УДК 801.009

В.Б. Толубко, д-р техн. наук, проф.,
Л.О. Литвиненко, здобувач,
КНУ імені Тараса Шевченка

РОЗРОБКА МОДЕЛІ АВТОМАТИЧНОГО СИНТАКСИЧНОГО АНАЛІЗУ І СИНТЕЗУ ТЕКСТУ В СИСТЕМІ МАШИННОГО ПЕРЕКЛАДУ

У статті запропоновано процедура формалізації синтаксису на основі знання-орієнтованого підходу. Особливістю процедури є декларативне представлення правил синтаксису і збереження змістової цілісності тексту. Показано практичну реалізацію запропонованого підходу при побудові систем машинного перекладу.

Ключові слова: автоматичний синтаксичний аналіз, декларативне представлення правил синтаксису, змістова цілісність тексту, система машинного перекладу.

Актуальність дослідження. Автоматичний синтаксичний аналіз (АСА) є невід'ємною складовою будь-якої системи автоматичної обробки текстової інформації, але повною мірою він реалізується в системах машинного перекладу (СМП). Не дивлячись на те, що роботи в напряму алгоритмізації синтаксису мови ведуться понад півстоліття [1-3], результати СМП ще досить скромні. Причини цього, на наш погляд, пов'язані із загальними проблемами автоматичної обробки природно-мовних текстів (ПМТ), яким притаманні (в тому числі і на синтаксичному рівні мови) багатозначність і невизначеність. Крім того, в межах побудови СМП дослідники розглядають, як правило, синтаксичну структуру окремих речень, а не текстів, що дозволило б і розв'язати значну частину невизначеності і багатозначності синтаксичного рівня мовної системи. Одним із шляхів підвищення якості автоматичного перекладу автори вважають реалізацію АСА на основі знання-орієнтованого підходу до розробки СМП.

Постановка задачі дослідження. Знання-орієнтований підхід до розробки СМП базується на принциповому положенні, що предметом аналізу виступають наявні в текстовій інформації знання про навколишній світ (предметну область).

Синтаксичний аналіз в СМП має багатофункціональне призначення, а саме:

- усунення лексико-граматичної омонімії, отриманої на етапі морфологічного аналізу;
- розпізнавання термінів і понять, що є словосполученнями;
- побудова синтаксичної структури речення.

Кінцевою метою синтаксичного аналізу є представлення синтаксичної структури речення, яке є придатним для семантичного аналізу.

Об'єктом аналізу синтаксичного рівня мовної системи є синтаксичні закономірності взаємодії лексем у межах речення і речень у межах цілісного тексту. Важливим для знання-орієнтованого підходу є збереження змістової цілісності тексту.

Вихідними даними АСА є результати роботи попередніх модулів: доморфемного і морфологічного аналізу та апріорно задані словники синтаксичних правил, які визначають ознаки синтаксичного поєднання лексем у словосполучення.

Загалом, між синтаксичною і семантичною структурою є однозначний зв'язок. Так, синтаксичні відношення не існують без семантичних, які в свою чергу реалізуються в заданій предметній галузі. Тому, інформаційна база даних включає словник лексико-семантичних валентностей дієслів, який обумовлює ознаки найбільш вірогідного оточення, та словник семантичних інтерпретацій, який на основі розпізнаних синтаксичних правил визначає стійкі словосполучення і поняття в заданій предметній області.

Викладення результатів дослідження.

Синтаксичний аналіз загалом передбачає 3 етапи:

1) контекстно-синтаксичний аналіз;
2) синтаксичний аналіз простого речення;
3) міжфразовий синтаксичний аналіз (складні речення аналізуються як частковий випадок міжфразового синтаксису).

Перший етап відповідає контекстно-синтаксичному аналізу. Задачею цього етапу є визначення іменникових груп в межах однієї синтагми, що можуть позначати терміни (цілісні поняття) в заданій предметній області (ПО). На цьому етапі аналізу речення прочитується з кінця, тобто від маркера, що позначає кінець речення, в межах кожної синтагми словосполучення перевіряються на відповідність контекстних синтаксичних правил (узгодження, керування, прилягання) й укладання в термінах граматики безпосередніх складників синтаксичних сполук, для кожного з яких визначається головне слово. Продемонструємо роботу контекстно-синтаксичного аналізу на прикладі такого речення:

Оборона України, захист її суверенітету, територіальної цілісності і недоторканності покладаються на Збройні Сили України.

Результат контекстно-синтаксичного аналізу представлений на рис.1.

```
MP=>
  MC=>
  [Rule У1, mainword Left]=>
Оборона[1*211000002]України[1*212000002]
(L06),
MC=>
захист[1*211000002/1*214000002]
[Rule У2, mainword Left]=>
її[5*212300000]
суверенітету[1*112000002/1*113000002]
  (L06),
```

```
 MC=>
 [Rule C1, mainword Right]=>
територіальної[2*212000000]цілісності[1*212000002]
і [24*000000000]
недоторканності[1*212000002]
покладаються[9*014322010/9*012522010/9*016322010]
[Rule C2, mainword Left]=>
на [23*004000000]
[Rule C1, mainword Right]=>
Збройні [2*224000000]
        [Rule У1, mainword Left]=>
Сили[1*221000002/1*224000002]України[1*212000002]
КР.
```

Рис.1. Результат контекстно-синтаксичного аналізу речення

З рис.1 видно, що на першому етапі виділилося чотири словосполучення (*оборона України, її суверенітету, територіальної цілісності, Збройні Сили України), для кожного словосполучення визначено правило, за яким сформовано словосполучення.* Визначення синтаксично поєднаних слів здійснюється за контекстно-синтаксичними правилами (узгодження, керування і прилягання), які представляються декларативно. Правила, які використані в даному прикладі, представлені в таблиці 1. В таблиці визначені: тип синтаксичного правила, лексико-граматичні ознаки їх прояву в контексті (на рис. 1 – це цифрова інформація, що супроводжує кожну лексему в квадратних дужках. Вона визначається на етапі морфологічного аналізу), головне слово для сполуки.

Таблиця 1

Форма представлення правил контекстного поєднання лексем у межах синтагми

Який клас	З яким класом	За якими ознаками			Тип СП	Головне слово
		рід	числ.	Відм.		
Правило узгодження						
2*	1*	+	+	+	С1	1*
1*	23*			+	С2	23*
…	**…**	**…**	**…**	**…**	**…**	**…**
Правило керування						
1*	1*			2	У1	1*/1
1*	1*			3	У2	1*/1
…	**…**	**…**	**…**	**…**	**…**	**…**

В даній таблиці використані такі позначення: 2* – код лексико-граматичного класу (1* – означає іменник; 2* – прикметник; 23* – прийменик); «+» означає, що словоформи мають однакове значення (це перші три позиції після *), 1*/1 визначає, що позицію головного слова у словосполученні, якщо коди лексико-граматичних класів однакові.

Результат контекстно-синтаксичного аналізу подається на модуль інтерпретації, який виявляє терміни і стійкі словосполучення за словником стійких словосполучень і формує єдине нероздільне поняття (для нашого прикладу: *Збройні Сили України*).

На другому етапі автоматичного синтаксичного аналізу (побудова дерева синтаксичного підпорядкування) визначаються підмет, присудок і другорядні члени речення. На цьому етапі також синтагми (якщо це просте речення) включаються до єдиної синтаксичної структури речення, при цьому розділові знаки убираються, оскільки при синтезі речення засобами вихідної мови, вони можуть передаватися інакше. Правила синтаксичного аналізу мають аналогічний вигляд (табл. 2), але аналізуються лише головні слова у визначених, на попередньому етапі словосполученнях. Алгоритм аналізу починає з виявлення підмета і присудка. Фрагмент правил, які використані у нашому прикладі, представлений в таблиці 2.

Таблиця 2

Форма представлення синтаксичних правил для другого етапу синтаксичного аналізу

Який клас	Через який кл.	З яким класом	За якими ознаками			Члени речення	
			рід	числ.	Відм.	ПІДМЕТ	ПРИСУДОК
Правила узгодження підмета і присудка							
1*		9*	+	+		1*	9*
1*+1*		9*		2		1*+1*	9*
Правила визначення однорідних членів речення							
1*	**,**	1*			**+**		
1*	24*(і)	1*			**+**		
Правила визначення другорядних членів речення							
9*	23*	1*			+	1*	ДОДАТОК
…	…	…	…	…	…	…	…

Визначені дієслівні синтаксичні конструкції, що спрацювали за таблицею 2, перевіряються на словнику лексико-семантичних валентностей. Цей словник може містити як прості синтаксичні конструкції (наприклад: *грати роль*, *мати значення*), так і складні (наприклад: *забезпечувати недоторканість від дій або впливу противника*).

Словник лексико-семантичних валентностей, з одного боку, дозволяє уникнути хибних синтаксичних конструкцій (з точки зору семантики), з іншого – сформулювати вимоги до довжини представлення стійких дієслівних конструкцій у перекладному словнику.

На останньому етапі аналізуються складні речення і текст загалом, оскільки, як вже зазначалося, для забезпечення адекватного перекладу важливим є збереження змістової цілісності тексту. Метою цього етапу є визначення так званих лексичних конекторів (тобто словоформ/словосполучень), які визначають на лексичному рівні логіко-семантичні відношення між реченнями в тексті. Так, приклад нашого речення був вибраний із тексту, що представлений на рис. 2. Курсивом на рисунку виділені лексичні конектори, через які реалізується семантична зв'язаність тексту, це можуть бути:

- повторювані слова (1 і 2 речення: *захист* – *захист*);
- анафоричні зв'язки (3 і 4 речення: *президент* – *він*) тощо.

Збройні Сили України

Захист суверенітету і територіальної цілісності України, забезпечення її економічної та інформаційної безпеки є найважливішими функціями держави, справою всього Українського народу. Оборона України, *захист* її суверенітету, територіальної цілісності і недоторканності покладаються на Збройні Сили України.

Президент України є Верховним Головнокомандувачем Збройних Сил України. *Він* призначає на посади та звільняє з посад вище командування Збройних Сил України, інших військових формувань, здійснює керівництво у сферах національної безпеки та оборони держави, очолює Раду національної безпеки і оборони України.

Рис.1. Приклад автентичного тексту

Виявлення конекторів є важливою ланкою для автоматичного перекладу, оскільки, наприклад, визначення відношень між 3 і 4 реченнями дозволяє правильно перекласти англійською мовою займенник ***він*** (he, а не it).

Безпосередньо класифікація і логіко-семантична інтерпретація конекторів відбувається на семантичному рівні аналізу тексту. Результати АСА, в свою чергу, є вхідними даними для автоматичного семантичного аналізу.

Висновки. Запропонований трьох-етапний синтаксичний аналіз дозволяє не тільки будувати синтаксичну структуру речень, але й зберігати (через міжфразовий синтаксис) цілісність тексту, що є дуже важливим для адекватності перекладу.

Занурення проміжних результатів (після кожного етапу) синтаксичного аналізу у ПО дозволяє забезпечити реалізацію знання-орієнтованого підходу до розробки СМП.

Декларативне подання правил синтаксичного аналізу дає можливість звести його в плані програмування до обробки таблиць та реалізувати відомий принцип програмування: відокремлення даних від алгоритму їх обробки, що робить його відкритим як щодо нових мов, так і щодо «нових» прикладних задач з обробки текстової інформації.

Список використаних джерел

1. Теньер Л., Основы структурного синтаксиса. – М., 1988. – 427 с.
2. Дарчук Н.П. Комп'ютерна лінгвістика (автоматичне опрацювання тексту). – К.: ВПЦ «Київський університет, 2008. – 351 с.
3. Севбо И.П. Графическое представление синтаксических структур и стилистическая диагностика. – К., 1981. – 112 с.

Надійшла до редколегії 12.02.13

В.Б. Толубко, д-р техн. наук, проф.,
Л.А. Литвиненко, соискатель,
КНУ имени Тараса Шевченко

РАЗРАБОТКА МОДЕЛИ АВТОМАТИЧЕСКОГО СИНТАКСИЧЕСКОГО АНАЛИЗА И СИНТЕЗА ТЕКСТА В СИСТЕМЕ МАШИННОГО ПЕРЕВОДА

В статье рассматривается процедура формализации синтаксиса на основе знание-ориентированного подхода. Особенностью процедуры является декларативное представление правил синтаксиса и сохранение смысловой целостности текста. Показана практическая реализация предложенного подхода при построении систем машинного перевода.

Ключевые слова: автоматический синтаксический анализ, декларативное представление правил синтаксиса, смысловая целостность текста, система машинного перевода.

V.B. Tolubko, Dr., Professor,
L.A. Lytvynenko, Candidate for a Degree

DEVELOPMENT OF THE AUTOMATIC PARSING AND TEXT SYNTHESIS MODEL IN THE MACHINE TRANSLATION

The article describes the procedure of formalizing syntax on the basis of a knowledge based approach. A specific feature of the procedure is a declarative representation of the rules of syntax and semantic preservation of the integrity of the text. The practical implementation of the proposed approach in the construction of machine translation systems is shown.

Keywords: automatic parsing, declarative representation of the rules of syntax, semantic integrity of the text, machine translation system.