7С

УДК 004.82, 004.434:004.94

О.П. Сирота
НТУУ «КПІ», м. Київ, Україна, sirota@ukrpost.com.ua

Обмежувально-продукційний метод подання знань для аналізу несуперечності текстів

У статті розглядається проблема автоматизованого аналізу несуперечності текстів та відповідності текстів нормативній базі. Пропонується здійснювати аналіз несуперечності тексту шляхом логічного аналізу із залученням знань про предметну область, природну мову та нормативну базу. Пропонується обмежувально-продукційний метод подання знань та розглядається його застосування для подання логічного еквівалента тексту та моделювання предметної області. Пропонується підхід до моделювання предметної області, який дозволяє здійснити аналіз несуперечності тексту. Зокрема цей підхід передбачає включення нормативної бази у формальний опис предметної області, а також створення лінгвістично-орієнтованої бази знань предметної області, що дозволяє виконувати аналіз семантичної правильності побудови тексту та аналіз відповідності тексту нормативній базі.

Вступ

Процес створення несуперечливих текстових документів, узгоджених із нормативною базою, потребує автоматизації при великому обсязі нормативної бази, великих обсягах та великій кількості текстових документів. Автоматизація цього процесу здатна знизити час підготовки текстових документів, а також підвищити їхню якість. Фактично ця задача означає автоматизований аналіз несуперечності тексту та аналіз відповідності тексту нормативній базі.

Метою роботи є розробка методу аналізу несуперечності текстів та відповідності текстів нормативній базі.

Суперечність текстів означає, що висловлення в тексті не відповідають законам нормативної бази, текст містить висловлення та його заперечення, тексти подають взаємовиключні факти, текст містить невірно побудовані висловлення. Питання суперечності текстів розглядається в лінгвістичних дослідженнях [1], [2].

Аналіз несуперечності тексту потребує залучення знань з предметної області, нормативної бази та природної мови, а також потребує виконання логічного аналізу тексту, тобто побудови його логічного еквівалента. Логічний аналіз тексту розглядається в [2], [3].

Таким чином, аналіз несуперечності тексту є аналізом несуперечності логічного еквівалента тексту із залученням формально поданих предметної області, природної мови, нормативної бази.

Зрозуміло, що для подання формального опису предметної області, мови, нормативної бази та для побудови логічного еквівалента тексту повинна використовуватись одна й та сама модель подання знань. Використання природної мови, залежність тексту від певного контексту, необхідність підключення до тексту нормативної бази висувають певні вимоги до формальної моделі подання знань, а саме:

– розширюваність, тобто можливість подати нові типи знань без зміни моделі знань;

– інтеграція різних формальних моделей подання знань, а саме: формальної логічної [4], продукційної [4] моделей, об'єктного підходу [5];
– повторне використання знань, що дозволить побудувати формальний опис нормативної бази, предметної області, які можна буде спільно та повторно використовувати.

Для подання формального опису предметної області та побудови логічного еквівалента тексту пропонується використати обмежувально-продукційний метод подання знань, який задовольняє всім вищеназваним вимогам.

У статті розглядаються: обмежувально-продукційний метод подання знань; підхід до моделювання предметної області, який дозволить здійснити аналіз несуперечності текстів; моделювання предметної області та логічного еквівалента тексту за допомогою обмежувально-продукційного методу. Показано, яким чином здійснюється аналіз несуперечності текстів та аналіз відповідності тексту нормативній базі.

Обмежувально-продукційний метод подання знань

Обмежувально-продукційний (*constraint-production*, *CP*) метод подання знань, надалі *CP*-метод, є модифікацією модельного методу подання знань [6]. На відміну від модельного методу, *CP*-метод розширяється можливістю завдання окремого механізму виводу для кожної моделі знань.

CP-метод дозволяє реалізувати різноманітні мови подання знань, контролювати правильність побудови знань та надати для кожної мови подання знань окремий механізм виводу.

CP-метод подання знань визначається наступними складовими:
– універсальні мови для подання знань;
– *CP*-моделі знань.

Універсальні мови складаються зі слів, побудованих за певним алфавітом, при цьому словами в такій мові будуть не довільні ланцюги символів алфавіту, а трійки символів. Такий підхід дозволяє використати ідеї семантичних мереж для подання знань у вигляді слів універсальної мови.

Нехай Ψ – множина всіх можливих символів об'єктів, призначених для подання об'єктів предметної області та мови. Нехай LIT – множина строкових рядків; N – множина натуральних чисел, R – множина дійсних чисел. Нехай $\Omega=\Psi \bigcup LIT \bigcup N \bigcup R$. Будемо називати Ω множиною всіх можливих алфавітних символів.

Визначення. *Алфавітом* є довільна не порожня підмножина множини Ω.

Визначення. Нехай Σ – алфавіт, $\Sigma \subseteq \Omega$. Будь-яку трійку з множини $\Sigma\times\Sigma\times\Sigma$ будемо називати *словом* в алфавіті Σ.

Визначення. Σ^* – множина всіх слів в алфавіті Σ.

Визначення. *Універсальною мовою L* в алфавіті Σ буде довільна підмножина множини Σ^*.

Уведемо клас функцій, які будемо називати *обмежувальними*. Призначення обмежувальної функції – визначити, чи задовольняє слово універсальної мови деяким обмеженням.

Визначення. Будь-яку функцію $\chi : 2^{\Omega^*}\times\Omega^* \to \{1, 0\}$ будемо називати *обмежувальною*. Будемо казати, що обмежувальна функція χ допускає слово ω

мови *L,* якщо $\chi(L, \omega)=1$. Обмежувальна функція χ не допускає слово ω мови *L,* якщо $\chi(L, \omega)=0$ або не визначено.

Для створення нових слів уводиться клас породжуючих функцій, які породжують по вхідній мові новий набір слів.

Визначення. Будь-яку функцію p: $2^{\Omega^*} \to 2^{\Omega^*}$ будемо називати *породжуючою*.

Ключовим поняттям *СР*-методу подання знань є *СР*-модель знань.

Визначення. *СР-моделлю знань M* з назвою μ буде шістка (μ, E, W, χ, p, A), де

- μ – назва моделі;
- E – скінченна множина символів, $E \subset \Psi$;
- W – множина слів;
- χ – обмежувальна функція;
- p – породжуюча функція;
- A – множина назв безпосередньо приєднаних моделей.

СР-модель M задає *повну множину символів* $\overline{E}$, *повну множину слів* $\overline{W}$, *повну обмежувальну функцію* $\overline{\chi}$. Ці поняття вводяться через *відношення приєднання*.

Функція p визначена на множині $\overline{W}$, $p(\overline{W}) \subset \overline{W}$.

Нехай П – множина всіх *СР*-моделей.

Визначення. На множині П виділяють *відношення приєднання* $\pi \subset \Pi \times \Pi$. Нехай $M=(\mu, E, W, \chi, p, A)$, $M_1=(\mu_1, E_1, W_1, \chi_1, p_1, A_1)$. $(M, M_1) \in \pi$ означає, що модель M *приєднує* модель M_1. $(M, M_1) \in \pi$, якщо $\mu_1 \in A$. Відношення π є ациклічним.

Нехай π^+ – транзитивне замикання відношення π.

Визначення. *Повна множина символів* $\overline{E}$ моделі M – це об'єднання множини символів моделі із множинами символів приєднаних моделей.

$\overline{E} = E \cup (\bigcup_{(M, M_1) \in \pi} E_1)$, де E_1 – множина символів моделі M_1.

Нехай $\Sigma = \overline{E} \cup LIT \cup N \cup R$ – алфавіт.

Визначення. Множина слів W *СР*-моделі M є мовою в алфавіті Σ, $W \subseteq \Sigma^*$.

Визначення. *Повна множина слів* $\overline{W}$ моделі M – це об'єднання множини слів моделі із множинами слів транзитивно приєднаних моделей.

$\overline{W} = W \cup (\bigcup_{(M, M_1) \in \pi^+} W_1)$, де W_1 – множина слів моделі M_1.

Визначення. *Повна обмежувальна функція* $\overline{\chi}$ для моделі M – це обмежувальна функція, яка допускає слово, якщо його допускає обмежувальна функція моделі M та обмежувальна функція кожної транзитивно приєднаної моделі.

$\overline{\chi}(L, \omega) = \chi(L, \omega) \wedge \prod_{(M, M_1) \in \pi^+} \chi_1(L, \omega)$, де χ_1 – обмежувальна функція моделі M_1.

Визначення. *СР*-модель знань M є *цілісною*, якщо повна множина слів моделі допускається повною обмежувальною функцією, тобто $\forall \omega \in \overline{W}$ $\overline{\chi}(\overline{W}, \omega)=1$.

СР-метод подання знань дозволяє використати формальну логічну, продукційну моделі подання знань, об'єктний підхід до подання знань та їх інтеграцію. Для цього необхідно визначити основні *СР*-моделі знань – функціональну, продукційну, аксіоматичну, об'єктну, а також модель логічних операторів. Кожна

модель уводить відповідні символи, слова, обмежувальну та продукційну функції. Призначення кожної моделі наведено в таблиці 1. Взаємозв'язок основних *СР*-моделей показаний на рис. 1.

Таблиця 1 – Призначення основних СР-моделей знань

***СР*-модель**	**Призначення**
Функціональна	Забезпечує використання відношень, предикатів тощо. Контролює правильність побудови термів.
СР-модель логічних операторів	Забезпечує використання логічних операторів. Не допускає формулу та її заперечення.
Продукційна	Забезпечує прямий та зворотний логічний висновок. Не допускає слова, які попадають у зону дії зворотного правила, якщо не виконуються необхідні умови.
Аксіоматична	Забезпечує завдання тверджень, що завжди істинні у предметній області. Забезпечує виконання аксіом.
Об'єктна	Забезпечує використання класів та екземплярів. Контролює правильність побудови об'єктних визначень.

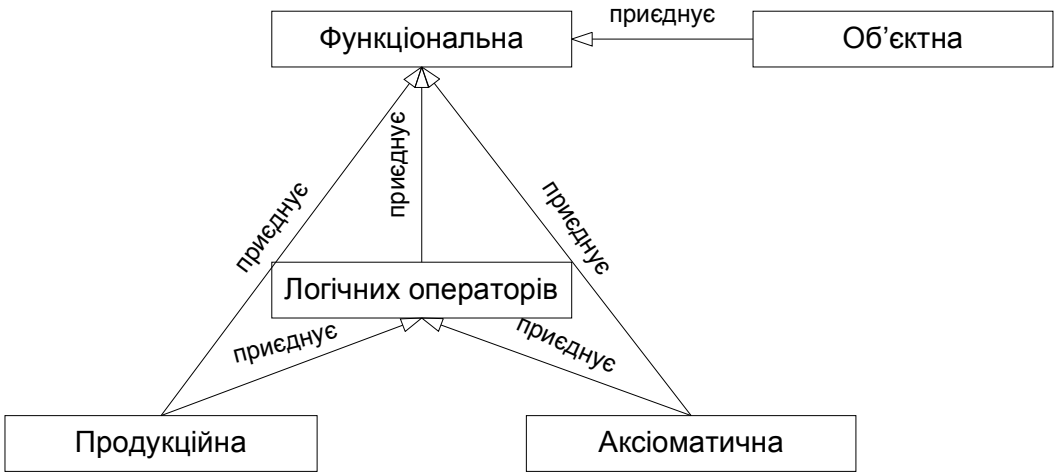

Рисунок 1 – Взаємозв'язок основних СР-моделей знань

Використання *СР*-методу для моделювання конкретних предметних областей полягає у створенні нової *СР*-моделі, приєднанні до неї необхідних основних *СР*-моделей та наповненні її знаннями.

У таблиці 2 показано, приєднання яких *СР*-моделей дозволяє використати різні формальні моделі подання знань та їх інтеграцію.

Таблиця 2 – Формальні моделі подання знань та СР-моделі

Формальна модель подання знань	***СР*-моделі**
формальна логічна	функціональна, логічних операторів, аксіоматична, продукційна
продукційна	продукційна
об'єктний підхід	об'єктна
об'єктний підхід, продукційна, використання аксіом	продукційна, об'єктна, аксіоматична

Аналіз несуперечності текстів

Аналіз несуперечності тексту пропонується виконувати шляхом аналізу несуперечності логічного еквіваленту тексту. Для цього використаємо бази знань (БЗ), які дозволяють відстежити зазначену несуперечність тексту.

Будемо називати *онтологією предметної області* базу знань, яка містить знання про предметну область.

Онтологія предметної області може бути подана як сукупність

- *концептуальної онтології предметної області,* тобто БЗ, яка містить поняття та відношення між ними;
- *нормативної онтології предметної області*, тобто БЗ, яка містить правила та закони взаємодії понять предметної області.

Це дозволяє визначити єдину концептуальну онтологію предметної області та приєднувати до неї різні нормативні онтології.

Для подання формального опису предметної області за допомогою *СР*-методу подання знань створюються наступні *СР*-моделі знань:

- $M_{DOMAIN\_CONCEPT}$ – *СР*-модель концептуальної онтології предметної області;
- $M_{DOMAIN\_NORMATIVE}$ – *СР*-модель нормативної онтології предметної області;
- M_{DOMAIN} – *СР*-модель онтології предметної області.

Взаємозв'язок зазначених *СР*-моделей поданий на рис. 2. Для їх побудови використовуються об'єктна, продукційна та аксіоматична *СР*-моделі знань.

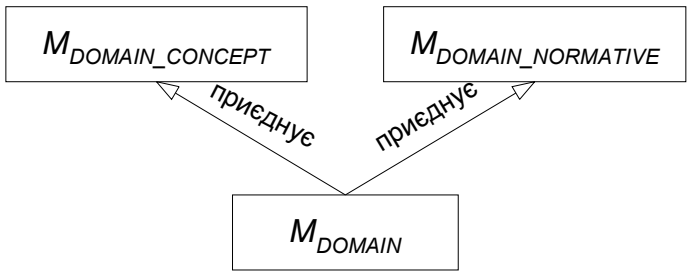

Рисунок 2 – СР-модель онтології предметної області

Логічний аналіз тексту полягає у побудові логічного еквівалента тексту. Виділяють два підходи до побудови логічного еквівалента тексту:

- у вигляді формули логічного числення (числення предикатів першого порядку або інтенсіональної логіки) із об'єктів предметної області;
- у вигляді семантичного графа, вершинами якого є об'єкти та відношення предметної області, а дугами – двомісні семантичні відношення [2] («*агент*», «*інструмент*», «*об'єкт*»).

У більшості випадків ці способи дозволяють еквівалентно відобразити семантичні властивості тексту, що аналізується. Але використання двомісних семантичних відношень має більше переваг, оскільки дозволяє розкласти *n*-місні відношення на більш прості, а також побудувати висловлення про взаємодію актантів *n*-місного відношення.

Отже, для побудови логічного еквівалента тексту є необхідною інтеграція понять та відношень предметної області із двомісними семантичними відношеннями. Для цього побудуємо наступні бази знань:

– *онтологія природної мови* – БЗ, яка містить формальний опис загальних лінгвістичних знань (лінгвістичні класифікації [2], двомісні семантичні відношення);
– *лінгвістично-орієнтована онтологія предметної області* – БЗ, яка містить для кожного поняття предметної області його лінгвістичну характеристику (місце в лінгвістичній класифікації та двомісні семантичні відношення).

Онтологія природної мови та лінгвістично-орієнтована онтологія предметної області не залежать від конкретної природної мови. На основі онтології природної мови може бути створено лінгвістично-орієнтовану онтологію будь-якої предметної області. На основі лінгвістично-орієнтованої онтології може бути побудовано базу знань про лексичний склад будь-якої природної мови.

Для подання цих онтологій за допомогою *СР*-методу подання знань створюються наступні *СР*-моделі знань:
– M_{LING} – *СР*-модель онтології природної мови;
– M_{LING_DOMAIN} – *СР*-модель лінгвістично-орієнтованої онтології предметної області.

Взаємозв'язок цих моделей між собою та з онтологією предметної області поданий на рис. 3. Для їх побудови використовуються об'єктна, продукційна та аксіоматична *СР*-моделі знань.

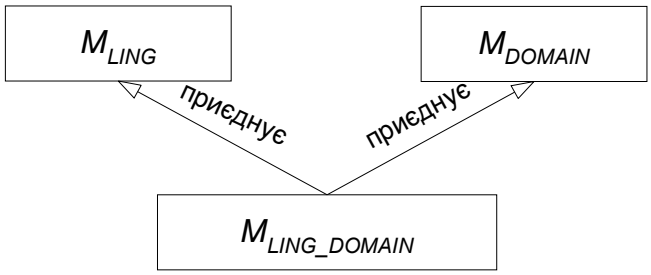

Рисунок 3 – СР-модель лінгвістично-орієнтованої онтології предметної області

Логічний еквівалент тексту будується у вигляді *СР*-моделі M_{TEXT}, яка приєднує лінгвістично-орієнтовану онтологію предметної області (рис. 4). Для побудови моделі використовуються об'єктна *СР*-модель та *СР*-модель логічних операторів.

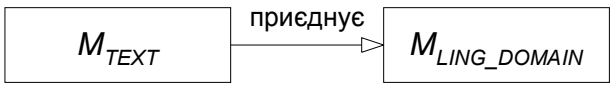

Рисунок 4 – СР-модель тексту M_{TEXT}

Під час побудови *СР*-моделі логічного еквівалента тексту відбувається аналіз її цілісності. Якщо логічний еквівалент тексту побудований невірно, є логічно суперечливим, або порушує закони нормативної онтології предметної області, то *СР*-модель не буде цілісною. Порушення цілісності відстежується обмежувальними функціями основних *СР*-моделей знань. Деякі випадки порушення цілісності моделі M_{TEXT} розглянуто в таблиці 3.

Таблиця 3 – Випадки порушення цілісності СР-моделі M_{TEXT}

Причина	Пояснення	Засоби контролю
Невиконання обмежень на сполучуваність понять	Обмеження на сполучуваність понять задаються в лінгвістично-орієнтованій онтології предметної області. Прикладами понять, які не можуть сполучуватись, є «пити» та «камінь», «читати» та «голку» тощо.	Обмежувальні функції об'єктної та аксіоматичної моделей
Антонімія	Антонімія означає, що один і той самий суб'єкт не може одночасно виконувати дії, які є антонімами. Відношення антонімії та вищевказана аксіома антонімії задаються в онтології природної мови. Антоніми задаються у лінгвістично-орієнтованій онтології предметної області.	Обмежувальна функція аксіоматичної моделі
Порушення нормативної бази	Порушення нормативної бази означає, що в тексті знайдено факти, які порушують закони нормативної бази.	Обмежувальні функції об'єктної та аксіоматичної моделей
Суперечність тверджень	Суперечність означає одночасне виконання твердження та його заперечення.	Обмежувальна функція логічної моделі

Отже, метод аналізу несуперечності текстів та їх відповідності нормативній базі полягає у

1) виділенні у формальному описі предметної області концептуального, нормативного та лінгвістично-орієнтованого прошарків;
2) використанні *СР*-методу подання знань для подання формального опису предметної області та побудови логічного еквівалента тексту.

Висновки

У статті розглянуто проблему автоматизованого аналізу несуперечності текстів та їх відповідності нормативній базі. Запропоновано здійснювати аналіз несуперечності шляхом логічного аналізу тексту із залученням знань про предметну область, природну мову та нормативну базу. Запропоновано підхід до моделювання предметної області, який полягає у виділенні у формальному описі предметної області концептуального, нормативного та лінгвістично-орієнтованого прошарків, що дозволяє залучати всю необхідну для аналізу несуперечності інформацію.

Запропоновано обмежувально-продукційний метод подання знань, який дозволяє інтегрувати різні формальні моделі подання знань та контролювати цілісність знань. Продемонстровано, що логічний еквівалент тексту та формальний опис предметної області можуть бути подані за допомогою цього методу.

Показано, що при використанні запропонованого підходу до моделювання предметної області та використанні обмежувально-продукційного методу як засо-

бу подання знань аналіз несуперечності тексту та аналіз відповідності тексту нормативній базі зводиться до побудови логічного еквівалента тексту.

Отримані результати дозволяють автоматизувати аналіз несуперечності текстів.

Література

1. Логический анализ языка: Противоречивость и аномальность текста.– М., 1990.
2. Кобозева И.М. Лингвистическая семантика: Учебник. – М.: Эдиториал УРСС, 2002. – 353 с.
3. Мельчук И.А. Опыт теории лингвистических моделей «Смысл-Текст». Семантика, Синтаксис. – М.: Наука, 1974. – 346 с.
4. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2001. – 384 с.
5. Буч Г. Объектно-ориентированное проектирование с примерами применения. – Киев: Диалектика, 1992.
6. Демченко О.М., Сирота О.П. Модельний метод подання знань в задачах автоматизованої обробки текстів природної мови // Системні технології. – 2004. – № 1(30). – С. 20-27.

Е.П. Сирота

Ограничительно-продукционный метод представления знаний для анализа непротиворечивости текстов

В статье рассматривается проблема автоматизированного анализа непротиворечивости текстов и соответствия текстов нормативной базе. Предлагается осуществить анализ непротиворечивости путем логического анализа текста с привлечением знаний о предметной области, естественном языке и нормативной базе. Предлагается ограничительно-продукционный метод представления знаний и рассматривается его применение для представления логического эквивалента текста и моделирования предметной области. Предлагается подход к моделированию предметной области, который позволит осуществить анализ непротиворечивости текста. В частности этот подход предусматривает включение нормативной базы в формальное описание предметной области и создание лингвистически-ориентированной базы знаний предметной области, что позволит выполнить анализ семантической правильности построения текста и анализ соответствия текста нормативной базе.

Constraint-Production Knowledge Representation Method for Text Consistency Analysis

The paper is dedicated to the problem of automated text consistency analysis and text to normative base correspondence analysis. It is proposed to implement text consistency and correspondence analysis via text logic analysis with the attraction of application domain, natural language and normative base knowledge bases. The constraint-production knowledge representation method is proposed and the method's applying to formal representing of text logic equivalent and application domain is considered. An approach is proposed to application domain modeling which will allow to perform text consistency analysis. In particular, an approach provides for to include normative base knowledge into application domain formal representation and to create linguistic-oriented knowledge base for application domain. This will allow to analyze text semantic compatibility and text to normative base correspondence.

Статья поступила в редакцию 29.06.2004.