Національний університет “Києво-Могилянська Академія”

Реферат

з курсу

“**Лінгвістичне забезпечення інтелектуальних систем**”

на тему:

# “Синтаксичний аналіз у системах автоматичного перекладу: концепції та алгоритми”

студентки ДКТ-5

Фіялка Світлани

Київ-1999

**Зміст:**

Компонент синтаксичного аналізу в системі ФРАП 3

Синтаксис і семантика в системі англо-російського перекладу 11

Алгоритм синтаксичного аналізу мовних текстів 18

Компонент синтаксичного аналізу в системі ФРАП

В системі французько-російського автоматичного перекладу (ФРАП) процес аналізу організований у вигляді послідовності працюючих компонент, незалежних одна від одної в тому смислі, що кожна з них здатна приймати на вхід і перетворювати лінгвістичну інформацію певного виду. Синтаксичний аналіз (СінАн) організований і описаний у вигляді процесу перетворень з заданою послідовністю дій.

Потужність системи АП

При порівнянні систем АП між собою важливим чинником виступає *потужність системи АП*, тобто її здатність використовувати при аналізі, перекладі і синтезі лінгвістичну інформацію певних типів, причому мається на увазі саме систематичне використання інформації в спеціальних структурах. Так, в системах автоматичного перекладу першого покоління систематично використовувався лише один тип інформації – морфологічна інформація. Системи третього покоління, про які й буде тут йти мова, виробляють щонайменше три вида інформації про текст – морфологічну, синтаксичну і семантичну інформацію.

Рівнева і “компонентна” стратегія побудови системи АП

При розробці системи АП третього покоління можливі дві стратегії, які пов’язані з двума поглядами на взаємодію різних видів інформації в тексті.

Перша стратегія – *рівнева* – передбачає ієрархічну організацію взаємодії різних видів інформації. При такому підході для кожного рівня будується представлення текста, в якому вся інформація, шо представлена в тексті, інтерпретується засобами специфічної для даного рівня мови. Процес аналізу будується як “переклад” представлення одного рівня в представлення наступного рівня. Для кожного рівня формується поняття правильної структури. Аналіз на кожному рівні складається з побудови всіх можливих для даного представлення структур за допомогою набору синтагм і в пошуці серед них правильних. Пошук правильних структур здійснюється за допомогою фільтрів, які можуть бути включені до складу синтагм або описані окремо у вигляді статичних правил. Такий підхід аналізу називається *фільтровим*. При рівневому підході кількість рівней не обмежується властивостями мовних одиниць. Вузлами структури спочатку виступають словоформи, потім значення слів. В ідеалі представлення на найглибшому рівні буде виступати експлікацією смислу в термінах елементарних смислових одиниць.

Інша стратегія – *неієрархічна* – виходить з того, що в тексті можна виділяти різні види інформації (синтаксичну, семантичну, комунікативну), і кожна з цих типів інформації має самостійну значимість, тобто не може бути перекладена в інший вид для даної глибини інтерпретації тексту. При цьому глибина інтерпретації визначається мінімальною одиницею аналізу. Мінімальною глибиною аналізу вважається та, при якій одиницею аналізу виступає повнозначне слово. Така степінь глибини має назву *першого етапу інтерпретації тексту*. На цьому етапі в кожному представленні тексту експліцірується лише один вид інформації: наприклад, в синтаксичному представленні – інформація про те, яким членом речення є слово і яке слово є головним по відношенню до нього, в семантичному представленні – інформація про значення слів і семантичних відношеннях між ними. При такому підході експлікацією інформації кожного виду, тобто побудовою відповідного представлення в системі аналізу, займається спеціальний компонент. Він може використовувати результати роботи інших компонентів, які сформульовані в зрозумілому для нього вигляді. При такому підході кількість основних компонентів системи співпадає з кількістю різних видів інформації.

Принципи і засоби СінАн

Теоретичною основою СінАн, що реалізований в системі ФРАП, є *граматика членів речення*. До засобів аналізу крім граматики відноситься також словник. В системі ФРАП на етапі аналізу використовуються 4 словника: *словник основ*, *словник оборотів*, *семантичний словник* та *словник конструкцій*. Словники оборотів і конструкцій використовуються у відповідних компонентах для аналізу словосполучень. Семантичний словник містить повний набір інформації різних видів про значення лексичної одиниці, а також засоби синтаксичної реалізації її валентностей. СінАн працює після компоненти СемАн (Семантичний аналіз).

Синтаксичне представлення (СінП)

СінП будується для фрази, оскільки саме вона є об’єктом СінАн. СінП складається з *вузлів* і *відношень*, що задані на множині вузлів.

Вузлами виступають лексичні одиниці, що мають при собі морфолого-синтаксичну інформацію, в тому числі сполучення службових слів з повнозначними та фразеологічні словосполучення, що об’єднані в один вузел на досинтаксичному етапі аналізу і в процесі СінАн, а також знаки пунктуації. Крім того введені штучні вузли.

На множині вузлів задано відношення лінійного порядку і відношення “несумісності”. В процесі СінАн між вузлами встановлюються зв’язки залежності з функціональною міткою та додаткові зв’язки ref та dist, а також відношення тотожності. Функціональні зв’язки утворюють синтаксичну структуру простих речень; зв’язки ref з’єднують штучний вузол, що символізує речення або фразу, з вершиною цього речення або з головним реченням фрази; зв’язок dist поєднує дві частини складного слова або парні знаки пунктуації. Відношення тотожності використовуються, наприклад, при аналізі твору.

Способи відображення неоднозначності результатів аналізу в СінП

В компоненті СінАн ситеми ФРАП для фрази будується одне представлення, яке і містить в собі всі види синтаксичної неоднозначності результатів аналізу. Нерозв’язана на ранніх етапах неоднозначність зберігається в представленні до тих пір, поки результати роботи наступного етапу аналізу не дозволять її виявити.

В СінП зберігаються такі неоднозначності:

* Неоднозначність функціональних зв’язків – наявність зв’язків, що порушують синтаксичну структуру.
* Неоднозначність результатів морфологічного аналізу словоформи – морфологічна омонімія. В цьому випадку використовується складний номер вузлу: перше число позначає порядковий номер вузла у фразі, інше – номер омоніма.
* Неоднозначність поділу на вузли. Для представлення цієї неоднозначності використовується відношення несумісності.

Компонент СінАн системи ФРАП

Задача компонента СінАн

Задачею компонента СінАн є побудова для фрази СінП в термінах граматики членів речення. Для цього необхідно: 1) сформувати множину вузлів – членів речення; 2) виявити синтаксичну функцію кожного з цих вузлів.

1. *задача.* Про деякі слова фрази заздалегідь відомо, що вони не є членами речення. Для французької мови це: допоміжні слова – частини аналітичних форм часу, залогу, ступенів порівняння, заперечувальні частки, компоненти фразеологічних словосполучень, які не можна розкласти, артикль, сполучники, прийменники. Крім того, заздалегідь відомо, що членами речення є прості речення в складі складного. Таким чином перша задача розбивається на дві підзадачі: 1) елімінування із множини вузлів тіх слів фрази, які не є членами речення; 2) виділення простих речень в складному з формуванням штучних вузлів, що заміщають підрядні речення в головному.
2. *задача.* Передбачає побудову синтаксичної структури.

Компонент СінАн системи ФРАП організований у вигляді процесу, в якому ці дві задачі виконуються одночасово.

Організація компонента СінАн

Компонент СінАн організований у вигляді блоків, які предназначені для розв’язання двух вище зазначених задач. Перша задача виконується в процесі роботи перших двух блоків СінАн: блок аналізу іменникових вузлів і блок аналізу складних речень. Друга задача починає виконуватися одночасово з першою для відповідних типів вузлів. Крім того, розв’язанням цієї задачі повністю зайнятий третій блок – блок аналізу простих речень. СінП можна вважати повністю закінченим лише після перевірки його узгодженості з сементичним словником і уточнення тих частин синтаксичної структури, для яких необхідні відомості, що витікають з результатів СемАн. Останній, четвертий блок СінАн працює з вузлами, синтаксична функція яких вже встановлена.

Інструменти аналізу, що використовуються в компоненті СінАн

Вхідною для компоненти СінАн є послідовність вузлів з морфолого-синтаксичною інформацією, що отримана на попередніх етапах. В процесі СінАн для обробки цієї інформації використовуються такі інструменти аналізу: *аналізатор*, *алгоритми*, *списки правил*, *позиційна таблиця*.

В системі ФРАП аналізатор використовується тричі. З його допомогою на лінійно упорядкованій множині вузлів встановлюються всі можливі зв’язки залежності, що указані в аналізаторі.

Для аналізу в термінах членів речення повністю фільтрована організація СінАн неможлива, оскільки вхідна для СінАн множина вузлів (результат морфологічного аналізу) не є множиною членів речення. Потрібен процес формування вузлів структури членів речення. Таким чином, перші два блоки за необхідністю описані як процес, тобто алгоритмічно. Опис у вигляді алгоритму третього блоку заснований на гіпотезі про значимість порядку обробки вузлів для автоматичної побудови структури в термінах членів речення. Алгоритми працюють з вже сформованою аналізатором множиною всіх можливих гіпотез про зв’язки вузлів, тобто метод фільтрів застосовується, но процесом їх застосування керують спеціальні алгоритми. Повністю фільтровим СінАн стає тільки в процесі взаємодії СінП з СемП.

Списки правил містять правила дозволу омонімії для різних етапів аналізу, типи узгоджуваності, деякі фільтри.

Позиційна таблиця починає формуватися в другому блоці аналізу. В ній міститься інформація про порядок слідування головних вершин та про позиційну характеристику в простому реченні тих його вузлів, для яких СінАн ще не завершений. Вона використовується для визначення правильності набору головних вершин для фрази в кінці роботи другого блоку і в складній фразі для визначення порядку обробки простих речень в процесі роботи третього блоку. Позиційна характеристика вузлів використовується в четвертому блоці.

Блоки аналізу компонента СінАн

Блок аналізу іменникових вузлів

В цьому блоці розв’язуються 4 задачі: 1) встіновлюються функціональні зв’язки app(a,b), де а – прикладка b, та attr(a,b), де а – узгоджене означення b; 2) аналізуються деякі види твору; 3) структурно відновлюються елідіровані вершини іменникових груп; 4) елімінуються з представлення прийменники, артиклі і проаналізовані сурядні сполучники.

В першій частині блоку синтаксичний аналізатор встановлює зв’язки app та attr. Встановлюються також допоміжні зв’язки prep(a,b), де а – ім’я або інфінітив, b - прийменник та det(a,b), де а – артикль, b – ім’я.

В дугій частині блоку аналіза іменникових вузлів з представлення фрази елімінуються прийменники, артиклі та проаналізовані сурядні сполучники, які поміщаються в інформацію к пов’язаним до них повнозначним словам; відповідно елімінуються і допоміжні зв’язки.

Блок аналізу складних речень

Перед початком роботи цього блоку в представленні аналізатором встановлюється зв’язок dist та всі можливі функціональні зв’язки, крім вже встановлених в першому блоці та сурядних.

Блок аналізу складних речень складається з двох частин.

В першій частині обробляються вузли двух видів: 1) особові форми дієслова, предикати, предикативні формули, тобто вузли, які можуть виконувати функцію присудка; 2) підрядні сполучники. Виясняється, які з головних вершин є вершинами підрядних речень, а які - ні. Одночасно в СінП формуються штучні вузли, що символізують підрядні речення, та зв’язок ref, а також формується та частина позиційної таблиці, в якій перелічені головні вершини по порядку їх слідування у фразі і для кожної з них вказано, чи є вона, за відомостями, отриманими в цій частині блоку, вершиною незалежного або підрядного речення.

В другій частині перевіряється правильність складу головних вершин СінП за позиційною таблицею. Представлення, що є правильними, пропускаються в наступний блок аналізу. До них відносяться ті представлення, в яких є не більше однієї незалежної головної вершини, а всі інші – вершини підрядних речень. В цій частині блоку затримуються і підлягають аналізу ті представлення, в яких більше однієї незалежної вершини або є вершини з недозволеною омонімією. Для таких вершин алгоритм звертається до СінП і аналізує лівий контекст. В результаті аналізу контексту можливі такі дії: 1) розділення складного речення на прості при наявності відокремлювача; 2) встановлення сурядного зв’язку між головним реченням фрази і одним з незалежних простих речень при наявності сурядного сполучника; 3) видалення номеру незалежної вершини з позиційної таблиці, що рівнозначно визнанню цієї вершини присурядненою частиною присудка одного з попередніх речень; 4) дозвіл омонімії сполучника або головної вершини. Після виконання однієї з цих дій представлення знову перевіряється на правильність складу головних вершин.

Блок аналізу простих речень

В цьому блоці продовжується оброблення того набору зв’язків, який був встановлений аналізатором перед роботою попереднього блоку і уточнений цим останнім.

Аналіз проводиться за допомогою позиційної таблиці, в якій поступово заповнюються номерами вузлів спеціально відведені для них місця. Для кожного символу простого речення в ній відведено п’ять позицій: 1) між початком речення і найвіддаленішим від присудка лівим актантом; 2) між найближчим до присудка лівим актантом і присудком; 3) між присудком і найближчим до нього правим актантом; 4) між двума актантами; 5) між найвіддаленішим від присудка правим актантом і кінцем речення.

Процес аналізу визначається почергово двума алгоритмами: перший, загальний, керує аналізом всієї фрази; другий – аналізом підрядних речень.

Вхідним пунктом для аналізу є присудок головного речення. Спочатку визначається перша позиція від початку речення до присудка (вважається, що жодного актанта слова ще не знайдено). Потім починається обробка цієї позиції. Якщо в ній є підрядні речення, то в дію вступає другий керуючий алгоритм і аналізується кожне з них в певному порядку. Номера проаналізованих вузлів викреслюються з позиційної таблиці. Коли аналіз підрядних речень закінчений, аналізуються прислівникові, дієприслівникові та інфінітивні синтаксичні обороти. Після цього позиція оброблюється підблоком аналізу твору, який складається з аналізатору, що встановлює лише сурядні зв’язки, та алгоритма обробки результатів роботи цього аналізатора. Тільки після цього серед залишених в позиції вузлів відшукуються актанти присудка і відбувається перерозподіл номерів вузлів в позиційній таблиці між першою та другою позиціями. Аналогічно оброблюється фраза праворуч від присудка.

В цьому блоці встановлюються такі функції вузлів: підмет, пряме доповнення, непряме доповнення. Встановлення інших актантів відбувається після порівняння СінП з семантичним словником, тому четверта позиція в цьому блоці, як правило, не заповнюється.

Четвертий блок СінАн

Цей блок складається з двох частин. В першій частині оброблюються придієслівні займенникові клітики. Ця частина являє собою алгоритм, що використовує правила аналізу, складені на основі способу опису використання цього типу одиниць у французькій мові, запропонованого Л.Н.Іорданською в доповіді на семінарі в ІНФОРМЕЛЕКТРО в 1978р. Новим в цьому способі опису є поняття синтаксично опорного слова. В системі ФРАП був використаний фактичний матеріал. Різниця полягає лише в тому, що в оригіналі правила були сформульовані для аналізу фільтрового типу, а в системі ФРАП вони використовуються по-іншому: елементом синтаксичної структури вважається зв’язок займенника з своїм опорним словом, а всі можливі семантичні господарі вираховуються за синтаксичними зв’язками у відповідності з правилами. Інформація про семантичних господарів використовується в СемАн при заповненні їх валентностей.

Друга частина четвертого блоку призначена для вирахування керуючих для тих вузлів, які є або можуть бути сирконстантами у відповідності з інформацією з позиційної таблиці. В якості керуючих для таких вузлів перераховуються всі можливі кандидати, що допущені синтаксичною структурою вже побудованою частиною СінП і властивостями самої сирконстанти. Друга частина четвертого блоку використовується також для уточнення СінП у відповідності з результатами СемАн.

Синтаксис і семантика в системі англо-російського перекладу

Лінгвістичною базою системи АРАП служить модель природньої мови “Смисл⬄Текст”, в рамках якої запропонований ряд теоретичних принципів та формальних інструментів, які так чи інакше використовуються в системі. В основу системи покладена формальна модель перекладацького процесу, яку схематично можна зобразити так:

T1

T2

C1

C2

П1

П2

?

Тут Т1 – текст першою мовою (вхідною), Т2 – текст другою мовою (вихідною), П1 та П2 – проміжне формальне представлення одного та іншого тексту, С1 та С2 – семантичний запис текстів.

Відповідно до цієї схеми процес перекладу складається з таких етапів: 1) аналіз вхідного тексту з наступною його заміною через проміжні представлення в семантичний запис (Т1=>C1), 2) власне переклад, що відбувається на рівні одного з формальних представлень тексту, що отримані при аналізі, і зводиться до вибору перекладних еквівалентів для одиниць, які утворюють дане представлення вхідного тексту; в результаті будується проміжне представлення для текста перекладу (П1=>П2), 3) синтез тексту перекладу: перехід від проміжного представлення до послідовності реальних словоформ та знаків пунктуації (П2=>Т2), 4) перевірка синтезованого тексту на наявність в ньому небажаної омонімії, яка могла виникнути в процесі синтезу (Т2=>П2), 5) оцінка адекватності перекладу шляхом аналізу проміжного тексту представлення (Т2=>С2) і порівняння отриманого при цьому семантичного запису С2 з семантичним записом С1 вхідного тексту (С1?C2)? 6) у випадку необхідності – редагування тексту перекладу за результатами проведених перевірок і порівнянь: при омонімічності тексту - пошук неомонімічного варіанту шляхом включення системи синонімічного перефразування (П2=>П2); при неадекватності перекладу – повернення до етапу вибору перекладних еквівалентів (пунктирна лінія) і перегляд або окремих компонентів, або всього етапу в цілому.

Для системи АРАП прийнятий порядок роботи, коли спочатку розробляється верхня частина схеми (лінія Т1=>П1=>П2=>Т2).

При цьому підході особливо важливе значення має вдалий вибір рівня П, до якого пред’являються дві протилежні вимоги. З одного боку він повинен бути достатньо “семантичний” (близький до СЗ), щоб в представленні тексту на цьому рівні експліцитно вказувалися або легко з нього виводилися відомості про те, від яких його елементів і яким чином залежить смисл даного тексту, а тим самим і вибір перекладацьких еквівалентів. З іншого боку рівень П повинен бути і достатньо поверховим (близьким до природнього представлення тексту), щоб не виникала потреба здійснювати при аналізі дуже багато надлишкових перетворень, які не впливають на кінцевий результат перекладу.

В системі АРАП в якості проміжного представлення тексту вибрана *комбінована синтаксична структура* (КСС).

Одиницею тексту, для якої в системі АРАП будується КСС і в межах якої здійснюється переклад, виступає речення. КСС речення представляє собою *граф* залежностей особливого роду, який поєднує в собі властивості поверхнево-синтаксичної і глибинно-синтаксичної структур моделі “Смисл⬄Текст”. В вершинах цього графу стоять повні глибинно-морфологічні предствалення (ГМП) словоформ та знаків пунктуації даного речення з інформацією про їх лінійне розташування в тексті відносно один одного у вигляді порядкових позначень. Ребрами графу є стрілки, що символізують синтаксичні та анафорічні зв’язки між словоформами (або знаками пунктуації), при чому відносно стрілок першого типу (синтаксичних) КСС речення виступає деревом. Ті одиниці або сполучення одиниць, які семантично не значимі і при перекладі не повинні отримувати самостійних еквівалентів, заключаються в КСС в кутові дужки.

Множина вершин КСС повинна бути ізоморфна множині словоформ та знаків пунктуації, що є в реченні. Винятки допускаються лише в двох випадках: 1) якщо в даному реченні є фразеологічні обороти, кожному з них ставиться у відповідність одна вершина, незалежно від того, скільки словоформ входить до складу даного обороту; співставлений цій вершині порядковий номер являє собою перелік номерів всіх вхідних словоформ; 2) якщо речення містить еліптичні конструкції, що порушують вимоги зв’язності синтаксичного дерева, вони в процесі побудови КСС перетворюються в повні структури, і в КСС можуть з’являтися ГМП деяких додаткових словоформ, що відсутні у вхідному тексті; вершини, що містять такі ГМП, маркуються порядковими номерами особливого виду (“фіктивними”).

Синтаксичні зв’язки, які встановлюються в КСС між ГМП словоформ та знаків пунктуації, являють собою глибинно-синтаксичні відношення. Вони можуть бути трьох основних видів:

* зв’язки сильного керування (актантні);
* зв’язки слабого керування (означальні);
* координативні або сурядні зв’язки.

Зв’язки сильного керування – це зв’язки між предикатами та їх актантами. Вони позначаються стрілками, що йдуть від предиката до актанта і несуть цифрові індекси. Індекс являє собою номер, під яким актантна валентність предикату, що реалізується даним зв’язком, записана в моделі керування, вказаної для цього предикату в словнику або в граматиці.

Зв’язки слабого керування – це зв’язки, означальні в самому широкому смислі. Вони відображаються стрілками з індексом m (modifying), і йдуть від значуваного слова до означення. На відміну від зв’язків сильного керування вони реалізують валентності не керуючого, а керованого слова.

Координативні зв’язки являють собою зв’язки між яким небудь словом та сурядним сполучником або знаком пунктуації, що має смисл сурядного сполучника. Вони реалізують валентність цього сполучника або знака пунктуації і відображаються стрілками з індексом c (co-ordinate), що напрвлений до сурядного сполучника. Зв’язок між цим сполучником та другим членом сурядної групи вважається зв’язком сильного керування, що реалізує другу валентність даного сполучника. Приклад:

пряма

і

зворотня

теореми

с

2

m

Анафорічні зв’язки – це зв’язки, що встановлюються між словами з співпадаємими денотатами. На анафорічні зв’язки не поширюються ті обмеження, які прийняті у відношенні структури власне синтаксичних зв’язків. Вони відображаються пунктирними лініями, направленими від слова до його тецедента. Приклад:

аналіз

тексту

і

його

результати

2

с

2

1

Врахування в КСС поверхових властивостей тексту забезпечується тим, що в ній, згідно з її визначенням, зберігаються, за винятком складових компонентів фразем (які при необхідності легко відновити по словниковим статтям самих цих фразем), всі одиниці і характеристики ГМП тексту (навіть до відомостей про порядок слів) – тобто, всі елементи тексту з точністю до явищ додаткової дистрибуції і вільного варьювання на морфологічному рівні. В КСС речення не опускаються прийменники, сполучники, частки, морфологічні характеристики, знаки пунктуації та інші одиниці, навіть якщо їх роль в даному реченні зводиться в основному до того, щоб слугувати показниками тих чи інших синтаксичних зв’язків. На відміну і від поверхових, і від глибинних синтаксичних структур моделі “Смисл⬄Текст”, в КСС такі одиниці не заміняються символами відповідних синтаксичних відносин, а лише доповнюються ними.

Що стосується “семантичності” КСС, то ця її властивість проявляється вже в тому, що семантично однозначному реченню співставляється в загальному випадку рівно одна КСС.

Послідовне включення в КСС всієї можливої поверхової інформації також сприяє її “семантичності”, відповідаючи вимозі максимально повного і дифіренційованого відображення в КСС значущих елементів тексту. Це пов’язано в першу чергу з тим, що мовні феномени, як правило, поліфункціональні і заздалегідь важко передбачити, яка з їх функцій буде головною в тому чи іншому тексті. Збереження відомостей про поверхневі вирази одиниць, що відносяться до смислової структури тексту, є надійною гарантією проти непередбачених втрат інформації у випадках, коли знайдені при аналізі поверхневі явища не тільки служать засобом оформлення виявлених глибинних одиниць, але й мають якийсь свій додатковий зміст.

Важливе значення в процесі власне перекладу має і можливість безпосереднього врахування порядку слів. В системі АРАП діє таке правило: при відсутності в КСС вхідного англійського речення додаткових показників комунікативної організації (наприклад, артиклі) для всіх словоформ, місце яких в лінійному представленні речення не визначається однозначно їх місцем в його синтаксичній структурі, порядкові номера цих словоформ, що вказані в співставлених їх вузлам синтаксичного дерева, в процесі заміни англійських одиниць їх перекладними еквівалентами зберігаються тими ж; при наявності в вхідній КСС артиклів та інших одиниць, в ролі перекладних еквівалентів яких виступають при перекладі на російську мову правила змін порядку слів, порядкові номера словоформ змінюються за відповідними правилами.

Проте найбільше узгодженість КСС з смисловою структурою тексту виявляється у характері використовуємих в ній синтаксичних зв’язків. Такі зв’язки являють собою глибинно-синтаксичні зв’язки. Сутність застосування саме глибинно-синтаксичних відношень, а не поверхневих полягає в тому, що ці зв’язки розглядаються як значимі елементи тексту, які задають смислові відношення між словоформами та іншими сегментними одиницями тексту. Основний принцип, якому повинна задовільняти КСС речення: всі синтаксичні зв’язки між словоформами, що відображаються в КСС повинні бути семантично обгрунтованими.

В рамках системи АРАП принцип семантичної обгрунтованості синтаксичних зв’язків конкретизується у вигляді такої вимоги: всі синтаксичні зв’язки, що встановлюються в КСС довільної реальної фрази, повинні реалізувати які-небудь семантико-синтаксичні валентності наявних у фразі лексичних або морфологічних одиниць. Самі ж ці валентності, в свою чергу повинні безпосередньо співвідноситися з тлумаченнями відповідних одиниць, і при їх виявленні повинна залучатися семантична інформація.

В системі АРАП залучається два важливих типа семантичної інформації: 1) семантичні описи, або тлумачення, значимих лексичних, морфологічинх і пунктуаційних одиниць; 2) відомості про семантичне сполучення цих одиниць одна з одною.

Тлумачення використовуються двух типів: словникові і контекстні. Словникове тлумачення може бути приписано будь-якій одиниці, що включена в словник, і задає ту частину її смислу, яка повністю або частково зберігається в різних випадках реалізації цієї одиниці в тексті. Контекстуальне тлумачення характеризує словоформу і являє собою той фрагмент семантичного запису тексту, який відповідає данній словоформі в даному конкретному її використання, визначаючи можливості здійснення з нею тих чи інших семантичних операцій. Контекстуальні тлумачення будуються з словникових одиниць, що утворюють ГМП цих словоформ, з урахуванням наявного лексико-синтаксичного контексту та можливих в цьому контексті семантичних явищ.

Відомості про сполучення в системі АРАП носять характер семантичних вимог. З їх допомогою можна досить повно охарактеризувати всі найбільш суттєві аспекти смислової структури тексту і задати всі основні семантичні операції над цим текстом, які можуть знадобитися в процесі перекладу.

Для формального запису використовуємих при цьому семантичних відомостей розроблена спеціальна символічна мова, яку можна реалізувати на ЕОМ. Вона заснована на використанні біля 100 “семантичних елементів” – символів, що позначають одиниці смислу. Семантичні елементи можуть з’єднуватися за певними правилами, утворюючи семантичні формули - семантичні дерева залежностей, які завдяки використанню формальних записів, зокрема, дужкового запису, можуть записуватися лінійно. На множині семантичних формул визначений ряд бінарних відношень, що моделюють смислові парадигматичні відношення в лексиці. В результаті в цих термінах можна записувати як словникові та контекстуальні тлумачення будь-яких одиниць, так і родо-видові лексико-семантичні класи, в які ці одиниці входять, а також семантичні інтерпретації їх синтаксичних валентностей і утворювані валентностні лексико-семантичні класи.

Ще однією важливою властивістю даного апарату є можливість спростити співвідношення між синтаксичним і семантичним представленням тексту у порівнянні з тим, як трактується це співвідношення в більшості розробок за моделлю “Смисл⬄Текст”. В зв’язку з тим, що структура семантичних формул зберігає основні властивості дерева залежностей, перехід від КСС тексту до його семантичного запису зводиться до простої заміни всіх слів і значимих морфологічних і пунктуаційних одиниць, наявних в КСС, їх контекстуальними тлумаченнями.

Прийнятий в системі АРАП підхід до опису синтаксису і семантики в їх взаємодії відповідає тим принципам, які закладені у формальну модель процесу перекладу, що є теоретичною базою цієї системи.

Алгоритм синтаксичного аналізу мовних текстів

Алгоритм, що буде тут розглядатися, являє собою частину більш складної системи, яка призначена для автоматичного синтаксичного аналізу текста. Під синтаксичним аналізом розуміється визначення смислових зв’язків між об’єктами тексту, тобто між словами в межах простих речень і між простими реченнями в складі складних.

Вся система в цілому складається, окрім алгоритму синтаксичного аналізу і деяких допоміжних алгоритмів (зокрема технічного характеру), з алгоритму морфологічного аналізу. Морфологічним аналізом називається обробка окремих словоформ, в результаті якої кожній словоформі ставиться у відповідність її інформація – характеристика, що відображає ті властивості словоформи, які необхідні для наступного синтаксичного аналізу. До початку синтаксичного аналізу увесь текст представляється у вигляді послідовності інформацій до словоформ, так що алгоритм синтаксичного аналізу має справу не з словоформами, а лише з відповідними інформаціями.

Для представлення результатів синтаксичного аналізу, тобто для відображення зв’язків між словами і між реченнями, використовується 31 відношення безпосередньої домінації (ВБД). Ці відношення бінарні; вони антирефлексивні, антисиметричні і антитранзитивні. Один з членів відношення БД вважається головним (господарем), інший – залежним (слугою). В простому реченні всі слова повинні бути залежними, крім присудка (який розглядається як “вершина” речення); в складному реченні повинні бути залежними всі речення, крім головного. Кожне слово повинно мати лише одного “господаря”, але може мати скільки завгодно “слуг”; кожне речення може мати двух “господарів” (обов’язково різного рівня: одного “господаря” – речення, а іншого “господаря” – слово в цьому реченні), а також скільки завгодно “слуг”. ВБД відображаються нумерованими стрілками, що направлені від “господаря” до “слуги”. Наприклад:

Побудувати

такий

алгоритм

ще

складніше

1

2

9

27

Можна сказати, що синтаксичний аналіз полягає в тому, щоб певним чином розставити в тексті нумеровані стрілки.

Смисл зв’язкам можна дати такий:

1. Предикативне (происходит совмещение, процесс окончен, естественно предположить, требовалось выяснить).
2. 1-е об’єктивне (решать задачу, хранение информации, принадлежащий к множеству, избегает ошибок).
3. 2-е об’єктивне (приписать букве индекс, сведение вычислений к сложениям).
4. 3-е об’єктивне (перевод текста с английского на русский язык, сравнить самолет с ракетой по скорости).
5. Означальне (обычная запись, вычислительный процесс, действовать машинально, чисто автоматически).
6. Вказівне (этот человек, тех решений).
7. Присвійне (орудие вычислителя, их комната).
8. Кванторне (все данные, любая буква, некоторые процессы).
9. Загальнокваліфікаторне (такие таблицы, такой подход).
10. Порядкове (первое слово, шестая строка, ХХ век, 1953 год).
11. Кількісне (пять страниц, трем отделам, из ста восьми шагов, много способов, сколько текстов).
12. Агентивне (решено машиной, исполнение романса певицей, вычисляется алгоритмом, наше отставание, его полет).
13. Субстантивно-атрибутивне (отверствие диаметром 6мм, столб высотой 2 м, такого рода утверждение).
14. Партитивне (блок устройства, елемент множества).
15. Загальногенетивне (лист бумаги, пример алгоритма, понятие алгебры, преимущество метода).
16. Обставинне (находиться позади барьера, квадрат слева, цифра записана на ленте, подается для реализации, однако выяснилось).
17. Суб’єктно-копредикативне (он вернулся усталым, умер стариком).
18. Об’єктно-копредикативне (нашли его усталым).
19. Аппозитивне (угол ABC, часть Г).
20. Порівняльне (сильнее первого мотора, более низкий, чем стержень С2).
21. Елективне (каждый из разрядов, две из колонок, многие из иероглифов).
22. 1-е призв’язочне (будет полным, этот четырехугольник есть квадрат, являющиеся достаточными).
23. 2-е призв’язочне (бывают трех типов, могут быть следующего рода).
24. 1-е допоміжне, або службове (более четкий, самые полные, не получил, видели бы).
25. 2-е допоміжне (сорок три, сремиться к нулю, физика и математика).
26. 3-е допоміжне, або відприйменникове (без двигателя, несмотря на отказ).
27. Обмежувальне (хотя бы один, только числа, лишь законченные работы).
28. Однорідне (операции и константы, целые или не целые числа, нумеруются, но не сдвигаются).
29. Квазіоднорідне (общая функциональная схема, обычного графического изображения).
30. Відсполучникове (как выход, как для решения).
31. Зіставлюване (чем дальше продвигаемся, тем яснее видим).

В процесі аналізу за допомогою певної послідовності операцій текст розбивається на певні частини, які ототожнюються з одиницями аналізу (в даній роботі це синтагми та інши змістовні сполучення); одиниці аналізу зібрані в особливому списку, де до них приписані вказівки, які необхідно виконати, щоб фіксувати (зображати) знайдені зв’язки (в даній роботі це вказівки про постановку нумерованих стрілок).

В тексті, що аналізується, розглядаються три типи відрізків: *гіпотетична словоформа* (вхідна) – послідовність літер між двума проміжками; знаки пунктуації також вважаються вхідними словоформами; *гіпотетичний сегмент* – послідовність словоформ між двума знаками пунктуації або сполучниками, перед якими немає коми; *гіпотетична фраза* – послідовність сегментів між двома крапками або іншими тотожніми знаками. В результаті роботи алгоритму ці гіпотетичні (вхідні) об’єкти перетворюються на кінцеві, або приведені: *приведена словоформа*, *приведений сегмент*, *приведена фраза*. Приведена словоформа – це загальна назва для всіх лексичних одиниць. Приведеними синтагмами називаються вирази з повною або частковою предикативністю, тобто прості речення, дієприслівникові і обособлювані означальні (зокрема прислівникові) звороти. Приведена фраза є аналогом самостійного речення – простого або складного, але не того, що є частиною іншого речення. Приведена фраза – це максимальний відрізок тексту, в межах якого враховуються синтаксичні зв’язки. Приклад:

**.**(1) **Н**(2)| **.**(3)**И**(4)| **.**(5)**Голубков**(6) **исследовал**(7) **еще**(8) **двадцать**(9) **шесть**(10) **видов**(11) ||| **,**(12) **так**(13) **как**(14) **его**(15) **первый**(16) ||| **и**(17) **второй**(18) **опыты**(19) **дали**(20) **результаты**(21) ||| **,**(22) **представленные**(23) **на**(24) **рис**(25)| **.**(26) **6**(27) ||| |.

Тут 27 гіпотетичних словоформ (пронумеровані), 4 гіпотетичних сегмента (відокремлені |||), 4 гіпотетичних фрази (відокремлені |), 17 приведених словоформ (підкреслені), 3 приведених сегмента (1 – до **так как**, 2 – до **представленные**, 3 – до кінця), 1 приведена фраза (все речення).

Необхідні відомості про текстові об’єкти представляються у вигляді інформацій. Інформація - послідовність ознак (граф), що приймають певне значення.

Більшість ознак, що утворюють інформацію, є синтаксичними розрізнюваними ознаками словоформ та сегментів.

Увесь процес аналізу зводиться послідовно на декількох рівнях. Спочатку з окремо взятих гіпотетичних словоформ, точніше з інформацій до морф, на які ці словоформи розбиваються, виводиться інформація до словоформ, потім з інформацій до словоформ виводиться інформація до гіпотетичних сегментів; і нарешті, з інформацій до гіпотетичних сегментів виводиться синтаксична структура фрази. Одночасно гіпотетичні об’єкти перероблюються на приведені.

Відповідно, аналіз підрозділяється на *формологічний* (розбір окремих словоформ і отримання інформацій до них) та *синтаксичний* (все інше). Синтаксичний аналіз, в свою чергу, поділяється на *внутрішньосегментний* (розбір окремих гіпотетичних сегментів і отримання інформацій до них) та *міжсегментний* (розбір всієї фрази в цілому).

В межах гіпотетичних сегментів багато зв’язків між словоформами не можуть бути встановлені зовсім або встановлюються неправильно, тому в інформації до сегменту застосовуються ϕ - ознаки (вказівки про невстановлені, але передбачувані зв’язки) та ρ - ознаки (вказівки про можливі помилки у встановлених зв’язках). Ці ознаки використовуються при міжсегментному аналізі.

Особливо треба відзначити λ - ознаки, які виробляються на всіх етапах аналізу і спочатку входять в інформації до сегментів, а звідти – в інформації до фраз. λ - ознаки – це вказівки про синтаксичні неоднозначності, які не можуть бути розв’язані навіть в межах фрази, без залучення загального смислу тексту.

Для пошуку одиниць синтаксичного аналізу – змістовних сполучень – всі вони оформлюються у вигляді шаблонних пошукових правил – конфігурацій. Конфігурації складаються з 5 частин: 1) містить порядковий номер конфігурації, в 2) записано перший і другий члени конфігурації, що задають члени змістовного сполучення, 3) відведена для допоміжної інформації, в 4) знаходиться основна інформація (назва стандартної операції, що виконує переробку інформацій та її параметри), 5) відведена для “переадресації”, тут записуються номери конфігурацій, до яких треба переходити у певних випадках.

Конфігурації зібрані в таблицю, в своєрідний “синтаксичний словник”. Окремо від таблиці існує “власне алгоритм” – інструкція по використанню словника. Власне алгоритм синтаксичного аналізу записується у вигляді набору стандартних алгоритмічних операторів.

Синтаксичний аналіз організований циклічно. При внутрішньосегментному аналізі кожний гіпотетичний сегмент оброблюється 5 раз. Для міжсегментного аналізу також пропонується 5 циклів.

“Власне алгоритм” та будова таблиці конфігурацій не залежать від конкретної мови.