Functional characteristics and morphological annotation of verbs in English abstracts

Irina Poliakova

The Laboratory of Computational Linguistics, Kyiv National Linguistic University, UKRAINE, Kyiv,
V. Vasylkivska Street 73,
E-mail: ririna.cl@gmail.com

As a result of the conducted research, the forms of the verbs which can be met in English abstracts were established and the morphological indications were determined. Namely:

- on the basis of the contextual analysis the formal indications of the verb in sentence were defined;
- on the basis of stated indications deferential signs of verbal forms were deduced which were assumed as a basis of the program of morphological analysis of verbs in English abstracts meaning the rules which the program operates with.

According to the abstract analysis results were identified that the verbs are used in passive voice no more often than in active voice, but the word combinations as ‘we present’, ‘we report’, ‘we introduce’, ‘we study’, ‘this paper presents’, ‘this paper describes’ are utilized. Thus, it can be summarized that the article and abstract were written by the author or co-authors together. It was also detected that infinite forms of verbs are used quite often; among them –ing forms and infinitives are more common.

To write the program of morphological analysis of the verbs in English abstracts the special approach which uses database of verbal forms was chosen. The program refers to the database and compares the words with those in the text of abstract finding potential verbal forms. Such approach eliminates the probability of appearance of lexico-grammatical homonymy ‘verb-noun’.

Created on the basis of the rules which determine the verbal forms and taking off the lexico-grammatical homonymy in the text, the program identifies about seventy percents of verbs.

The program makes the chart with found verbs, their morphological indications, morphological code and context.

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

Функціональні характеристики та морфологічне кодування англійських дієслів в анотаціях

Ірина Полякова

Лабораторія комп’ютерної лінгвістики, Київський національний лінгвістичний університет,
УКРАЇНА, м. Київ, вул. В. Васильківська, 73,
E-mail: ririna.cl@gmail.com

У даній роботі представлено спосіб морфологічного кодування з використанням бази даних дієслівних форм. Такий підхід зменшує ймовірність виникнення лексико-граматичної омонімії "дієслово-іменник" і дозволяє визначити близько сімдесяти відсотків дієслівних форм тексту. Як результат виводиться таблиця зі знайденими дієсловами, їх морфологічними ознаками, морфологічним кодом та контекстом.

Ключові слова – комп’ютерний морфологічний аналіз, дієслівна форма, диференційна ознака, контекстний аналіз, анотація, лексико-граматична омонімія.

I. Вступ

Морфологічний аналіз пов'язаний з добуванням зі слова інформації про його будову, синтаксичні та морфологічні характеристики або значення морфологічного складного слова. [4]

Проблема комп’ютерного морфологічного аналізу (КМА) досі є важливим питанням у сфері опрацювання природної мови. Метою КМА є розуміння внутрішнього механізму утворення словоформ. Морфологічний аналізатор може забезпечити виявлення важливої інформації для таких комп’ютерних лінгвістичних задач як лематизація, синтаксичний аналіз, машинний переклад, пошук інформації, кодування (або розмітка) та багато інших. [3]

Оскільки модуль автоматичного морфологічного аналізу є обов’язковим для всіх систем автоматичного опрацювання тексту, було вирішено створити програму присвоєння морфологічних кодів дієслівним формам. Для створення такої програми необхідні правила, що базуються перш за все на диференційних ознаках дієслівних форм, за допомогою яких комп’ютер відрізняє одну форму дієслова від іншої та приписує їй морфологічний код.

II. Формальні диференційні ознаки дієслівних форм

Диференційна ознака – це формальна чи змістовна риса певної одиниці чи групи одиниць, яка відрізняє її від інших одиниць. Диференційні ознаки використовуються для опису чи обчислення системи одиниць та їх форм. Кожна одиниця характеризується індивідуальним набором диференційних ознак. [2] Так, наприклад, для інфінітиву можна виділити такі формальні диференційні ознаки, як:

1) перед інфінітивом теперішнього часу активного стану (indefinite active infinitive) вживається частка “to”; 2) перед інфінітивом теперішнього часу пасив-

ного стану (indefinite passive infinitive) вживається допоміжне дієслово “be”; 3) перед перфектним інфінітивом пасивного стану (perfect passive infinitive) вживається “have been”.

Причому герундій і дієприкметник теперішнього часу об’єднуються в інгову форму. Звичайно, такий підхід не є ефективним, оскільки перед іменником, на –ing не завжди вживається артикль. [2] У випадках з дієприкметником минуло часу було використано контекстний аналіз, який враховує безпосередньо текстові умови вживання мовної одиниці, протиставляючи позамовному фактору виявлення значення мовних елементів. Ознака структурного оформлення мовної одиниці є невід’ємною рисою контексту. [1] Завдяки контекстному аналізу були встановлені такі правила:

1) якщо перед participle II стоїть займенник (we) або сполучник (then, when, which), то це минулий час активного стану (past indefinite active voice); 2) якщо перед participle II стоїть допоміжне слово was/were, то це минулий час пасивного стану (past indefinite passive voice); 3) якщо перед participle II стоїть допоміжне слово have/has, то це перфект теперішнього часу активного стану (present perfect active voice); 4) якщо перед participle II стоїть have been/has been, то це перфект теперішнього часу пасивного стану (present perfect passive voice); 5) якщо перед participle II стоїть had/had been, то це перфект минулого часу активного/пасивного стану (past perfect active/passive voice).

Форму дієслова можна також визначити за положенням артикля: якщо ж артикль стоїть після такої словоформи, то це або інфінітив, або теперішній час активного стану (present indefinite active voice).

III. Етапи опрацювання тексту анотації

Після завантаження тексту анотації з файлу або копіювання його у вікно для тексту анотації (тобто ліве вікно у вікні програми), цей текст проходить такі етапи опрацювання:

- всі пунктуаційні знаки окрім тире в тексті заміняються на пробіл; два пробіли підряд заміняються на один;
- за пробілом слова в тексті розділяються на окремі елементи і записуються в масив;
- відкривається з’єднання з базою даних, здійснюється запит на вибірку всіх записів з бази даних;
- елементи масиву порівнюються з записами бази даних;
- програма знаходить одинакові елементи (тобто дієслівні форми з бази даних та масиву;
- якщо потрібно, то перевіряється лівостороннє оточення дієслівної форми;
- якщо була використана функція “Analyze”, залежно від оточення та колонки в базі даних, до якої належить дієслівна форма, визначаються її морфологічні ознаки та приписується код;
- результати аналізу записуються в таблицю;
- якщо була використана функція “Find Possible Verb Forms”, то знайдена дієслівна форма виділяється червоним.

Для наочності на рисунку подано інтерфейс програми, де виконано обидві функції програми. Як видно, у лівому вікні показані результати використання першої функції; праве вікно розміщує результати використання другої функції. У правому вікні розміщена таблиця, де в алфавітному порядку виписані дієслівні форми, їх морфологічні характеристики (час, стан, особа, число), морфологічний код і контекст у якому зустрілася окрема дієслівна форма.

Інтерфейс програми з прикладом виконаного завдання.

Висновок

На сучасному етапі правильність результатів програми можна вважати вищими за 60 %, оскільки використання бази даних зменшує ймовірність омонімії між іменником та інговою формою. Також, завдяки контекстному аналізу знімається значний відсоток випадків омонімії між іменником множини та дієсловом теперішнього часу третьої особи однини, і дієприкметником минулого часу та деякими особовими формами дієслова.

Література

[1] Дарчук Н.П. Комп’ютерна лінгвістика (автоматичне опрацювання тексту): підручник/ Н.П. Дарчук. – К.: Видавничо-поліграфічний центр "Київський університет", 2008. – С.13 - 94.

[2] Милых Н.Г. Глоссарий грамматических терминов/ Н.Г. Милых, В.И. Перебейнос, Э.П. Рукина // Английское спряжение: Система и функционирование (справочник). – К.: КНЛУ, 2003. – С. 6 - 43.

[3] Tang X. English Morphological Analysis with Machine-learned Rules. – http://www.aclweb.org/anthology/

[4] Trost H. Morphology/ H. Trost// The Oxford Handbook of Computational Linguistics/ [ed. By R.Mitkov]. – Oxford: Oxford University Press, 2003. – P.25 - 48.