УДК 681.3.016

С.Ф. Теленик, д.т.н., проф.
Р.В. Смічик, аспір.
Національний технічний університет України "КПІ"

ОБРОБЛЕННЯ ТЕКСТІВ ПРИРОДНОЇ МОВИ НА ОСНОВІ МОДЕЛЕЙ РОЗУМІННЯ В АВТОМАТИЗОВАНИХ СИСТЕМАХ УПРАВЛІННЯ

Стаття присвячена проблемі розробки природномовного інтерфейсу в автоматизованих системах управління. Запропоновано алгоритм аналізу текстової інформації, в якому моделюється розуміння людиною природномовних висловлювань. Подана концепція та характеристика алгоритму, вказані напрямки подальшої роботи з обробки текстів на основі моделей розуміння.

Вступ

Проявом зростаючої ролі інформації в сучасному суспільстві є створення систем її оброблення, автоматизованого управління і прийняття рішень. Людство нагромадило величезні обсяги інформації, найбільш поширеною формою подання якої є природна мова (ПМ). Перевагами ПМ в порівнянні зі штучними мовами, що застосовуються в автоматизованих системах на даному етапі розвитку інформаційних технологій, є гнучкість і універсальність. Таким чином, зростає необхідність в системах зберігання, аналізу і синтезу саме природномовних текстів.

Труднощі оброблення текстів ПМ полягають в наявності величезної кількості структур (словосполучень, фраз, речень), що можуть бути породжені. Це унеможливлює застосування описового методу аналізу. Крім того, одна і та ж структура може мати декілька смислових значень, що породжує проблему неоднозначності.

Існує два види неоднозначності: лексична неоднозначність, пов'язана зі здатністю слова виражати різні значення, і структурна неоднозначність, яка полягає в здатності речення, фрази або словосполучення виражати різні значення в залежності від ситуації, навколишньої обстановки, цілей суб'єктів обміну інформацією, культурних особливостей суспільства, що користується конкретною ПМ, тощо.

Для розв'язання зазначеної проблеми недостатньо застосування тільки граматичних правил і обмежень. Необхідне застосування позамовних знань про контекст і ситуацію, цілі і плани користувача автоматизованої системи, зовнішню обстановку, а також загальнофонових знань (про світ, події, процеси тощо).

Отже, виникає комплексна проблема оброблення текстів ПМ на основі знань в автоматизованих системах.

У статті запропонований підхід до розв'язання цієї проблеми шляхом імітації людського розуміння текстової інформації. Подається концепція алгоритму оброблення текстів ПМ на принципах штучного інтелекту на основі лінгвістичних моделей і моделей подання знань, наводиться сам алгоритм, аналізуються його особливості в загальній схемі розв'язання проблеми аналізу, інтерпретації, вербалізації, інших проблем комп'ютерної лінгвістики.

1. Представлення знань

У комп'ютерній лінгвістиці для розв'язання проблем оброблення текстів ПМ ідея залучення знань уже не вимагає додаткових аргументів [1–3]. Для розв'язання тих же проблем в автоматизованих системах виправданим буде застосування:

1) знань про мову:
   - граматики і лексичної семантики;
   - принципів мовного обміну інформацією;
2) позамовних знань про:
   - контекст і ситуацію;
   - цілі і плани адресата;
   - цілі і плани автора висловлювання;
   - фонових знань про світ (події, процеси тощо).

Найбільш ефективним методом подання знань зарекомендували себе бази знань. Для їх реалізації все частіше використовуються об'єктно-орієнтовані СУБД, ефективні саме для подання найпоширеніших в комп'ютерній лінгвістиці семантичних мереж. З одного боку, знижуються вимоги до спеціального програмного забезпечення систем оброблення ПМ, з іншого боку, тоді семантична мережа може подавати різноманітні знання про типи об'єктів та зв'язків, наприклад, один з нижніх рівнів мережі може бути виділеним для зберігання словників ПМ.

Для зручності будемо виділяти в БЗ декілька типів семантичних мереж: дій (в т.ч. додаткових дій, що виражаються дієприслівниками); понять (сутностей); узагальнених ознак (прислівників) тощо. Спеціальними назвемо ознаки, які в термінах граматики виражаються природною мовою прикметниками або дієприкметниками. Спеціальні ознаки не виділяються в окрему підмережу, оскільки вони за визначенням можуть в текстах ПМ стосуватися тільки сутностей, і тому для підвищення ефективності роботи алгоритму пошуку повинні зберігатися разом з ними в одній логічній області. Узагальнені ж ознаки наділені універсальністю, тобто можуть доповнювати значення не тільки сутностей, але і дій, спеціальних ознак та інших узагальнених ознак (наприклад, *крок* ***назад*** – ознака сутності, *підіймати* ***вгору*** – ознака дії, *вельми* ***спірне*** *питання* – ознака спеціальної ознаки, *нескінченно* ***багато*** – ознака узагальненої ознаки). Тому для них виділяємо окрему семантичну підмережу, а зв'язок з іншими підмережами буде здійснюватися по зовнішньому ключу.

Імітація людського розуміння полягає в заповненні певної моделі розуміння – структури внутрішньої репрезентації вхідного тексту – за допомогою аналізу текстів ПМ, насамперед пошуку в БЗ, застосування певних правил і обмежень, незалежних або залежних від конкретної ПМ, тощо. Структура внутрішньої репрезентації вхідного тексту складається із структури внутрішніх репрезентацій речень і деякої загальної структури [5]. Перша включає пропозиційне ядро, модальність, прагматичний зміст, граматичне значення речення, експресивно-стилістичне забарвлення, нормативні наслідки та персональність, а друга – їх узагальнення для всього тексту, насамперед узагальнена характеристика внутрішньої репрезентації речень.

Оскільки смислове значення тексту, що обробляється, залежить від смислового значення його складових – параграфів, складних і простих речень, простих складових складних речень і слів, – структура внутрішньої репрезентації тексту ПМ подається у вигляді ієрархії, на кожному рівні якої виділяються атрибути складових тексту, які заповнюються значеннями в процесі роботи алгоритму аналізу. Детальний опис структури БЗ та структури внутрішньої репрезентації тексту є темою окремої публікації, тож розпочнемо з власне алгоритму аналізу текстів ПМ.

2. Концепція алгоритму аналізу тексту ПМ

Початковими припущеннями для алгоритму є умова граматичної правильності вхідного тексту і обмеження використання алгоритму для аналізу текстів наукового та ділового стилів.

Входом алгоритму є природномовний текст, а результатом роботи алгоритму повинна бути заповнена структура внутрішньої репрезентації тексту.

Коротко розглянемо основні ідеї, покладені в основу алгоритму. По-перше, в основу структуризації операцій алгоритму покладена структуризація тексту ПМ. Заповнення структури відбувається знизу вгору (від значення слів до значення тексту) з передбачуваними поверненнями на більш низький рівень, якщо на ньому при переході на наступний рівень залишалися незаповнені атрибути. Для запропонованого алгоритму важливе виділення наступних рівнів смислового аналізу вхідного тексту на ПМ:

1) слів;
2) простих речень (та простих складових складних речень);
3) речень;
4) абзаців;
5) тексту.

По-друге, алгоритм відрізняє комплексне врахування граматичних, семантичних, прагматичних та інших аспектів тексту, поєднуване з основним акцентом на аналіз семантики. Базовою формою обміну інформацією на ПМ є речення. В ньому слова, якими називаються дії, поняття та їх властивості, об'єднані за певними правилами для вираження деякої думки.

По-третє, управління роботою алгоритму аналізу здійснюється на основі структури речення, семантичною основою якої є дія. Структура речення представляється у вигляді дерева, де коренем є слово, що виражає дію. Даний вибір пояснюється тим, що дії набагато рідше, ніж сутності, згадуються в повідомленні на ПМ неявно, а неявне вказування передбачає роботу алгоритму вгору за вказаною нами ієрархією тексту з незаповненими атрибутами на більш низькому рівні, що збільшує кількість можливих варіантів інтерпретації значення, які підлягають перевірці.

Виділення в складному реченні простих складових дозволяє визначити додаткову інформацію про основну дію (у разі складнопідрядного речення) або виділити рівноправні дії, що відбуваються одночасно або послідовно (у разі складносурядного речення).

Виділення в тексті параграфів використовується переважно для визначення семантичних суб'єкта і об'єкта тексту і границь пошуку антецедентів для анафоричного аналізу.

У алгоритмі, що пропонується, до інформації, витягнутої з БЗ, застосовуються граматичні, семантичні і прагматичні правила визначення значення кожного елемента структури внутрішньої репрезентації.

Можливим результатом роботи алгоритму є структура розуміння, деякі атрибути якої мають більше ніж одне значення, що означає дійсну неоднозначність через недостатність інформації, що міститься в тексті.

3. Алгоритм

Спочатку подамо власне алгоритм аналізу у традиційному узагальненому вигляді.

Крок 1 Розділення тексту на слова.
Крок 2 Для кожного слова пошук і читання словникових статей БЗ.
Крок 3 Об'єднання слів в речення.
Крок 4 Для кожного речення:
- 4.1 Визначення типу (повідомлення/запит).
- 4.2 Розділення на прості складові.
- 4.3 Визначення типів зв'язків простих складових складного речення.

Крок 5 Об'єднання речень в параграфи.
Крок 6 Об'єднання параграфів в текст.
Крок 7 Для всіх простих речень і всіх простих складових складних речень:
- 7.1 Заповнення актантів лексичних об'єктів "спеціальна ознака".
  - 7.1.1 Застосування граматичних правил.
  - 7.1.2 Застосування семантичних правил.
- 7.2 Заповнення актантів сутностей.
  - 7.2.1 Застосування граматичних правил.
  - 7.2.2 Застосування семантичних правил.
- 7.3 Заповнення актантів дій.
  - 7.3.1 Застосування граматичних правил.
  - 7.3.2 Застосування семантичних правил.
- 7.4 Визначення предикації.

Крок 8 Для всіх речень вхідного тексту:
- 8.1 Визначення семантичних суб'єкта і об'єкта речення.
- 8.2 Визначення прагматичного змісту.

Крок 9 Визначення семантичних суб'єкта і об'єкта тексту.
Крок 10 Для всіх речень вхідного тексту:
- 10.1 Визначення анафоричних посилань.
- 10.2 Для кожного виділеного посилання:
  - 10.2.1 Визначення області пошуку антецедента.
  - 10.2.2 Застосування граматичних правил.
  - 10.2.3 Застосування семантичних правил.
- 10.3 Визначення властивостей речення.

Крок 11 Визначення властивостей тексту.

Тепер розглянемо детальніше окремі кроки алгоритму.

1. Розділення тексту на слова проводиться розбиттям речення на частини в місцях розташування знаку "пробіл". Отримані частини вважаються словами. Заповнення рівня слів структури внутрішньої репрезентації знайденими в тексті об'єктами-словами (заповнюються атрибути "порядковий номер слова в тексті", "ім'я" – рядок символів з тексту, що є словом, без пробілів з можливим кінцевим розділовим знаком).

2. Для кожного слова (кожного об'єкта рівня слова) пошук і аналіз словникових статей у БЗ. Якщо для слова внаслідок відпрацювання запиту до семантичної мережі видано більше одного об’єкта (випадок неоднозначності) прочитати всі статті із заголовком, рівним початковій формі значення атрибута "ім'я" відповідного об'єкта. Запам'ятати граматичну, семантичну і прагматичну інформацію про слово з БЗ у відповідних атрибутах об'єкта-слова. Для слів з неоднозначністю прийняти за правильний варіант той об'єкт, який має велику імовірність використання, а при рівних імовірностях або їх відсутності – перший за списком об'єкт, знайдений в БЗ. Випадок відсутності в словнику слів, присутніх в тексті, пошуку словарної статті для слів, що введені в тексті з помилками, та інші питання надійності інтерфейсу є темою інших публікацій і не є предметом даної статті.

3. При об’єднанні слів в речення спочатку будемо вважати, що всі речення складаються з однієї простої складової, і послідовно перебираємо всі об'єкти-слова, виділені в тексті. Заповнення рівня простої складової.

4. Для кожного речення:

4.1. Визначення типу (повідомлення/запит) проводиться простою перевіркою знаку кінця речення. Тип повідомлення (первинна інформація, розширення інформованості, команда) визначається на етапах 8.2 (визначення прагматичного змісту речення), 10.3 (визначення властивостей речення).

4.2. Розділення на прості складові становить собою заповнення рівня речення. Етап повністю може бути проведений засобами граматичного аналізу рівня простих складових, але вхідною інформацією для цього етапу є характеристики слів, визначення яких може вимагати не тільки граматичного, але і семантичного аналізу. Такими характеристиками є частини мови кожного слова ПМ. Приклад граматичного правила для визначення простої складової: позначимо через

$$S = \{W, W_{pb}, W_{pe}\} \tag{1}$$

множину слів в реченні S, де W_{pb} – множина, що складається з першого слова речення та слів, що знаходяться безпосередньою після знаків "кома", "крапка з комою", "двокрапка" або "тире"; W_{pe} – множина, що складається з останнього слова речення та слів, які знаходяться безпосередньо перед знаками "кома", "крапка з комою", "двокрапка" або "тире"; W – множина, що складається зі слів, які не потрапили до жодної з множин W_{pb}, W_{pe}. Тут і далі позначатимемо прописними літерами W з індексом чи без нього підмножину слів речення, рядковим символом w з індексом – одне будь-яке слово речення.

Через

$$\begin{aligned} &S_{sub} = \{\{W_{sub} \subset W\}, \{w_{pb} \in W_{pb}\}, \{w_{pe} \in W_{pe}\}\}: \\ &w_{pb}.index < (w_{sub} \in W_{sub}).index < w_{pe}.index \wedge \\ &|W_{sub}| = w_{pe}.index - w_{pb}.index - 1 \end{aligned} \tag{2}$$

позначимо множину простих складових речення, де *index* – це атрибут, значення якого є порядковим номером слова в реченні.

Тоді граматичне правило, що дозволяє визначити, чи є підмножина слів S_{subk} простою складовою речення, формулюється наступним чином:

$$\begin{aligned} &S_{subk} = \{w_b \in W_{pb}, W_m \subset W, w_e \in W_{pe}\} \subset S_{sub}: \\ &\left(\begin{aligned} &\exists w_x \in W_m : w_x.sm = 'instance' \wedge \exists w_y \in W_m : w_y.sm = 'action' \wedge \\ &\left(\begin{aligned} &W_k \subset S_{sub} : \forall (w_k \in W_k) w_k \notin S_{subk} \\ &\vee \exists w_x \in S \setminus S_{subk} : w_x.sm = 'instance' \wedge \exists w_y \in W_m : w_y.sm = 'action' \end{aligned} \right) \end{aligned} \right). \end{aligned} \tag{3}$$

В (3) через *sm* позначений атрибут "семантична характеристика", значення якого 'instatnce' мають об'єкти-слова, що виражають в даному реченні сутності, а 'action' – об'єкти-слова, що виражають дії.

Тобто, підмножина розташованих одне за одним слів речення є простою складовою, якщо: 1) перший член підмножини є першим словом речення або словом, перед яким стоїть кома, крапка з комою, тире або двокрапка; 2) останній член підмножини є останнім словом речення або словом, після якого стоїть кома, крапка з комою, тире або двокрапка; 3) серед слів підмножини, що аналізується, знайдуться два слова, одне з яких виражає сутність (в синтаксичних термінах іменник чи займенник), а друге виражає дію; 4) в реченні вже виділена хоча б одна проста складова або серед решти підмножин, перший член яких є першим словом речення або словом, перед яким стоїть кома, крапка з комою, тире або двокрапка, а останній член підмножини є останнім словом речення або словом, після якого стоїть кома, крапка з комою, тире або двокрапка, знайдуться два слова, одне з яких виражає сутність або є займенником, а друге виражає дію.

4.3. Можливі значення типів зв'язків простих складових складного речення: підрядна і сурядна. Для підрядних зв'язків визначаються їх граматичні значення: визначальне, обставинне, місця, часу, умови, причини, мети, наслідку, способу дії, міри і ступеня, порівняння, поступки. Прикладом правила визначення типу зв'язків простих складових є наступне правило: нехай

$$S = \{S_{sub1}, S_{sub2}, ..., S_{subn}\}, S_{sub2} = \{w_k, ..., w_m\} -$$

речення, яке складається з n простих складових, а друга проста складова містить слова речення з порядковими номерами від k до m. Тоді правило виглядатиме наступним чином:

$$r\_subord(S_{sub1}, S_{sub2}), S_{sub2}.r\_type = 'place': w_k.name = 'де', \tag{4}$$

в якому через *r_subord* позначено відношення підпорядкованості простих складових речення, де перший аргумент – підпорядковуюча складова, другий аргумент – підпорядкована; *r_type* – атрибут "граматичне значення підрядного зв'язку"; *name* – атрибут "ім'я", що фактично є записом слова в даному реченні.

Наведене правило використовується для визначення складнопідрядного зв'язку двох простих складових, граматичне значення якого є "*place*" – підрядний зв'язок місця.

5. Об'єднання речень в абзаци – це процес заповнення рівня абзацу. Етап не є обов'язковим, але може бути корисний при визначенні семантичних суб'єкта і об'єкта вхідного тексту і розв'язанні анафоричних посилань. Проводиться простим розбиттям тексту в місцях розташування символу абзацу.

6. Об'єднання абзаців в текст являє собою заповнення рівня тексту. Якщо текст не був розбитий на абзаци, текст вважається складеним з одного абзацу, а всі речення – такими, що входять в один абзац.

7. На даному етапі в кожному простому реченні і кожній простій складовій проводиться з'єднання всіх виділених об'єктів в смислові групи, для чого визначаються актанти лексичних об'єктів "спеціальна ознака", "сутність" і "дія". Тому важливий саме такий порядок визначення актантів, оскільки будь-яка спеціальна ознака є кандидатом на заповнення актанта сутності, а будь-яка сутність є кандидатом на заповнення актанта дії. Внаслідок виконання етапу всі об'єкти рівня слова, що належать даному простому реченню або простій складовій, повинні заповнити конкретну актантну позицію одного з об'єктів-дій.

7.1. Для заповнення актантів лексичних об'єктів "спеціальна ознака" даного речення прово-диться аналіз всіх об'єктів рівня слова, що належать даному реченню, атрибут "частина мови" яких дорівнює "прикметник", "дієприкметник", на наявність актантів. Кандидати на заповнення актантних позицій об'єктів "ознака" – об'єкти рівня слів, атрибут "частина мови" яких дорівнює "прислівник", "дієприслівник", "сутність" або "займенник".

7.1.1. При застосуванні граматичних правил визначаємо кандидатів на заповнення актантів. Наведемо приклад граматичного правила для відбору кандидатів на заповнення позиції "загальна ознака". Нехай проста складова або просте речення, що розглядається, складається з *m* слів (5):

$$S_{sub} = \{w_1,...,w_k...,w_m\}, \tag{5}$$

де *k*-те слово виражає об'єкт "спеціальна ознака", для якого треба знайти кандидати на заповнення актантної позиції. Правило заповнення можна записати, використовуючи функцію умови *iif*:

$$\begin{aligned} &iif\,(\exists w_p : (w_k.index+1 = w_p.index) \wedge (w_k[len(w_k)] \neq ',') \vee \\ &(w_k.index = w_p.index+1) \wedge (w_p[len(w_p)] = ','), w_k.gen\_det = w_p, \\ &w_k.gen\_det\_list.add(w_p : (w_k.index > w_p.index) \wedge (w_p[len(w_p)] \neq ','))) \end{aligned} \tag{6}$$

Тобто, якщо серед слів простого речення або складової, що розглядається, знайдеться слово w_p, що задовольняє умови, записані в першому аргументі функції *iif*, актантна позиція "загальна ознака" (*gen_det*) заповнюється цим словом, інакше заповнюємо список кандидатів на заповнення цієї ознаки (*gen_det_list*) словами, що задовольняють умову, записану в третьому аргументі функції *iif*. Через *len* позначено функцію довжини довільного рядка символів.

7.1.2. Семантичні правила застосовуються до виділених на попередньому кроці кандидатів. Актантну позицію займе кандидат, що відповідає семантичним правилам. Для прикладу знаходження актанта на позицію "загальна ознака" (див. етап 7.1.1) наведемо семантичне правило визначення актанта серед виділених кандидатів. Нехай

$$w_p \in w_k.gen\_det\_list\,. \tag{7}$$

Тоді

$$\begin{aligned} &w_k.gen\_det = w_p : \\ &\neg\exists w_a, i : (w_a.SPart = 'N' \wedge w_a.act\_pos[i] = w_k \wedge \\ &\exists w_b, i : w_b.SPart = 'N' \wedge w_b.act\_pos[i] = w_a). \end{aligned} \tag{8}$$

Тобто загальною ознакою спеціальної ознаки w_k буде такий лексичний об'єкт w_p, який не є ознакою сутності, що заповнила актантну позицію другої сутності. В (8) через *SPart* позначений атрибут "частина мови", значення '*N*' якого мають слова-сутності (іменники та займенники), через *act_pos*[*i*] позначений елемент списку актантних позицій лексичного об'єкта.

7.2. Для заповнення актантів сутностей проводиться аналіз всіх об'єктів рівня слова, які належать об'єкту рівня простого речення, що аналізується, або простої складової, атрибут "частина мови" яких дорівнює "іменник" або "займенник". Кандидати на заповнення актантних позицій сутності – лексичні об'єкти "спеціальна ознака" (прикметник, дієприкметник), "узагальнена ознака" (прислівник, дієприслівник) і "сутність" (іменник, займенник) разом із з'єднувачем – об'єктом рівня слова, атрибут "частина мови" якого дорівнює "прийменник", – що безпосередньо стоїть в реченні перед самою сутністю або її актантом, що має найменший порядковий номер в реченні або простій складовій.

7.2.1. Застосуванням граматичних правил визначаються кандидати на заповнення актант-них позицій сутностей. Вигляд граматичних правил для даного етапу аналогічний наведеному прикладу граматичного правила для визначення актантів лексичних об'єктів "спеціальна ознака" (див. опис етапу 7.1.1).

7.2.2. Семантичні правила застосовуються до виділених на попередньому кроці кандидатів. Актантну позицію займе кандидат, що відповідає семантичним правилам. Вигляд семантичних правил для даного етапу аналогічний наведеному прикладу семантичного правила для визначення актантів лексичних об'єктів "спеціальна ознака" (див. опис етапу 7.1.2).

7.3. Для заповнення актантів дій проводимо аналіз всіх об'єктів рівня слова, атрибут "частина мови" яких дорівнює "дієслово". Кандидати на заповнення актантних позицій дій – будь-який об'єкт рівня слова, не приєднаний на попередніх етапах кроку 7 до жодного іншого об'єкта.

7.3.1. Застосуванням граматичних правил визначаються кандидати на заповнення актантних позицій. Вигляд граматичних правил для даного етапу аналогічний наведеному прикладу граматичного правила для визначення актантів лексичних об’єктів "спеціальна ознака" (див. опис етапу 7.1.1).

7.3.2. Семантичні правила застосовуються до виділених на попередньому кроці кандидатів. Актантну позицію займе кандидат, що відповідає семантичним правилам. Вигляд семантичних правил для даного етапу аналогічний наведеному прикладу семантичного правила для визначення актантів лексичних об’єктів "спеціальна ознака" (див. опис етапу 7.1.2).

7.4. Визначення предикації дій в простих реченнях або простих складових: таксоно-метрична, реляційна, характеризуюча, оцінювальна, екзистенційна. Предикація визначається шляхом перевірки в БЗ атрибута "предикація" вказаної дії або, якщо цей атрибут дорівнює *null*, найближчого пращура дії, що має цей атрибут заповненим. Детально структура БЗ для оброблення текстів ПМ в автоматизованих системах управління викладена в працях, які знаходяться на стадії опублікування.

8. Для всіх речень вхідного тексту:

8.1. На кроці визначення семантичних суб'єкта і об'єкта речень визначається локальна тема – тема речення, яка в термінах синтаксису є її граматичною основою. Тема – це семантичний суб'єкт. Семантичним суб’єктом є агенс основної дії простого речення чи основної дії підпорядковуючої простої складової. Для складносурядного речення семантичним суб’єктом буде множина об’єктів, оскільки в цьому випадку може бути вказано дві чи більше рівноправні дії. Семантичний об'єкт визначається розбором об'єктів-сутностей, що заповнили одну з актантних позицій дії, агенсом якого є семантичний суб'єкт речення.

8.2. Прагматичний зміст речення визначається за модальними словами, що присутні в реченні, і може мати значення інформативності, спонукання до дій (прохання, вимога, наказ), розширення інформованості адресата, зміни стану. Проблема визначення прагматичного змісту є досить складною, тому ми залишаємо її розв’язання для подальших праць.

9. Визначення семантичних суб'єкта і об'єкта тексту – це виділення теми тексту. Дана проблема описується в ряді публікацій, одна з найсучасніших – робота [7].

10. Даний етап є логічним продовженням етапу 8. Таке розбиття етапу розбору рівня речення на дві частини пояснюється необхідністю застосування результатів етапу визначення семантичних суб'єкта і об'єкта тексту для наступних кроків розбору рівня речення.

10.1. Анафоричними посиланнями є всі актантні позиції, заповнені займенниками і займенниковими словами (займенники і абстрактні слова типу "питання", "предмет" тощо [4]). Оскільки займенникові слова можуть і не бути анафоричними посиланнями, то потрібне застосовування правил визначення анафоричного посилання. Одним з таких правил може бути твердження, що займенникове слово є посиланням, якщо у нього не заповнені актантні позиції уточнюючого типу (належність тощо).

10.2. На даному кроці проводимо визначення об'єктів, на які вказують посилання, виявлені в реченні.

10.2.1. Визначення області пошуку антецедента проводиться за допомогою методу розповсюдження позначок тексту, описаного в [8].

10.2.2. Застосуванням граматичних правил визначаються антецеденти-кандидати.

10.2.3. Семантичні правила застосовуються до виділених антецедентів-кандидатів. Граматичні та семантичні правила для анафоричного аналізу викладені в [6].

10.3. Визначення властивостей речення на основі семантичних властивостей слів: узагаль-нена модальність, узагальнений прагматичний зміст, узагальнене експресивно-стилістичне забарвлення, нормативні наслідки, персональність.

11. Визначення властивостей тексту проводиться на основі виділених властивостей речень. Характеристика та правила визначення властивостей речень та тексту, частина яких описана в [5], є предметом подальших досліджень.

4. Реалізація і властивості алгоритму

Поданий в попередньому розділі алгоритм є загальною структурою аналізу тексту ПМ. В ній описані кроки, які необхідно реалізувати в програмному коді розбору вхідної множини природномовних речень. Конкретна реалізація даної структури пов’язана з рядом проблем, серед яких ітеративний характер алгоритму (необхідність переходу на попередні кроки у випадку неоднозначності), великий об’єм БЗ, який потребує особливого підходу, складна структура самої внутрішньої репрезентації. Крім того, реалізація даного алгоритму передбачає розробку та застосування спеціальних правил співвідношення об’єктів різних систем моделювання природномовних речень і текстів. Склад таких правил безпосередньо залежить від загальної кількості речень природної мови, що можуть породжуватись. Оскільки відомо, що природною мовою може бути сформована нескінченна кількість речень, то і складність перебору правил не може бути однозначно оціненою.

Основна робота алгоритму аналізу тексту ПМ сконцентрована на рівні речень. Отже, оцінимо складність розбору речення даного алгоритму в залежності від кількості слів, а точніше від кількості

слів, представлених частинами речення. Рівень речення в алгоритмі представлений кроками 7, 8, 10 алгоритму.

В найважчому випадку кількість операцій аналізу на кроці 7.1 (заповнення актантів об'єктів "спеціальна ознака") дорівнює:

$$C_{7.1} = \sum_{i=1}^{m} \left(\left(n_{ADJi} + n_{VADJi} \right) \times \left(n_{ADVi} + n_{VADVi} + n_{Ni} + n_{PRONi} \right) \right), \tag{9}$$

де $C_{7.1}$ – складність кроку 7.1; m – кількість простих складових речення (для простого речення m = 1); n_{ADJi}, n_{VADJi}, n_{ADVi}, n_{VADVi}, n_{Ni}, n_{PRONi} – відповідно кількість прикметників, дієприк-метників, прислівників, дієприслівників, іменників та займенників i-ї простої складової.

Даний випадок є найважчим, оскільки, якщо одна із спеціальних ознак на перших кроках буде пов'язана за якимось однозначним правилом з одним з об'єктів речення, то пов'язані об'єкти більше не розглядаються, і кількість операцій кроку зменшуватиметься в кожному такому випадку.

Кількість операцій аналізу на кроці 7.2 (заповнення актантів сутностей):

$$C_{7.2} = \sum_{i=1}^{m} \left(\left(n_{Ni} + n_{PRONi} \right) \times \left(n_{ADJi} + n_{VADJi} + n_{Ni} + n_{PRONi} - 1 \right) \right). \tag{10}$$

Кількість операцій аналізу на кроці 7.3 (заповнення актантів дій):

$$C_{7.3} = \sum_{i=1}^{m} \left(n_{Vi} \times \left(n_{NONAi} \right) \right), \tag{11}$$

де n_{Vi} – кількість дій; n_{NONAi} – кількість не приєднаних на попередніх етапах кроку 7 слів i-ї простої складової.

Кількість операцій аналізу при отриманні семантичного суб'єкта на кроці 8.1 дорівнює:

$$C_{8.1subj} = \sum_{i=1}^{m} n_{Ag}\ , \tag{12}$$

де m – кількість простих складових в складносурядному реченні або, якщо речення складнопідрядне, кількість підпорядковуючих речень (для простого речення m = 1); n_{Ag} – кількість слів, що заповнили актантну позицію дії "агенс".

Кількість операцій аналізу при отриманні семантичного об'єкта на кроці 8.1 дорівнює:

$$C_{8.1obj} = \sum_{i=1}^{k} n_{AP}\ , \tag{13}$$

де k – кількість виділених семантичних суб'єктів; n_{AP} – кількість актантних позицій дії, агенс якої є семантичним суб'єктом.

Кількість операцій аналізу на кроці 8.2 визначається кількістю модальних слів у реченні.

Складність кроку 10 дорівнює:

$$C_{10} = \left(n_{PRON} + n_{WPRON} \right) \times \left(n_{N-} \right), \tag{13a}$$

де n_{WPRON} – кількість займенникових слів [4]; n_{N-} – кількість сутностей, що не є займенниковими словами в реченні.

Загальна кількість операцій аналізу природномовного речення дорівнює:

$$C_S = C_{7.1} + C_{7.2} + C_{7.3} + C_{8.1subj} + C_{8.1obj} + C_{8.2} + C_{10}\ . \tag{14}$$

Таким чином, якщо не враховувати наявність правил міжмодельного співвідношення об'єктів та кількість повернень на попередні кроки при розв'язанні неоднозначностей, оцінка складності алгоритму виражена поліноміальною функцією.

Оцінимо тільки кількість повернень при розборі речення. Позначимо через n_N кількість сутностей, n_{adj} – кількість спеціальних ознак, n_{adv} – кількість загальних ознак в реченні, а через r_N, r_{adj}, r_{adv} – кількості можливих атрибутів в структурі розуміння, які можуть бути заповнені сутністю, спеціальною ознакою чи загальною ознакою. Тоді в найважчому випадку кількість повернень дорівнює:

$$C_{Back} = r_N^{n_N} + r_{adj}^{n_{adj}} + r_{adv}^{n_{adv}}\ , \tag{15}$$

що свідчить про досить неприємну експоненційну оцінку алгоритму. Простий підрахунок для сутностей, які є найбільш поширеними елементами речення (воно містить в середньому від п'яти сутностей), дозволяє відчути характер оцінки.

Крім того, при оцінці складності алгоритму необхідно використовувати коефіцієнт, що враховує здатність одного слова виступати в реченні різними частинами мови. Цей коефіцієнт залежить від конкретної природної мови. Для української чи російської мови він буде досить невеликим порівняно, наприклад, з англійською мовою, в якій майже кожне слово може бути як мінімум двома частинами мови.

Отже виникає важлива проблема підвищення ефективності роботи алгоритму, розв'язання якої вимагає розроблення і використання в процесі аналізу спеціальних правил, які дозволяли б вилучити в процесі його роботи заздалегідь некорисні кроки.

Висновки

Загальною метою наших досліджень є автоматизація обробки текстової інформації і синтезу знань та створення системи аналізу та синтезу текстів ПМ на основі моделей розуміння. Така система, основні складові якої – підсистеми аналізу, смислової інтерпретації та вербалізації, має бути посередником між автоматизованою системою управління та користувачем, підтримуючи діалог на ПМ.

Запропоновано алгоритм, який розроблений для застосування в підсистемі аналізу текстової інформації, розглянуті його властивості та особливості реалізації.

Розвиток досліджень пов'язаний з уточненням структури БЗ і внутрішньої репрезентації тексту, розробленням більш ефективних процедур для окремих його кроків, побудовою системи правил, застосування яких дозволить суттєво прискорити аналіз, тощо.

ЛІТЕРАТУРА:

1. Новое в зарубежной лингвистике: Вып. 24. Компьютерная лингвистика: Пер. с англ. / Под ред. Б.Ю. Городецкого. – М.: Прогресс, 1989. – 432 с.
2. Новое в зарубежной лингвистике: Вып. 23. Когнитивные аспекты языка: Пер. с англ. / Под ред. В.В. Петрова. – М.: Прогресс, 1988. – 430 с.
3. Моделирование языковой деятельности в интеллектуальных системах / Под ред. А.Е. Кибрика, А.С. Нариньяни. – М.: Наука, 1987. – 280 с.
4. *Семенова С.Ю.* Семантические поля словаря РОСС: опыт заполнения, анализ дескриптивных возможностей (Материалы к унификации словарных описаний) // Труды Международного семинара "Діалог 2000" по компьютерной лингвистике и ее приложениям. – http://www.dialog-21.ru
5. *Теленик С.Ф., Сичная А.А., Сичной А.Н.* Естественно языковый интерфейс в адаптивной технологии // Проблемы программирования. – 1999. – № 1. – С. 118–129.
6. *Carbonell J.G., Brown R.D.* Anaphora Resolution: A Multi-Strategy Approach // In proceedings of 12th International Conf. of Computational Linguistics (COLING'88), Budapest (Hungary). – P. 96–101.
7. *Kitani T., Eriguchi Y., Hara M.* Pattern Matching and Discouce Processing in Information Extraction from Japanese Texts // Journal of Artificial Intelligence Research. – 1994. – V. 2. – P. 89–110.
8. *Palomar M., Martinez-Barco P.* Computational Approach to Anaphora Resolution in Spanish Dialogues // Journal of Artificial Intelligence Research. – 2001. – V. 15. – P. 263–287.

ТЕЛЕНИК Сергій Федорович – доктор технічних наук, професор, завідувач кафедри автоматики і управління в технічних системах Національного технічного університету України "Київський політехнічний інститут".

Наукові інтереси:
– штучний інтелект;
– автоматизація програмування та проектування;
– математична логіка;
– комп'ютерна лінгвістика.

Тел.: (044) 236-42-99, 241-70-39.

E-mail: telenik@acts.ntu-kpi.edu.ua

СМІЧИК Руслана Вадимівна – аспірант кафедри автоматики і управління в технічних системах Національного технічного університету України "Київський політехнічний інститут".

Наукові інтереси:
– штучний інтелект;
– комп'ютерна лінгвістика.

E-mail: lana_smichyk@ukr.net

Подано 10.10.2002

С.Ф. Теленик, Р.В. Смичик

Обработка текстов естественного языка на основе моделей понимания в автоматизированных системах управления

Статья посвящена проблеме разработки естественно-языкового интерфейса в автоматизированных системах управления. Предложен алгоритм анализа текстовой информации, в котором моделируется человеческое понимание естественноязыковых высказываний. Дана концепция и характеристика алгоритма, указаны направления дальнейшей работы в области обработки текстов на основе моделей понимания.

S.F. Telenyk, R.V. Smichyk

Natural Language Processing Based on Understanding Models for Computer-Aided Management Systems

The subject of this paper is the problem of developing natural language interface in computer-aided management systems. An algorithm of written text analysis has been proposed which emulates human understanding of natural language utterances. We have introduced the concept and the description of the algorithm and also the directions for further work have been pointed out.