УДК 004.4'414(045)

Сич М.Ю.

МЕХАНІЗМ ВЗАЄМОПЕРЕТВОРЕННЯ ПРИРОДНО-МОВНИХ СЛОВАРНИХ КОНСТРУКЦІЙ

Національний авіаційний університет

Пропонується опис механізму трансформації текстової природно-мовної інформації. Механізм заснований на перетворенні словарних конструкцій у графи з подальшою модифікацією наборів значень їх вершин та дуг на основі правил заміни. Правила трансформації формулюються у вигляді пар взаємозамінних словарних конструкцій на природній мові. У якості моделі знань обрана семантична мережа з фіксованим набором зв'язків між вершинами.

Вступ

Пошук інформації сьогодні є однією зі звичайних справ сучасної людини. Цивілізована зайнята людина робить це кожен день, вже навіть не помічаючи цього. Найчастіше цей пошук пов'язаний з текстовою природно-мовною інформацією. Більшість пошукових систем мережі Інтернет, локальних мереж та ресурсів, різноманітних автоматизованих інформаційно-довідкових служб перш за все націлені на обробку та роботу з неформалізованою інформацією, що записана за допомогою звичайній повсякденної форми – природної людської мови.

Робота з природно-мовною інформацією ініціює цілий ряд складнощів. З одного боку інформаційні системи повинні працювати з запитами користувачів, глибоко проникаючи у їх зміст, сягаючи суті інформації, з іншого боку, роботи систем пошуку інформації має бути швидкою.

Однією з ключових проблем при пошуку у природно-мовній інформації є багатоманітність фактів. Одна і та сама подія може бути описана різними констатаціями слів. Смисли окремих слів при цьому різна, проте суть і зміст всього виразу або словарної конструкції (СК) лишається єдиним. І при пошуку у певному тексті у відповідь на пошуковий запит цілком можливо, що шукана інформація у тексті є, проте через різницю у констатації у запиті і у інформації пошук не дасть позитивного результату. Суть інформації все ще лишається недосяжною для пошукових систем.

Постановка задачі

Задача полягає у створенні механізму для взаємоперетворення текстових природно-мовних словарних конструкцій. Результатом роботи повинна стати можливість пояснювати системам інтелектуальної обробки природно-мовної інформації одні словарні конструкції через інші простою повсякденною мовою, без необхідності формалізації нових даних чи їх кодування.

Взаємоперетворення словарних конструкцій

За своєю суттю певний вислів або речення є словарною конструкцією, у якій смисли слів пов'язані певним характером взаємодії.

У даній роботі ми використаємо для представлення знань семантичну мережу [1]. Множина значень вершин – вся сукупність словоформ певної мови. Назвемо цю множину *Words*. Множина значень дуг (зв'язків) – сукупність граматичних властивостей словоформ у вигляді питань: хто?, що?, який?, яка?, де?, коли?, скільки? і т.п. Цю множину назвемо Questions.

Механізм автоматичної трансформації текстової інформації у цю модель знань, тобто фактично алгоритм синтаксичного та семантичного аналізу текстової інформації, описаний у [2].

Головною вершиною певного зв'язку назвемо вершину, з якої починається дуга зв'язку.

Підлеглою вершиною зв'язку назвемо ту, яка є кінцем дуги цього зв'язку.

Батьківськими зв'язками вершини будемо називати всі зв'язки даної вершини, у яких вона є підлеглою.

Дочірні зв'язки вершини - всі зв'язки вершини, для яких вона є головною.

Словарна конструкція є орієнтованим графом з певним значенням вершин та значенням взаємозв'язків між ними. Тому перетворення або трансформація певної словарної конструкції в іншу зводиться до заміни одного фрагменту графу іншим зі збереженням зовнішніх зв'язків.

Для коректного перетворення потрібна бути задана певна відповідність між різними фрагментами графу. Цю відповідність можна представити у вигляді пари взаємозамінних графів з певними вершинами та зв'язками між ними. Ця пара означає, що при знайденні у інформації уривку, що відповідає по структурі одному з фрагментів пари, його можна замінити другим уривком, відповідно трансформувавши у потрібну форму.

Побудова, а точніше задання взаємозамінних графів здається складним та громіздким завданням. Потрібно задати властивості вершин та характер взаємодії між ними, потім зберегти дані структури у певну базу, звідки при необхідності можна буде їх діставати та співставляти з інформацією. Задача ця під силу технікам та розробникам.

Проте вирішити цю проблему достатньо просто. Адже дані графи відповідностей можуть бути заданими звичайними природномовними словарними виразами або конструкціями. І вже на основі цих виразів будуть побудовані графи еквівалентних словарних конструкцій. Для зрозумілості приведемо приклад.

Візьмемо вислів «Цей чоловік завершить справу». За значенням він еквівалентний виразу «Цей чоловік доведе справу до кінця». Еквівалентними фрагментами речень є «закінчить справу» та «доведе справу до кінця». Відповідно взаємозамінними словарними конструкціями є пара: «закінчити справу» та «довести справу до кінця». Як бачимо, взаємозамінні графи можуть бути задані не технічним кодуванням, а простими природномовними виразами. Особливістю цього способу задання є як раз те, що значення вершин у самому реченні і у частинах правила трансформації є тими самими, характер взамозвязків між ними теж той самий у обох випадках. Це проілюстровано на рис.1.

Інформаційна словарна конструкція початкова (ІСКП) – вислів або речення, що містить у собі певну інформацію на природній мові, що має аналізуватись на предмет заміни словарних конструкцій (рис.1а). Множину вершин цієї СК назвемо Inf.

Первинна словарна конструкція (ПСК) – частина правила взаємоперетворення, що у даний момент аналізується на предмет знаходження відповідності до неї інформаційної СК (рис.1в). Множину вершин СК назвемо First.

Вторинна словарна конструкція (ВСК) – частина правила взаємоперетворення, на яку має бути замінена інформаційна СК, що відповідає первинній СК (рис.1г). Множину вершин назвемо *Second*.

Слід зазначити, що оскільки ми приймаємо, що частини правила взаємоперетворення СК є рівноправними, то вони обидві можуть бути як первинними, так і вторинними.

Вхідна словарна конструкція (ВхСК) – інформаційна СК, що при аналізі з первинною СК була прийнята як відповідна до неї (вершини 3,4 на рис.1а). Множину вершин цієї СК назвемо *Input*.

Вихідна словарна конструкція (ВихСК) – інформаційна СК, фрагмент графу, що отримується з вторинної СК шляхом підстановки у неї фактичних значень вершин з вхідної СК. Фактично, вихідна СК є новою словарною конструкцією, побудованою за допомогою правила трансформації (вершини 3-6 на рис. 1б). Множина вершин – *Output*.

Інформаційна словарна конструкція результуюча (ІСКР) – вислів або речення, що містить у нову словарну конструкцію, отриману на основі вторинної СК правила трансформації та початкової інформаційної СК (рис. 1б). Множина вершин – *InfOut*.

Таким чином, для опису процесу трансформації словарних конструкцій ми використаємо такі множини:

$$Inf = \{Inf_i \mid Inf_i \in Words\};$$

$$First = \{First_i \mid First_i \in Words\};$$

$$Second = \{Second_i \mid Second_i \in Words\};$$

$$Input = \{Input_i \mid Input_i \in Words\};$$

$$Output = \{Output_i \mid Output_i \in Words\};$$

$$InfOut = \{InfOut_i \mid InfOut_i \in Words\};$$

Послідовність операцій по трансформації словарних конструкцій приведена на рис.2.

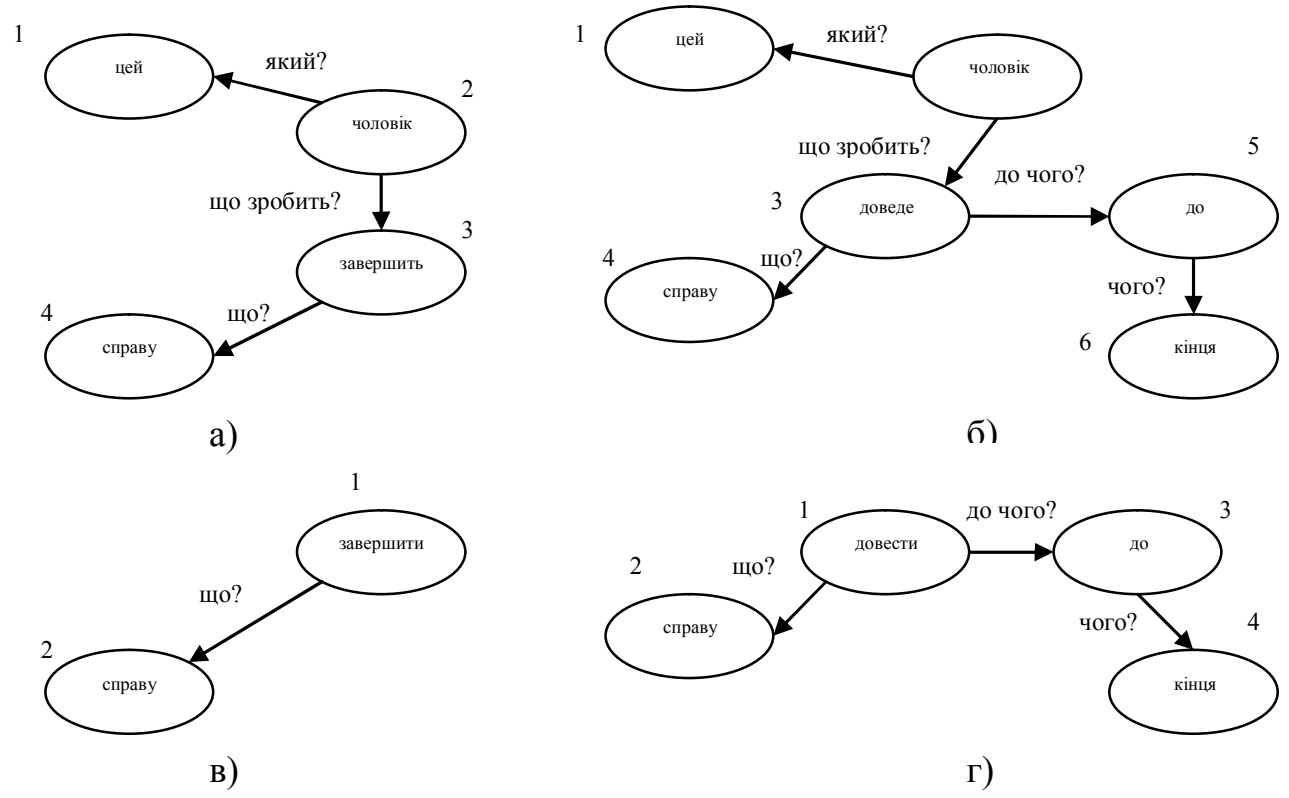

Рис. 1. Трансформації словарних конструкцій.
а) початкова інформаційна словарна конструкція; б) результуюча інформаційна словарна конструкція; в) первинна частина правила трансформації; г) вторинна частина правила трансформації

Рис. 2. Послідовність етапів трансформації словарної конструкції

Етап 1. Захоплення вхідної словарної конструкції.

На цьому етапі відбувається співставлення інформаційної СК та первинної СК. Результатом має стати знаходження відповідності між вершинами ІСК та ПСК, тобто кожному елементу $First_i$ має бути поставлений у відповідність елемент Inf_i, тим самим сформувавши множину *Input*.

Етап 2. Формування вихідної словарної конструкції.

На етапі формування *ВихСК* відбувається створення нової словарної конструкції, що базується на вторинній СК. Це по суті процес побудови нових формулювань інформації у вигляді графу, тобто заповнення множини *Output* на основі *Second*.

Етап 3. Видалення вхідної словарної конструкції з загального графу інформації.

Вершини інформаційної СК, що були поставлені у відповідність первинній СК, тобто входять у вхідну СК, вилучаються з графу інформаційної СК, оскільки вони будуть замінені на вершини нової сформованої вихідної СК.

Етап 4. Додання вихідної словарної конструкції до загального графу інформації.

У граф інформаційної СК додаються вершини вихідної СК (отримана на *етапі 2*). Цей процес відбувається шляхом передачі зовнішніх батьківських зв'язків основи графу вхідної СК на основу графу вихідної СК. На цьому етапі фактично формується множина *InfOut*.

Етап 5. Зчитування нової інформаційної словарної конструкції.

На цьому етапі відбувається перетворення дерева графу отриманої на попередньому етапі ІСК у природно-мовну текстову словарну конструкцію, адже додання нових вершин у ІСК приводить до зміни її структури, тому необхід-

не формулювання нової СК на основі правил граматики.

У ході обробки інформаційних словарних конструкцій, як вже було сказано, обидві частини правила виступають спочатку у ролі первинної СК та вторинної СК. У разі, якщо при накладанні однієї частини правила на інформаційну СК була знайдена відповідність, відбувається заміна на словарну конструкцію з другої частини правила. Якщо відповідність знайдена не була, для аналізу використовується друга частина правила у якості первинної.

Структура правила взаємоперетворення словарних конструкцій

Правило взаємоперетворення складається з двох природно-мовних словарних конструкцій (ПСК) (табл. 1). Кожна така ПСК має два типи вершин: постійні та змінні.

Постійні вершини або *визначеності* мають визначений смисл слова.

Таблиця 1. Структура правила трансформації

№ п/п	Словарна конструкція №1	Словарна конструкція №2
1	завершити справу	довести справу до кінця
2	хтось народився у чомусь	щось є батьківщиною когось
…		
n	перевести подих	віддихатися

Змінні вершини або *невизначеності* не мають чіткого смислу, а тільки задають основні властивості вершини. Власне самі вершини мають смисл, проте ці смисли розглядаються як позначення вершин-шаблонів, яким мають бути поставлені у відповідність фактичні інформаційні вершини при аналізі конкретних ІСК. Невизначеності поділені на групи, як і сутності, які вони представляють. Для задання змінних вершин можна використати займенники невизначеного типу, деякі дієслова, а також прислівники, що використовуються для позначення місця, часу, і т.д. без їх точного називання. Невизначеності по групам сутностей представлені у табл.2.

Використовуючи змінні вершини з'являється можливість вказувати на сутності реального світу, не називаючи їх точно, тим самим показуючи взаємозв'язки між ними. Відбувається абстрагування від точних смислів слів і задається лише характер вершини. Таким чином розширюються можливості задання правил трансформації словарних конструкцій і для ситуацій, коли певна група сутностей реального світу має однаковий характер взаємодії і повний перерахунок їх є складним і громіздким процесом.

Невизначеності у правилах є *точками дотику* різних частин правила, що дають можливість перенести сутності з однієї констатації або словарної конструкції у іншу. Часто можуть складатися ситуації, коли при заданні правила необхідно використати один тип невизначеність кілька разів. У такому випадку можна використати додаткові довільні смисли-маркери для виділення невизначеностей. Це можуть бути прикметники, числівники. Таким чином, для визначеності відповідності певної змінної вершини вторинної СК до первинної, потрібно знайти невизначеності з тими самими наборами уточнюючих слів-маркерів.

Таблиця 2. Змінні вершини для представлення різних груп сутностей

№ п/п	Група	Невизначеності
1	Об'єкти	Хтось, щось, хто-небудь, що-небудь, і т.д.
2	Атрибути об'єктів	Якийсь, який-небудь, і т.д.
3	Дії	Робити щось, робити що-небудь, і т.д.
4	Атрибути дій	Десь, колись, кудись, де-небудь, коли-небудь, куди-небудь, і т.д.
5	Числа	Скількись, котрийсь, скільки-небудь, і т.д.

Захоплення вхідної словарної конструкції

Захопленням вхідної словарної конструкції назвемо процес знаходження відповідності між графом первинної СК з правила трансформації та графом фактичної інформаційної СК. У ході цього процесу необхідно знайти для вершин словарної конструкції з правила транс-

формації відповідні вершини у інформації з тим самим характером взаємодії між ними.

Процес порівняння графу первинної словарної конструкції (ПСК) і графу інформаційної словарної конструкції (ІСК) починається з виділення основи дерева графу ПСК. *Основою дерева графу* будь-якої словарної конструкції будемо вважати вершину, що не має посилання на себе з інших вершин, тобто не існує дуги взаємозв'язку між вершинами графу, яка була б направлена на дану вершину.

Основа графу ПСК послідовно накладається на кожну вершину дерева графу ІСК для знаходження можливої відповідності між вершинам ПСК та ІСК. Цей процес є рекурсивним. При виявленні відповідності вершини ПСК певній інформаційній вершині процес порівняння має повторитись для кожної вершини з множини підлеглих до вершини ПСК. Відбувається спуск від основи дерева ПСК по гілкам до листя. Графічно процес захоплення вхідної СК представлений на рис.3 та рис.4.

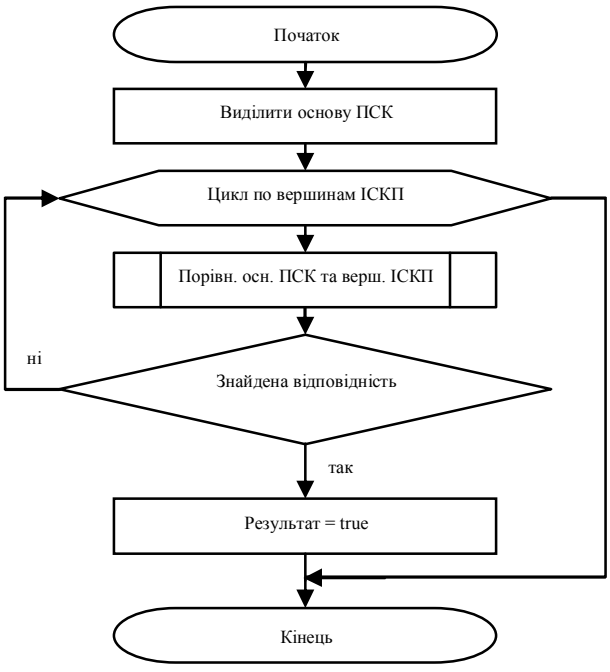

Рис. 3. Схема процесу захоплення вхідної словарної конструкції

При аналізі певної вершини ПСК повинні бути перевірені атрибути самої вершини і характер батьківського взаємозв'язку вершини. Ці величини повинні співпадати у вершин первинної СК та інформаційної СК.

Перевірка рівності батьківських взаємозв'язків відбувається на основі типу самого взаємозв'язку, такі, як наприклад, зв'язок, що вказує на сутність (хто?, що?), або на ознаку сутності (який?), і т.д.

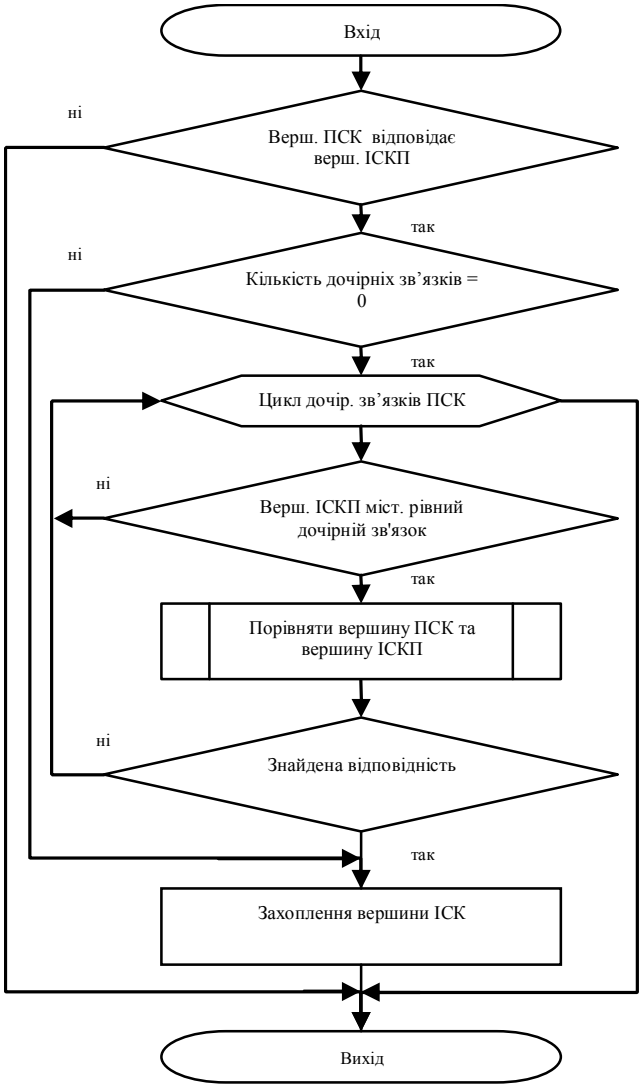

Рис.4. Схема порівняння на відповідність вершин ПСК та ІСКП

Додаткові атрибути зв'язку, що вказують на числа, роди і інше не перевіряються, оскільки їх нееквівалентність допускається.

Скажімо, при аналізі прив'язки «завершити справу» вона має бути визнана рівною таким прив'язкам, як «завершить справу», «завершать справу», «завершать справи» і інші. Характер взаємозв'язків зберігається, проте є різниця у додаткових атрибутах, у даному випадку у числах.

Захоплення вершин ІСК має різні принципи, оскільки правило трансформації містить у собі два типи вершин (постійні та змінні), що мають різні властивості.

Захоплення визначеності відбувається за простою рівністю батьківського взаємозв'язку та смисла слова.

Захоплення невизначеності.

1. Рівність типу взаємозв'язку з батьківською вершиною.

2. Рівність морфологічних властивостей вершин ПСК та ІСК.

При аналізі вершин невизначеностей потрібно перевіряти рівність батьківських взаємозв'язків та рівність властивостей самих вершин. Аналізовані властивості вершин носять переважно морфологічні властивості – відмінок, рід, число та ін.

Формування вихідної словарної конструкції

Вихідна словарна конструкція будується на основі вторинної СК правила трансформації. Таким чином, структура графа вихідної СК буде повністю співпадати з вторинною СК. Вторинна СК по суті є опорним каркасом, праобразом майбутньої нової вихідної словарної конструкції і являє собою фрагмент графу, яким має бути замінений уривок інформаційного графу, що входить до вхідної словарної конструкції.

Вторинна СК складається з постійних та змінних вершин з вказаними взаємозв'язками між ними. Невизначеності тут слугують для можливості поєднання, або знаходження *точок дотику* з первинною СК. Постійні вершини фактично вносять у СК нові смисли слів і дозволяють перебудувати вхідне речення.

Основні концепції виведення або формування вихідної словарної конструкції:

1. Основи первинної та вторинної СК мають співпадати по властивостям. Якщо основою ПСК є вершина певної групи сутностей, то основою ВСК має бути та сама група, інакше побудувати вихідну словарну конструкцію буде неможливо, точніше стане неможливим процес передачі зовнішніх зв'язків вхідної конструкції на вихідну.

2. Вторинна словарна конструкція пов'язана з первинною тільки невизначеностями та основою графу СК. При побудові вихідної словарної конструкції з вхідної беруться значення вершин, що були захоплені невизначеностями первинної СК. Смисли інших вершин та характер взаємозв'язків між вершинами беруться з вторинної СК.

3. Морфологічні властивості вершин-слів вихідної СК мають бути приведені до властивостей вхідної СК. Вершини точок дотику вхідної СК, що були передані у вихідну СК мають зберегти свої морфологічні властивості(кількість, рід, особа, і т.д.), а властивості вершин у вихідній СК, що прив'язані до них, але смисли яких були взяті з вторинної повинні бути змінені для відповідності першим. Тобто потрібно привести їх по числам, особам і т.д.

Процес починається з основи вторинної СК. На основі її властивостей будується нова вершина – основа вихідної СК, тобто перший елемент множини *Output.* Після цього для кожної вершини дочірніх зв'язків $Second_i$ має бути побудована нова вершина для елементу $Output_i$ з тим самим характером взаємозв'язку. Далі відбувається спуск на один рівень до наступної вершини і операції повторюються.

Графічно послідовність етапів процесу формування вихідної СК представлена на рис.5 та рис.6. Рис.5 містить загальну схему, а рис.6 – рекурсивну частину, що має викликатись для пар вершин $Second_i$ та $Output_i$.

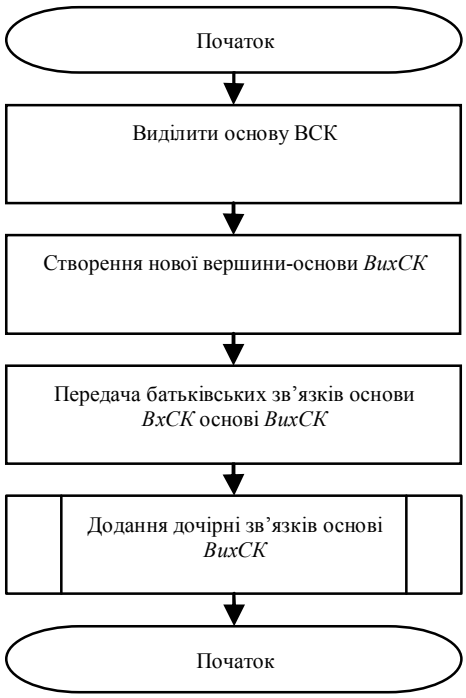

Рис.5. Схема процесу формування вихідної СК

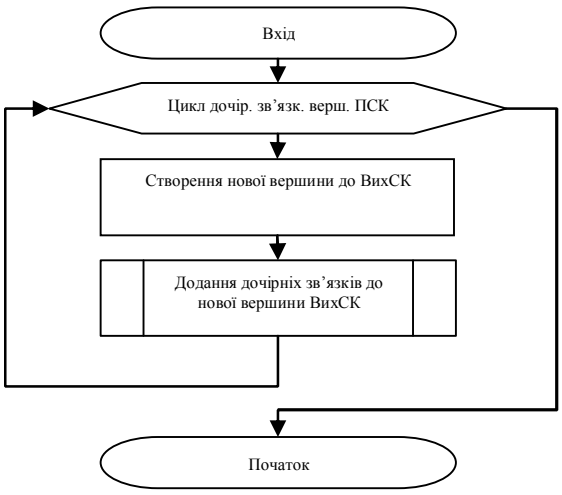

Рис.6. Схема додання дочірніх зв'язків до і-ої вершини ВихСК

Формування кожної нової вершини $Output_i$ вихідної СК відбувається на основі властивостей вершини $Second_i$ вторинної СК (для якої зараз будується вершина ВихСК), вершини батьківського зв'язку $Output_{i-1}$, а також вершини $Input_i$, у разі якщо $Output_i$ є невизначеністю.

Після визначення потрібних атрибутів вершини відбувається звернення до словника словоформ, з якої і отримується потрібне слово. Конкретні джерела властивостей нової вершини $Output_i$ приведені у табл.3.

Видалення вхідної словарної конструкції з загального графу інформації

Після того, як вихідна словарна конструкція була сформована, необхідно з графу початкової інформаційної СК вилучити вершини вхідної СК. При цьому вилучаються вершини вхідної СК та взаємозв'язки між ним, проте тільки внутрішні, тобто ті, початок і кінець яких є частинами захопленої ВхСК.

Додання вихідної словарної конструкції до загального графу інформації

Додання вихідної СК до графу інформаційної СК відбувається фактично прив'язкою основи ВихСК до ІСКП з видаленими вершинами ВхСК. Батьківські зв'язки основи вхідної СК мають бути перенаправлені на основу вихідної СК. Графічно це виглядає так, ніби вершину основи нової СК суміщають з вершиною основи вхідної СК.

Зовнішні зв'язки вершин всієї вхідної СК мають бути збережені у вершин вихідної СК. Це означає, що усі взаємозв'язки, що мають початок або кінець у захоплених у інформаційній СК вершинах, але на іншій стороні яких знаходяться незахоплені вершини, мають перейти у вихідну СК. Для всіх вершин вхідної СК ці зв'язки є дочірніми або підлеглими. Для основи *ВхСК*, а отже і *ВихСК*, це ще й батьківські зв'язки. Скажімо, якщо присудок потрапив у вхідну СК і є її основою, а підмет не потрапив, то очевидно, що у вихідній СК основа залишиться присудком для всього речення і має зберегти зв'язок з підметом.

Таблиця 3. Джерела властивостей вершин графу вихідної СК

Тип верш. *ВихСК*	Джерело властивостей	Властивості		
		Об'єкт	Ознака об'єкту	Атрибути об'єктів
Постійні	$Second_i$	Смисл, част. мови, рід, число, відмін.	Смисл, частина мови	Смисл, част. мови, доконаність
	$Output_{i-1}$		Число, рід, відмінок	Число, особа
Змінні	$Second_i$	Відмінок	Смисл, частина мови, рід, відмінок	Смисл, част. мови, рід, доконаність
	$Input_i$	Смисл, част. мови, числ, рід		
	$Output_{i-1}$		Число	Число, особа

Висновки

Представлена базова концепція роботи механізму трансформації може бути використана у найрізноманітніших типах систем по інтелектуальній обробці текстової природно-мовної інформації.

Оскільки набір правил взаємозаміни може розширюватись з використанням природномовних словарних конструкцій, то навчати систему, що використовує описаний у статті механізм може безпосередньо кінцевий користувач системи, що не обов'язково повинен мати технічні знання у області програмування.

Список літератури

1. Частиков А.П., Гаврилова Т.А., Бєлов Д.Л. Проектирование экспертных систем. – С.-П.: БХВ-Петербург, 2003. – 393 с.

2. Сич М.Ю. Алгоритм семантичної обробки текстової інформації. // Проблеми інформатизації та управління: зб. наук. праць. – К.: НАУ, 2009. – Вип.1(25). – С. 159–164.