriction>
      </simpleType>
    </attribute>
  </complexType>
 </element>
</schema>
```

Рис. 1. Структура файлу clause.xml

На рис. 2 зображено діалогове вікно з назвою і атрибутом мітки. Значення можна автоматично вибрати з випадного списку.

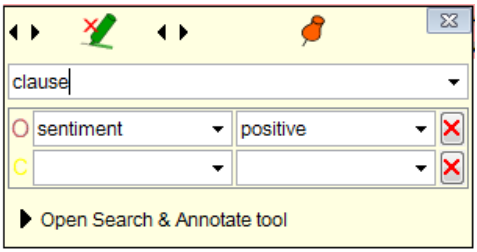

Рис. 2. Діалогове вікно з атрибутом мітки clause в середовищі Gate 7.0

У файл конфігурацій creole.xml також внесено відповідний запис про цю мітку:

```
<AUTOINSTANCE>
    <PARAM NAME ="xmlFileUrl" VALUE ="resources/schema/clause.xml" />
</AUTOINSTANCE>
```

Власне анотування текстових повідомлень

Із метою створення сентимент-анотованого корпусу відгуків українською мовою було залучено студентів кафедри прикладної лінгвістики Національного університету "Львівська політехніка". Для них було розроблено розрахункову роботу в межах курсу "Комп'ютерна лінгвістика". Метою роботи стало ознайомлення студентів із практикою створення корпусів та їх анотування.

Кожен студент отримав методичні вказівки щодо роботи в середовищі GATE, докладний опис розробленої схеми анотування з поясненнями і прикладами, а також власне набір відгуків у текстовому форматі. Після анотування студенти зберігали кожен відгук у форматі xml.

Результатом виконання розрахункової роботи став корпус сентимент-анотованих відгуків обсягом 600 відгуків у зручному для використання форматі. Цей корпус, однак, потребує додаткової перевірки через суттєву кількість помилок.

Приклад анотованого відгуку

Тетянка

24.04.2009, 22:29

а я "Цукерню" люблю і кафешку з тортиками на Дудаєва....

Відгук складається з ідентифікатора автора, дати і часу написання та власне тексту. Після анотування цей відгук було збережено у форматі xml. Структура xml-файлу містить інформацію про кодування, відомості про сам документ; власне відгук із розміткою анотацій і, нарешті, опис міток. Відгук із розміткою анотації має такий вигляд:

<TextWithNodes><Node id="0" />Тетянка
<Node id="8" />
<Node id="9" />
<Node id="11" />24.04.2009, 22:29<Node id="28" />
<Node id="30" />
<Node id="32" />а<Node id="33" /> <Node id="34" />я<Node id="35" /> "<Node id="37" />Цукерню<Node id="44" />" <Node id="46" />люблю<Node id="51" /> <Node id="52" />і<Node id="53" /> <Node id="54" />кафешку <Node id="61" /> <Node id="62" />з<Node id="63" /> <Node id="64" /> тортиками<Node id="73" /> <Node id="74" />на<Node id="76" /> <Node id="77" />Дудаєва<Node id="84" />....<Node id="88" />
<Node id="89" /> <Node id="90" /></TextWithNodes>

Приклад мітки clause у форматі xml:

```
<Annotation Id="7" Type="clause" StartNode="32" EndNode="88">
<Feature>
    <Name className="java.lang.String">sentiment</Name>
    <Value className="java.lang.String">positive</Value>
</Feature>
</Annotation>
```

Приклад мітки word у форматі xml:

```
<Annotation Id="11" Type="word" StartNode="46" EndNode="51">
<Feature>
    <Name className="java.lang.String">part_of_speech</Name>
    <Value className="java.lang.String">v</Value>
</Feature>
<Feature>
    <Name className="java.lang.String">lemma</Name>
    <Value className="java.lang.String">любити</Value>
</Feature>
<Feature>
    <Name className="java.lang.String">sentiment</Name>
    <Value className="java.lang.String">positive</Value>
</Feature>
<Feature>
    <Name className="java.lang.String">emotion</Name>
    <Value className="java.lang.String">joy</Value>
</Feature>
</Annotation>
```

Із наведених вище прикладів можна побачити, що кожна мітка містить початкову і кінцеву позиції фрагменту тексту, анотованого цією міткою, а також назви і значення її атрибутів, якщо такі є. Таке xml-дерево можна легко розпізнати і застосовувати для майбутніх досліджень.

Створення тонального словника

Створений сентимент-анотований корпус відгуків українською мовою дозволяє згенерувати основну частину тонального словника.

Кожне тональне слово в корпусі приведено до нижнього регістру й записано до словника. Таким чином, словник містить

слова і словосполучення, інформацію про частиномовну належність кожного слова, його тональність (позитивну чи негативну) й емоційне забарвлення, якщо таке є. До словника також записується початкова форма слова з такими ж атрибутами. Такі відомості надають можливість застосувати тональний словник для здійснення емоційно-смислового аналізу: у випадку відсутності потрібної словоформи в тональному словнику, слову присвоюватиметься тональність його початкової форми.

Оскільки службові слова ізольовано не несуть емоційного забарвлення, словник міститиме лише самостійні частини мови. Окрім початкових форм слів, словник міститиме й інші словоформи. Це пов'язано з тим, що використовуючи лише початкову форму слова, ми втрачаємо морфологічну інформацію, яка може бути корисною для подальшого емоційно-смислового аналізу. Наприклад, слова *люблю* і *любив* можуть мати різну тональність. Якщо в першому випадку тональність швидше за все буде позитивною, то в другому випадку може бути менш позитивною чи навіть негативною.

Окрім позитивних та негативних слів і словосполучень, тональний словник також міститиме слова-інвертори (наприклад, *не*, *немає*, *неможливо*) і слова-підсилювачі (*дуже*, *надзвичайно*, *безмежно*, *вкрай*).

Таким чином, за допомогою розробленого сентимент-анотованого корпусу відгуків можна згенерувати тональний словник, який міститиме позитивні та негативні слова і словосполучення, слова-інвертори та слова-підсилювачі.

Перспективою презентованого дослідження є подальше поповнення такого словника, зокрема шляхом залучення словників синонімів і антонімів. Надалі цей словник буде використано для реалізації емоційно-смислового аналізатора, метою якого є визначення емоційного забарвлення повідомлень українською мовою.

1. *Давыдов А. А.* Системная социология: Opinion Mining / А. А. Давыдов. – М., 2009. – Режим доступу: http://www.isras.ru/index. php?page_id=1024. 2. *Пазельская А.* Метод определения эмоций в текстах на русском языке / А. Г. Па-

зельская, А. Н. Соловьев // Компьютерная лингвистика и интеллектуальные технологии. Сб. научных статей. Вып. 10 (17). – М., 2011. – С. 510–522. 3. *Agrin N.* Developing a Flexible Sentiment Analysis Technique for Multiple Domains / Nate Agrin. – 2006. – Режим доступу: http://courses.ischool.berkeley.edu/i256/f06/projects/agrin.pdf. 4. Deep sentiment analysis with attensity analyze optimises Lloyds' customer service. – Режим доступу: http://www.attensity.com/wp-content/uploads/2010/09/LloydsSuccessStory.pdf. 5. *Denecke K.* Using SentiWordNet for Multilingual Sentiment Analysis / Kerstin Denecke // ICDE Workshops. – 2008. – P. 507–512. 6. *Ekman P.* Basic Emotions / Paul Ekman // Handbook of Cognition and Emotion. – John Willey & Sons Ltd, 1999. – P. 45–60. 7. *Kasper W.* Sentiment Analysis for Hotel Reviews / Walter Kasper // Proceedings of the Computational Linguistics-Applications Conference. – Poland, 10/2011. – P. 45–52. 8. *Moilanen K.* Multi-entity Sentiment Scoring / Karo Moilane, Stephen Pulman // Proceedings of Recent Advances in Natural Language Processing (RANLP 2009). – Bulgaria (Borovets). – September 14–16 2009. – P. 258–263. 9. *Shapiro S.* Natural Language Tools for Information Extraction for Soft Target Exploitation and Fusion / Stuart C. Shapiro. – NY, 2007. – P. 36–37. – Режим доступу: http://www.cse.buffalo.edu/~shapiro/Papers/shaaxt07.pdf. 10. Using GATE Developer. – Режим доступу: http://gate.ac.uk/sale/tao/splitch3.html#chap:developer. 11. *Yessenov K.* Sentiment Analysis of Movie Review Comments / Kuat Yessenov, Sasa Misailovic. – Massachusetts Institute of Technology, Spring 2009. – Режим доступу: http://people.csail.mit.edu/kuat/courses/6.863/report.pdf.

Представлены все этапы создания сентимент-аннотированного корпуса украиноязычных отзывов и тонального словаря на его основе.

Ключевые слова: сентимент-анализ, тональный словарь, сентимент-аннотированный корпус, средства для аннотирования текстов.

This paper deals with the implementation of sentiment-annotated corpus of Ukrainian reviews and the creation of sentiment dictionary based on it. The paper describes all the stages of sentiment-annotated corpus creation in detail.

Keywords: sentiment analysis, sentiment dictionary, sentiment-annotated corpus, tools for annotating texts.

Стаття надійшла до редакції 1.09.2012