The structural types of predicative centre of English sentence

Pavlo Hudan

The laboratory of computational linguistics, Kyiv National Linguistic University, UKRAINE,
Kyiv, 73 Velyka Vasylkivska Str.,
E-mail: pasha_hudan2@mail.ru

The article describes the stages of compiling the classification of structural types of predicative centre of English sentence and deducing the algorithm of automatic identification of predicative centre of the clause with the determination of its type according to the classification, which was implemented in the program by means of the programming language C#.

Having examined the functioning of the predicative centre on the research sources, the structural types of subject and predicate were defined which were based on their formal expressing. The predicative centre was examined as a unity of the subject and the predicate, therefore the denotation of the structural type of the centre consisted of the indication of type of subject and predicate. According to the 500 sentences of research material 792 predicative centers were identified. The amount of the structural types which were met equals 56. The most frequent type of the predicative centre is the subject, which is expressed by pronoun + simple predicate (active voice). Quite frequent types appeared to be the subject expressed by the noun in singular + simple predicate (passive voice) and subject expressed by the noun in singular + compound nominative predicate.

Having analyzed the structural types and means of expressing the predicative centre, the rules were complicated, which became the basis of the algorithm of automatic identification of predicative centre of the sentence. The rules were grounded on the method of distributive analysis and formal approached to the analysis of sentence structure: the methods of searching the point of reference, method of predicting analysis and methods of searching the limit-type signal. The compiled algorithm at the entrance receives the text with morphological marking.

The created program uses the ruled of indication of the sentence predicative centre which are saved in the external database. This gives the opportunity to make changes in the program operation and also alter the marking of the morphological codes which are used in the text at the entrance without interfering into the program code.

Переклад зроблено Горьковою Н.Г., центр іноземних мов «Universal Talk», www.utalk.com.ua

Структурні різновиди предикативного центру англійського речення

Павло Худан

Лабораторія комп'ютерної лінгвістики, Київський національний лінгвістичний університет, УКРАЇНА,
м.Київ, вул.Велика Васильківська, 73,
E-mail: pasha_hudan2@mail.ru

У статті описано етапи укладання класифікації структурних типів предикативного центру англійського речення та створення алгоритму автоматичного виділення предикативного центру речення з установленням його типу відповідно до класифікації, Алгоритм реалізовано у програмі мовою програмування C#. Встановлено структурні типи суб'єкта та предиката на підставі їх формального вираження. Предикативний центр розглядався як єдність суб'єкта та предиката, тому позначення структурного типу центру складалося з позначення типу суб'єкта та типу предиката. На 500 реченнях дослідного матеріалу було виділено 791 предикативний центр. Кількість структурних типів, що зустрілися, дорівнює 56.

Укладено правила, що лягли в основу алгоритму автоматичного виділення предикативного центру речення. Правила базувалися на методі дистрибутивного аналізу та формальних підходах до аналізу структури речення: методі пошуку опорних точок, методі передбачувального аналізу та методі пошуку граничних сигналів. Укладений алгоритм на вході отримує текст з морфологічною розміткою.

Створена програма використовує правила встановлення предикативного центру речення, що зберігаються у зовнішній базі даних. Це дає змогу вносити зміни до роботи програми, а також змінювати позначення морфологічних кодів, що використовуються у тексті на вході без втручання до коду програми.

Ключові слова – предикативний центр, структурний тип, класифікація, автоматичне виділення, алгоритм, суб'єкт, предикат.

I. Вступ

На сучасному етапі все більше уваги приділяється предикативному центру речення у контексті автоматичного синтаксичного аналізу, оскільки синтаксичний аналіз є базою для автоматизації таких процесів як: реферування тексту, машинний переклад. Автоматичний семантичний аналіз – одна з головних задач комп'ютерної лінгвістики – спирається в першу чергу на синтаксичний аналіз, тому що синтаксичні зв'язки між словами речення вказують також на їх семантичну зв'язаність. Водночас, синтаксичний аналіз дає змогу уточнювати морфологічний аналіз, оскільки для зняття омонімії в межах деяких омонімічних пар необхідно знати синтаксичну функцію проблемного слова в реченні.

Першим завданням синтаксичного аналізу речення є виділення його предикативного центру. Саме предикативний центр є базою для всіх зв'язків у межах речення – група підмета і група присудка включають

всі другорядні члени речення, з'єднуючи всі елементи в одну закінчену думку. Отже, від того, наскільки якісно встановлено предикативний центр речення, залежить весь подальший автоматичний синтаксичний аналіз [2, с. 78].

II. Створення класифікації предикативних центрів

Для встановлення структурних типів предикативного центру англійського речення було здійснено синтаксичний аналіз 500 речень наукового тексту.

У ході аналізу встановлювалися предикативні центри в простих та складних реченнях. Суб'єкт позначався одинарними квадратними дужками, предикат – подвійними. Слова або словосполучення, що стояли між конструктивними елементами суб'єкта або предиката, вилучалися за допомогою одинарних фігурних дужок. Така розмітка дала змогу в подальшому автоматично вилучити всі предикативні центри з речення в окрему базу даних для подальшого їх структурного аналізу. Розмічене речення мало такий вигляд:

While [the availability] of tagged corpora [[is {generally} scarce]], [[there are]] certain NLP [tools] [that] [[are accurate]] enough to allow automatic annotation.

Після аналізу усіх знайдених предикативних центрів за складом їх елементів та способами їх вираження, були виділені структурні типи суб'єкта та предиката, наведені у таблицях 1 та 2. Кожному з типів суб'єкта та предиката було поставлено у відповідність умовне позначення, що складається з літер латинського алфавіту.

Таблиця 1

СПОСОБИ ВИРАЖЕННЯ СУБ'ЄКТА У ТЕКСТІ НАУКОВОГО СТИЛЮ АНГЛІЙСЬКОЇ МОВИ

Тип	Позначення
Іменник у формі однини	NSg
Іменник у формі множини	NPl
Іменник, власна назва	PN
Займенник	Prn
Іменник у складі конструкції з словом на позначення кількості або частини цілого	NPh
Займенник *it* з інфінітивом у якості семантичного підмета	ItInf
Займенник *it* з підрядним реченням у якості семантичного підмета	ItCl

Оскільки англійська мова вважається мовою з достатньо фіксованим порядком слів [4, с. 2], правило, за яким суб'єкт займає позицію перед предикатом, часто використовується під час аналізу структури речення. Проте, існують випадки, коли порядок слів у реченні змінюється [5, с. 2]. Цим обґрунтовано виділення конструкції there + be в окремий тип предиката, оскільки при ній змінюється порядок елементів у предикативній парі.

Таблиця 2

СПОСОБИ ВИРАЖЕННЯ ПРЕДИКАТА У ТЕКСТІ НАУКОВОГО СТИЛЮ АНГЛІЙСЬКОЇ МОВИ

Тип	Типові структури	Позначення
Простий предикат, активний стан	*verb*	SPA
	be + Ving	
	have + PII	
	do + not + Infinitive	
Простий предикат, пасивний стан	*be + PII*	SPP
	have + been + PII	
Складений іменний предикат	*be + noun/adjective/ (to)Infinitive*	CNP
Складений іменний предикат з іменною частиною, вираженою підрядним реченням	*be + that + Clause*	IsThat
Складений модальний предикат, активний стан	*modal verb + Infinitive*	CVMPA
	be + able + to + Infinitive	
Складений модальний предикат, пасивний стан	*modal verb + be + PII*	CVMPP
Складений модальний іменний предикат	*modal verb + be + noun/adjective*	CMNP
Предикат, виражений конструкцією *there + be*	*there + be*	ThPh

Предикативний центр речення розглядається як єдність суб'єкта та предиката. Тому, для створення класифікації структурних типів предикативного центру було зіставлено таблицю 1 та таблицю 2 (кожен тип суб'єкта сполучається с кожним типом предиката). Так був отриманий список всіх теоретично можливих предикативних центрів з одним суб'єктом та одним предикатом, що містить 56 типів центрів. Проте, на матеріалі дослідження було знайдено лише 29 таких центрів. Це означає, що існують правила сполучуваності даних типів суб'єкта та предиката.

Важливо зауважити, що велика кількість () предикативних центрів містила більше одного суб'єкта та/або предиката. Тому, для запису предикативного центру за допомогою позначень наведених вище, була використана наступна формула:

$$ST_1+ST_2+\ldots+ST_n+PT_1+PT_2+\ldots+PT_m \quad (1),$$

де ST – тип суб'єкта, PT – тип предиката, n – кількість суб'єктів у предикативному центрі, m – кількість предикатів у предикативному центрі, + – зв'язок між суб'єктом і предикатом.

Список всіх центрів (791 предикативний центр), представлених у вигляді формули (1), було згруповано за типами. Найбільш частотні центри наведені у таблиці 3.

Таблиця 3

ЧАСТОТНІ ПРЕДИКАТИВНІ ЦЕНТРИ АНГЛІЙСЬКОГО РЕЧЕННЯ НА МАТЕРІАЛІ У 500 РЕЧЕНЬ

Тип центру	Кількість
Prn+SPA	145
NSg+SPP	101
NSg+CNP	89
NSg+SPA	84
NPl+SPP	58
NPl+SPA	42
NPl+CNP	24
PN+PN+SPA	23

III. Алгоритм автоматичного виділення предикативного центру речення

Цей етап вже передбачає виконання автоматичного синтаксичного аналізу. Загальний алгоритм виділення предикативного центру речення було побудовано на базі формально-граматичних методів синтаксичного аналізу, описані Ю. Д. Апресяном [1, с. 232-252]. Методи пошуку опорних точок та передбачувального аналізу стали кістяковими для вищезгаданого алгоритму.

Так, метод опорних точок передбачає, що в реченні існують такі одиниці, які можна часто однозначно визначити. Ці одиниці є опорними точками для подальшого аналізу і стають опорними для встановлення функцій інших одиниць речення. За опорні точки було взято модальні дієслова, особові форми дієслова *to be*, певні особові форми повнозначних дієслів та ін.

Передбачувальний аналіз полягає в тому, що поява певної одиниці викликає появу іншої, що є або її логічним завершенням, або щільно з нею пов'язана. Таким чином, наприклад, поява модального дієслова вимагає появи дієслова, що є другою частиною предиката.

За допомогою згаданих методів, було встановлено набір правил, що лягли в основу алгоритму автоматичного виділення предикативного центру. Такий алгоритм передбачає, що кожна словоформа вхідного речення має морфологічний код (див. працю Н. П. Дарчук: «… як мінімум, нам необхідно навчити комп'ютер визначати частину мови та її граматичні характеристики … тобто він має здійснити автоматичний морфологічний аналіз. Без цієї інформації про частини мови і форму слова АСА неможливий» [3, с. 99]).

Оскільки більшістю опорних точок були перші елементи предикатів (особові форми дієслова *to be*, модальні дієслова і т.д.), алгоритм починає роботу с пошуку та встановлення типу предиката. Наступним кроком є знаходження суб'єкта, що відповідає цьому предикату. Достатньо сталий порядок слів англійської мови дає можливість з високою вірогідністю знаходити суб'єкт у лівому напрямку від предиката. Виключеннями будуть питальні речення та речення з екзистенціальними зворотами *there is/are/…*, в яких порядок слів є інверсійним.

IV. Створення програми

Як було згадано вище, автоматичний синтаксичний аналіз неможливий без попереднього морфологічного аналізу. З цієї причини, алгоритм і програма використовують у своїй роботі дані про граматичні класи та форми слів.

Створена програма налаштована на текст з морфологічною розміткою, виконаною автоматичним морфологічним аналізатором UCREL CLAWS5, доступним за адресою http://ucrel.lancs.ac.uk/claws5 tags.html.

На вхід створеної програми подається речення, що має такий вигляд:

Acquisition_NN1 of_PRF the_AT0 subcategorization_NN1 frames_NN2 of_PRF verbs_NN2 is_VBZ described_VVN in_PRP section_NN1 21.4_CRD ._.

З речення бачимо, що морфологічний код приписано кожній словоформі (а також розділовому знаку речення) через знак «_». Програма виконує розбиття вхідного речення на список словоформ та список їх морфологічних кодів. У подальшому програма звертається до словоформи або до її морфологічного коду, залежно від умови конкретного правила встановлення предикативного центру.

Одним з етапів програми є пошук словоформ, що входять до списку складених прийменників, сполучників та прислівників. Їх морфологічний код відрізняється від звичайних тим, що в кінці коду дописано дві цифри: кількість компонентів у складеному слові та номер конкретного компонента складеного слова. Наприклад, складений сполучник according to представлено в тексті у такому вигляді: «according_PRP21 to_PRP22».

Список словоформ або граматичних кодів складених прийменників та сполучників зберігається у окремій таблиці у базі даних. Алгоритм виконує пошук кожної словоформи або граматичного коду у реченні, перевіряє наявність інших компонентів поряд із знайденою словоформою та представляє їх у вигляді одного слова зі знаками «+» замість пробілу. Цьому слову приписується відповідний граматичний код (без цифр вкінці), що задається у відповідній комірці бази даних.

Таким чином, сполучник according to буде мати такий вигляд: «according+to_PRP». До цієї таблиці можна занести список багатокомпонентних термінів або ідіоматичних сполук (максимальна довжина – 3 слова), для того щоб програма розглядала їх як одну словоформу. Таке рішення спрощує роботу алгоритму, якому не доводиться вирішувати під час аналізу структури речення, чи є слово, наприклад, компонентом складеного прийменника або багатокомпонентного терміна.

Алгоритм реалізовано таким чином, що всі правила знаходяться у окремій таблиці бази даних. До цієї таблиці заносяться також вказівки для роботи алгоритму: які морфологічні коди шукати при знаходженні опорної точки, які морфологічні коди ігнорувати, чи додавати знайдену словоформу до складу суб'єкта/предиката, чи є дана словоформа початком нового суб'єкта/предиката.

До кожної комірки таблиці можна заносити не лише морфологічний код словоформи з «_» на початку, але і окремі словоформи. За необхідністю введення декількох морфологічних кодів/словоформ, вони перераховуються через кому з пробілом. Це дає змогу точніше налаштовувати роботу алгоритму. Рис.1 дає

уявлення про таблицю бази даних, що задає правила встановлення предиката.

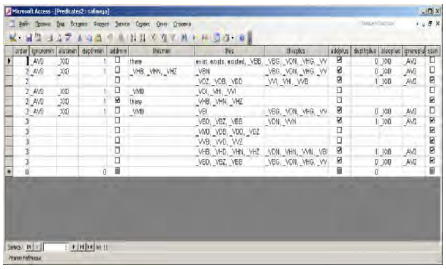

Рис.1 Набір правил встановлення предиката

Зовнішнє зберігання правил встановлення предикативного центру речення спрощує подальшу роботу з програмою – велику частину налаштування та коригування можна виконувати не втручаючись до коду самої програми. Крім того, за необхідності, можна повністю або частково змінити морфологічні коди, що використовуються програмою, переписавши таблицю, заповнюючи комірки відповідними новими кодами морфологічних класів слів.

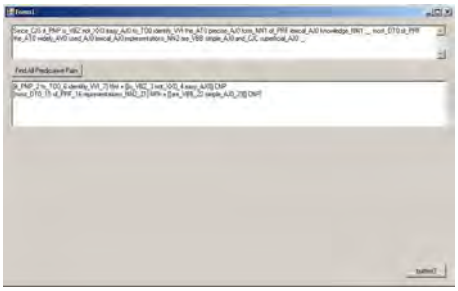

Рис.2 Інтерфейс програми.

Недолік програми полягає у тому, що результат частково залежить від результатів морфологічної розмітки тексту. Помилка у встановленні класу однієї словоформи часто призводить до неправильної роботи всього алгоритму. Інший недолік – неможливість виявлення однорідних членів речення.

Висновок

У статті описано етапи укладання класифікації структурних типів предикативного центру речення, створення алгоритму автоматичного встановлення предикативного центру речення та його реалізація у програмі.

Література

[1] Апресян Ю. Д. Идеи и методы современной структурной лингвистики (краткий очерк) / Ю. Д. Апресян. – М.: Просвещение, 1966. – 305 с.

[2] Дарчук Н. П. Автоматическое установление подлежащих и сказуемых в тексте / Н. П. Дарчук // Автоматизация анализа научного текста. – К.: Наук. думка, 1984. – С. 78-98.

[3] Дарчук Н. П. Комп'ютерна лінгвістика (автоматичне опрацювання тексту): підручник / Н. П. Дарчук. – К.: Видавничо-поліграфічний центр «Київський університет», 2008 – 351 с.

[4] Covington M. A. A Dependency Parser for Variable-Word-Order Languages / Michael A. Covington. – Athens: The University of Georgia, 1990. – 36 p.

[5] Primus B. Word Order [Електронний ресурс] / Beatrice Primus. – Режим доступу: ebookbrowse.com/primus-cels-word-order-pdf-d78397478