|  |  |
| --- | --- |
| Національний університет "Києво-Могилянська Академія" | |
|  | Департамент Коп‘ютерних Технологій  Кафедра інформатики |
| **Семантико-синтаксична модель автоматичного перекладу** | |
|  | Контрольна робота з курсу "Лінгвістичне забезпечення інтелектуальних систем" за осінній триместр 1999/2000 н. р.р. студентів ДКТ-5  Підгорного Свєтослава, Сігаєвої Марини |
|  | Викладач: доц. Штерн Ірен Борисівна |
| Київ 1999 | |

План

1. Система французько-російського автоматичного перекладу
2. Етапи роботи системи:
   1. підготовка тексту та графематичний аналіз;
   2. морфологічний аналіз;
   3. попередній синтаксичний аналіз;
   4. синтаксичний аналіз;
   5. семантичний аналіз;
3. Підготовка до синтезу й синтез;
4. Висновки
5. Література

Система французько-  
російського  
автоматичного  
перекладу

Розглянемо автоматичний переклад на прикладі системи ФРАП (французько-російського автоматичного перекладу). Розробка цієї системи стала значним кроком вперед у розвитку систем автоматичного перекладу.   
Система ФРАП складається з чотирьох компонентів – графематичного, морфологічного, синтаксичного та семантичного. Повний аналіз має проходити через усі чотири компонента, синтез – через три (без графематичного).   
Головний компонент у системі – семантичний, він визначає структуру та межі всіх інших компонентів. Основою семантичного компоненту є спеціально розроблена семантична мова. Але система ФРАП не досягає повного розуміння тексту. Реально досяжні границі розуміння обмежено в рамках одного окремого речення. Тобто відсутнє звертання до змісту інших речень чи до енциклопедій відповідної галузі знань.

Етапи роботи  
системи

При роботі система ФРАП проходить етапи графематичного, морфологічного, попереднього синтаксичного, синтаксичного та семантичного аналізу. Розглянемо детальніше роботу на кожному з етапів.

Підготовка тексту та  
графематичний аналіз

Деякі особливості реального тексту можуть створювати значні проблеми при автоматичному перекладі. Як-то: наявність різного роду формул, схем, креслень, діаграм тощо. Найкраще рішення – опустити їх при подальшому аналізі, залишивши в тексті спеціальну позначку, а потім вставити у відповідні місця при синтезі вихідного тексту.   
Крім того, слід виділити різного роду посилання (на інші сторінки, приклади, малюнки тощо). Вони не входять в структуру речення, а їх значення можливо доведеться змінити у вихідному тексті перекладу.  
Може виявитися, що реальний текст є складно-організованою структурою, тобто в нього можуть бути вставлені інші тексти (анотація, зміст, виноски, підписи до малюнків, цитати, бібліографія). Деякі з таких текстів доведеться прив‘язати до певних елементів (наприклад, підписи до малюнків), деякі передрукувати у вихідному тексті без перекладу (бібліографія).

Отже, для того, щоб врахувати властивості реального тексту потрібно ввести етап графематичного аналізу. Крім вже зазначених дій графематичний аналіз визначає та перевизначає границі слів. Так від слів відділяються розподільники, що стояли поруч: лапки, дужки, крапки, коми. Проводиться обробка слів, що пишуться через дефіс або апостроф. Формується гіпотетична морфолого-синтаксична та семантична інформація для слів-формул: наприклад, 25-ий буде представлено як слово 25 з підкласом "порядковий числівник" і семантичною характеристикою "квантор (25, х)", що дозволить йому далі брати участь у синтаксичному та семантичному аналізі.

Морфологічний аналіз

Морфологічний аналіз проводиться двічі: для слів, основи яких знайдено у словнику, та для слів, основи яких не знайдено. Морфологічний аналіз полягає у виявленні інформації про форму даного слова за його закінченням, яке залишилося після відсікання основи. Якщо ж слово непізнане, то на основі аналізу його кінцевих літер формується гіпотетична інформація щодо класу, до якого воно може належати. Також проводиться попереднє зняття морфологічної омонімії на основі розгляду лінійного контексту.

Попередній синтаксичний  
аналіз

Оскільки головним завданням семантичного компоненту є інтерпретація вузлів та зв'язків синтаксичної структури, вузли синтаксичної структури повинні максимально наближатися до одиниць семантичної структури. Це означає, що вузлами синтаксичної структури повинні бути лише повнозначні слова, а не допоміжні слова чи службові частини мови. Завдання попереднього синтаксичного аналізу й полягає в тому, щоб утворити вузли майбутньої синтаксичної структури. Для цього використовують техніку локального аналізу. В один вузол збираються звороти на основі спеціального словника зворотів. В один вузол збираються також аналітичні граматичні дієслівні комплекси. При цьому вузол представлено повнозначним словом комплексу, а вся інформація щодо часу, числа, особи, наявності заперечення тощо переходить в інформацію до вузла. Також утворюються іменні вузли, де інформація про прийменник, артикль, детермінатив стає частиною інформації про вузол.

Синтаксичний аналіз

Теоретичною основою синтаксичного аналізу є граматика членів речення. Крім граматики до засобів аналізу також належать чотири словника: словник основ, словник зворотів, семантичний словник і словник конструкцій.   
Головне завдання синтаксичного аналізу — побудувати всі зв'язки між вузлами в реченні на основі заданого Аналізатора. Аналізатор не використовує ніякої семантичної інформації, тобто зв'язки між вузлами мають лише функціонально-синтаксичний характер. Цим пояснюється їх зовсім невелика кількість (основних зв'язків тільки 14, з різновидами – близько 30).   
Інша частина синтаксичного аналізу – розбиття складних речень на прості, тобто побудова максимальних одиниць синтаксичної структури. Ця процедура проводиться за допомогою введення в синтаксичну структуру нетермінального вузла символу речення. При цьому підрядні та сурядні сполучники стають, подібно до прийменників при іменниках, частиною граматичної характеристики нетермінальних вузлів.   
Наступне завдання синтаксичного аналізу – з'ясувати позиції для всіх вузлів, що не мають свого місця в синтаксичній структурі (наприклад, слово стоїть у позиції 2 між підметом та присудком). За результатами такого позиційного аналізу вираховуються зв'язки між усіма "слабкими", тобто не зв'язаними Аналізатором вузлами.   
Ці гіпотетичні зв'язки далі інтерпретуються в семантичному компоненті системи.   
Частина синтаксичного компоненту, що виконує перераховані вище функції, називається первинним синтаксичним аналізом. Первинний синтаксичний аналіз встановлює та переводить в синтаксичну структуру лише "сильні" зв'язки, а для "слабких" прораховує всі позиційно можливі варіанти.   
Результатом роботи первинного синтаксичного аналізатору буде структура, що в загальному випадку не є деревом: у ній можуть бути розриви та альтернативні зв'язки. Така структура називається первинним синтаксичним представленням.   
Після роботи первинного синтаксичного аналізатору починається робота первинного семантичного аналізатору, що інтерпретує зв‘язки первинного синтаксичного представлення, а також гіпотетичні слабкі зв‘язки. Інтерпретовані (підтверджені) зв‘язки розглядаються вторинним синтаксичним аналізатором, який враховує властивості правильної синтаксичної структури. У результаті відкидаються зайві синтаксичні інтерпретації первинного синтаксичного аналізу, а також вводяться нові синтаксичні зв‘язки, що відповідають інтерпретованим слабким зв‘язкам.   
Отже, наявність семантичного компоненту в системі суттєво впливає на структуру та межі синтаксичної складової системи. Границю між синтаксисом та семантикою зсунуто в бік семантики.

Семантичний аналіз

Бажано, щоб у системі було щонайменше два рівня семантичного аналізу: первинний, що дає інтерпретацію вузлів та зв‘язків, утворених на попередніх рівнях, у термінах спеціальної синтаксичної мови, і власне сам семантичний аналіз, що працює на основі заданої граматики змісту.   
У системі ФРАП реалізовано лише первинний семантичний аналіз. До його функцій належать:

1. **Інтерпретація синтаксичних вузлів**, тобто вибір лексичних значень слів, або утворення семантичних вузлів. Ця процедура проводиться на базі семантичного словника, в якому кожному синтаксичному вузлу відповідає множина семантичних вузлів.
2. **Інтерпретація сильних синтаксичних зв‘язків** (зв‘язків, встановлених Аналізатором). Вона відбувається на основі словарної інформації про моделі керування кожного семантичного вузла. Модель керування в даній системі визначається як набір смислових валентностей семантичного вузла разом з відповідними морфолого- синтаксичними способами реалізації їх у фразі. Розпізнавання моделі керування у фразі технічно зводиться до застосування словарних фільтрів до структурних та семантичних характеристик вузлів фрази та зв‘язків, знайдених на всіх попередніх етапах.
3. **Фільтрація синтаксичних зв‘язків.** Інтерпретуючи синтаксичні зв‘язки, встановлені Аналізатором, семантичний аналіз підтверджує їхню правильність. Відмовляючи деякому зв‘язку в інтерпретації, семантичний аналіз переводить його з підкласу сильних (куди він потрапив через формальну подібність до сильних) до підкласу слабких, які інтерпретуються пізніше.
4. **Перерозподіл вузлів і зв‘язків при перекладі.** Він потрібен на тих ділянках, що не піддаються прямому перекладу (вузол – у вузол, зв‘язок – у зв'язок). Такі ділянки є конструкціями, та перекладаються за допомогою спеціального словнику конструкцій.
5. **Вибір російського перекладного еквівалента.** Для кожного семантичного вузла в словнику визначено російський еквівалент. Як умови вибору російського еквівалента можна використовувати будь-які відомості, отримані при роботі системи. Так наприклад, навіть тип смислових опущень, що супроводжують деяке слово у фразі, є смисловим контекстом, що підказує той чи інший переклад.
6. **Інтерпретація слабких зв‘язків.** Для слабких (гіпотетичних) зв‘язків сформована таблиця їхніх позицій та можливих синтаксичних "хазяїв". Інтерпретація слабких зв‘язків необхідна для коректного перекладу, оскільки вони не перекладаються через моделі управління.   
   Інструментом аналізу слабких зв‘язків є словник відносин. На вхід подається: сам лексико-граматичний елемент, що пов‘язує два вузли (це може бути прийменник, сполучник чи розподільний знак), набір граматичних та синтаксичних характеристик для слабокерованого вузла та його можливого хазяїна. На виході – смислове відношення, а також російські синтаксичні кореляти. Також може видаватися уточнений французький синтаксичний зв‘язок.   
   При неповній інформації на вході буде видано наближений, грубий переклад.
7. **Експлікація смислової структури речення.** Проявляються значення вузлів, називаються іменами смислових відношень встановлені у фразі синтаксичні зв‘язки. Експлікація проводиться і для тих частин смислу, що не виражені у реченні, але передбачаються структурою висловлювання або смислом використаних у ньому слів.

Підготовка до  
синтезу й синтез

На вхід синтезу повинно передаватися семантичне представлення у вигляді дерева з російськими семантичними вузлами. Проте, результат первинного семантичного аналізу не завжди має вигляд дерева. Тому для підготування семантичного представлення до синтезу потрібно добудувати це дерево та прийняти рішення про вибір російських еквівалентів. Слова, не знайдені у французькому словнику, будуть надходити без перекладу.   
Синтез російської фрази має пройти через такі етапи:

* побудувати синтаксичне представлення російської фрази;
* визначити лінійну послідовність вузлів;
* визначити морфологічні характеристики вузлів;
* провести морфологічний синтез.

Джерелом побудови російського синтаксичного представлення є, насамперед, інформація про моделі управління російських слів з російського семантичного словника. Оскільки опис моделей управління у французькому та російському словниках задано через набір смислових валентностей, на етапі побудови російського синтаксичного представлення природньо долаються невідповідності між цими наборами. Інші етапи синтезу значно простіші.

Висновки

Процес перекладу тексту — це складний інформаційний процес, який не можна звести до простої заміни слів однієї мови словами іншої. Необхідно проводити аналіз структури, враховувати синтаксичну, а найголовніше – семантичну будову тексту.   
Для системи ФРАП розробниками було передбачено кілька режимів роботи:

* послівний переклад;
* синтаксичний переклад;
* семантичний переклад, який проходить через усі попередні етапи.

Зрозуміло, що відносно якісний переклад може забезпечити тільки третій режим роботи, що має спеціальні механізми розуміння змісту тексту.

Література

1. Н. Н. Леонтьева, С. Л. Никогосов. "Система ФРАП как информационная система", "Актуальные вопросы практической реализации систем автоматического перевода", 1982, Москва, стр. 134 — 166.
2. Е. Г. Соколова, И. М. Кудряшева. "Компонент синтаксического анализа в системе ФРАП, "Актуальные вопросы практической реализации систем автоматического перевода", 1982, Москва, стр. 167 — 183.