­­МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ “ЛЬВІВСЬКА ПОЛІТЕХНІКА”

­­­

Кафедра САПР

**Особливості та основні засоби для комп’ютерного автоматичного перекладу**

Контрольна робота №1

з курсу: “Основи інформаційних технологій”

для студентів базових напрямків 6.0804 “Комп`ютерні науки”

Виконав студент гр. КНз-11

Чалий Михайло

­­

Львів 2012

Зміст

[Особливості та основні засоби для комп’ютерного автоматичного перекладу 3](#_Toc347225702)

[Машинний переклад заміною тексту 4](#_Toc347225703)

[Машинний переклад з семантичним розбором 5](#_Toc347225704)

[Статистичний машинний переклад 7](#_Toc347225705)

[Гібрідні техніки перекладу 9](#_Toc347225706)

[Основні засоби 10](#_Toc347225707)

[Майбутні техніки машинного перекладу 12](#_Toc347225708)

[Перелік літератури 14](#_Toc347225709)

[Адендум. Приклад перекладу за допомогою правил 15](#_Toc347225710)

Особливості та основні засоби для комп’ютерного автоматичного перекладу

Комп’ютерний автоматичний переклад (інша назва Машинний Переклад, чи скорочено МП) це комп’ютерна галузь яка опікується перекладом з однієї природної мови на іншу. Перекладу підлягають як письмові тексти так і усні. Найпростіші техніки перекладу базуються на заміні частин тексту відповідниками з іншої мови, нажаль результати таких перекладів дуже погані, адже мови мають відмінності не тільки в словах, а ще і в структурі речення, фраз, тощо. Кращі результати дають техніки, що базуються на семантичному розборі речення, це дає змогу отримувати пристойні результати з невеликими затратами ресурсів. Остання і мабуть найкраща на цей час техніка, це статистичний машинний переклад. Статистичний машинний переклад використовує великі об’єми мовних пар. Отже далі розглянемо кожну з цих технік більш детально.

# Машинний переклад заміною тексту

Машинний переклад заміною тексту з’явився як результат еволюції автоматизованого перекладу (автоматизований переклад - система тільки допомагає перекладати текст). Це найпростіша в реалізації техніка, адже маючи потрібний словник, програма просто замінює відповідні слова чи частини речення. Найбільшу популярність отримала на час раннього розвитку комп’ютерів, тоді головною проблемою для реалізацій ставала висока вартість пам’яті. При всіх недоліках, ця техніка має дуже важливу відмінність, вона дає найбільшу швидкість і найменше використання ресурсів. На цей час машинний переклад заміною тексту не використовують за прямим призначенням. Існує декілька сценаріїв використання, найпоширеніший це індексування відповідностей тексту, текстам на інших мовах. Так наприклад, найбільші пошукові системи як Google та Bing мають можливість шукати ігноруючи мову запиту та проіндексованого контенту. Таку саму можливість мають і індексатори для повнотекстового пошуку для Microsoft SQL Server, Apache Lucene та інші.

# Машинний переклад з семантичним розбором

Головною проблемою перекладів заміною є той факт, що всі мови мають відмінності в побудові речень. Це означало що переклад заміною міг задовольнити принаймні розуміння сенсу, але був незрівняний з перекладом людини. Наступним кроком еволюції машинного перекладу став переклад з семантичним розбором. Спочатку ця техніка використовувала просто набори правил (Rule - based Machine Translation), які покривали найпоширеніші трансформації один в один використовуючи словникову інформацію і аналіз граматичних правил конкретних мов. Такі системи будуються на основі лінгвістичного опису двох природніх мов (двомовних словників або інших баз даних, що містять морфологічну, граматичну та семантичну інформацію), формальних граматик і власне алгоритмів перекладу. Якість перекладу залежить від обсягів лінгвістичних баз даних (словників) і глибини опису природних мов, тобто, необхідний облік максимальної кількості особливостей граматичної структури як вхідної, так і вихідної мови. На заваді формалізованим правилам, стала як не дивно неформальність мови. Тоді як майже завжди є формальна частина тексту, для побудови перекладача природньої мови, не формальний аспект ставав дуже важливим. Приблизно в той самий час, поширення почали набувати алгоритми штучного інтелекту, це а також тогочасні розробки в лінгвістиці, дали змогу побудувати реалізації семантичного розбору який повністю покривав і формалізовану частину мови, і що найголовніше не формалізовану.

Систем на основі формалізованих правил ділять на дві окремих гілки. Системи типу Transfer - припускають морфологічний, синтаксичний і семантичний аналіз тексту на мові входу; перетворення в структуру вихідної мови (TRANSFER) і як результат синтез тексту вихідною мовою. І системи Interlingua - припускають аналіз вхідного тексту в термінах метамови і синтез метаструктур тексту вихідною мовою.

Порівняно з іншими техніками, техніки на базі правил мають такі перваги:

* Синтаксична і морфологічна точність,
* Стабільність і передбачуваність результату,
* Можливість налаштування на предметну область.

Головні недоліки:

* Трудомісткість і тривалість розробки,
* Необхідність підтримувати і актуалізувати лінгвістичні бази даних.

Майже всі системи машинного перекладу, що брали початок як десктопні застосунки, використовують переклад на базі правил. Так найвідоміші поточні реалізації мають Systran, Linguatec і платформи GramTrans та Apertium.

# Статистичний машинний переклад

З розвитком алгоритмів навчання, а також зі значним здешевленням комп’ютерних ресурсів, з’явилась можливість реалізації іншої техніки. Теорія статистичного перекладу сягає 17го сторіччя, вже тоді була подана ідея можливості перекладу тексту за допомогою великої кількості прикладів. Ця технологія заснована на пошуку найбільш ймовірного перекладу речення з використанням даних, отриманих з двомовної сукупності текстів. Такі системи перекладу будуються на основі порівняння дуже великих обсягів «паралельних» текстів, текстів, які містять речення на одній мові і відповідні їм речення на іншому. Такими текстами, наприклад, володіють веб пошукові системи на кшталт Google чи Bing. Інший приклад двомовної сукупності текстів це парламентські звіти, які являють собою протоколи дебатів в парламенті. Двомовні парламентські звіти видаються в Канаді, Гонконгу та інших країнах; офіційні документи Європейського економічного співтовариства видаються на 11 мовах; а Організація Об'єднаних Націй публікує документи на декількох мовах. Як виявилося, ці матеріали є безцінними ресурсами для статистичного машинного перекладу.

Цікаво що статистичний машинний переклад має властивість «самонавчання», адже чим більше в розпорядженні є паралельних текстів і чим точніше вони відповідають один одному, тим краще результат статистичного машинного перекладу.

Підрозділом статистичного перекладу також є краудсорсінговий переклад, він завжди використовується як доповнення інших технік, і дає можливість навчати систему правильним перекладам великій групі людей, в краудсорсінгових системах перекладу використовується неявне голосування за варіанти. Дуже показовий представник статистичного перекладу з використанням краудсорсінгу є Google Translate.

Переваги систем статистичного перекладу:

* Природність перекладу,
* Відносна легкість в реалізації при достатній кількості «паралельних» текстів
* Незалежність технології від мовних пар

Недоліки:

* Обмеженість «паралельних» текстів в природі,
* Невміння справлятися з морфологією і синтаксисом,
* Спотворення інформації (дублювання, пропуск, підміна інформації).

Традиційно компанії що володіють або мають доступ до великих об’ємів данних і розробляють такі системи: Google, SDL Language Weaver, Microsoft, Asia Online, IBM.

# Гібридні техніки перекладу

Описані техніки перекладу мають свої недоліки і складності, а також досягли певної межі у своєму розвитку, розробники рішень технік машинного перекладу розраховують на технологічний прорив за рахунок створення гібридних технологій. Наприклад об’єднання технік перекладу за правилами та статистичного перекладу. Такий підхід дозволяє взяти сильні сторони обох технологій (граматичну точність при перекладі по правилах і природність статистичного).

# Основні засоби

Технічні проблеми які доводиться вирішувати розробникам систем машинного перекладу можна поділити на такі:

1. Великі, або дуже великі об’єми даних необхідних наприклад для статистичного перекладу
2. Ефективність алгоритмів обробки даних та перекладу
3. Чистота вхідних даних

Об’єми текстів які дали можливість існувати системам на основі статистичного перекладу та навчанню штучного інтелекту вимірюються десятками і сотнями терабайтів, швидкість доступу до даних також відіграє важливу роль. Дуже часто для будування таких систем використовується NoSQL рішення, найпоширеніші представники яких э MongoDb та CouchDb. Дані в таких базах даних зберігаються у вигляді структурованих документів (наприклад JSON або BSON), це дає змогу зберігати тексти у вигляді дерев з дуже швидкими доступом як до документів, так і до окремих частин тексту. Для швидкого доступу, часто використовуються індекси або карти ідентифікаторів. Одним з дуже важливих факторів зберігання великих об’ємів даних є реплікація і шардінг. І реплікація і шардінг дозволяють зберігати данні на великій кількості фізичних машин, а от працювати з ним як з єдиним цілим. Чудовим прикладом такого використання баз даних є Google, наприклад для зберігання та обробки інформації Google Translate використовується більше 50 тис комп’ютерів.

Швидкість та якість перекладу, залежить не тільки від об’ємів даних, не останню роль тут мають алгоритми обробки. Один з алгоритмів який дозволяє обробляти великі об’єми даних і робити агрегацію, є Map-Reduce. Спрощенно дозволяє декомпозувати великий об’єм роботи на маленькі і швидкі компоненти, а це в свою чергу дозволяє виконувати роботу на великій кількості машин, а на прикінці зібрати результати в одну відповідь. Як з будь яким алгоритмом Map-Reduce має багато окремих реалізацій, хотілося б виділити дві Apache Hadoop та Microsoft Dryad. Перша є класичною реализацією Map-Reduce, у той час як друга це використання SQL подібного синтаксису для Map-Reduce.

Навіть маючи добре сховище даних, маючи чудові алгоритми, цього буде не достатньо для побудови сучасної системи перекладу, всі техніки використовують початкові данні, так для статистичного перекладу початкові данні потрібні завжди, а для систем базованих на правилах, потрібні данні для навчання штучного інтелекту. Тут на сцену виходять декілька окремих процесів. Найголовніший це ETL (Extract Transform Load) процес який готує дані для використання системою. На етапі Extract дані витягаються з джерела, наприклад з HTML сайтів, або стрічок новин, або як сказано вище з записів дебатів в парламенті, на цьому етапі проходить первинна обробка, наприклад для HTML документів прибираються внутрішня розмітка, картинки, тощо. На етапі Transform данні конвертуються в потрібний для нас формат. Також на цьому етапі може додаватися метаінформація. На етапі Load дані завантажуються в систему перекладу чи систему навчання.

# Майбутні техніки машинного перекладу

Всі описані в цій роботі техніки машинного перекладу використовують текст як в якості вхідних даних так і для результатів. Наразі це далеко не єдині формати. Наприклад дуже поширеним є використання систем що поєднують розпізнавання тексту (OCR) з графічних даних і перекладач. Такі системи можуть опрацьовувати інформацію в реальному режимі часу, і вже існують системи augmented reality, які дозволяють бачити переклад тексту, який просто знімається відеокамерою мобільного телефону.

Інший чудовий приклад це синтезування голосу, в цьому випадку після процесу перекладу, текст озвучується (наприклад Microsoft Sound Framework). Існують також техніки розпізнавання голосу, але на поточний момент вони ще потребують великих комп’ютерних потужностей, а тому не є дуже поширеним. Для роботи з голосом сьогодні використовуються гібридні методи. Так наприклад дуже поширена схема коли голос конвертується в хвильову форму, передається на сервера, там за допомогою статистичних технік знаходяться відповідники, результат в хвильовій формі передається на мобільний пристрій і там вже синтезується. Така схема працює в будь якому сучасному телефоні (Apple Siri, Google Voice Search, Bing Search), але все більша і більша частина обробки голосу переноситься на пристрій.

Наступним еволюційним кроком, є переклад аудіо\відео потоків. Так Microsoft Research, на конференції BUILD показали прототип системи, в якій відеорозмова між двома абонентами велась на різних мовах, але кожен співрозмовник чув свою мову. За рахунок того що використовувались статистичні алгоритми, систему вдалось зробити нечуттєвою до конкретної мови (наприклад демонстрація була зроблена на прикладі перекладу з Мандаринської мови на Французьку і навпаки).

# Перелік літератури

* [Hutchins, W. John](http://en.wikipedia.org/wiki/John_Hutchins); Somers, Harold L. (1992). [*An Introduction to Machine Translation*](http://www.hutchinsweb.me.uk/IntroMT-TOC.htm). London: Academic Press. [ISBN](http://en.wikipedia.org/wiki/International_Standard_Book_Number) [0-12-362830-X](http://en.wikipedia.org/wiki/Special:BookSources/0-12-362830-X)
* Wikipedia - Rule-based machine translation
* Wikipedia - Statistical machine translation
* Wolfram Alpha – Статистичні методи обробки даних
* MongoDb Reference Documentation – Introducing document oriented database
* MapReduce: Simplified Data Processing on Large Clusters (Google, OSDI'04: Sixth Symposium on Operating System Design and Implementation)

# Адендум. Приклад перекладу за допомогою правил

Приклад перекладу з Англійської на Німецьку:

*A girl eats an apple.* Source Language = English; Demanded Target Language = German

Для перекладу потрібно:

* Англо-німецький словник для кожного слова
* Правила що описують Англійські речення
* Правила що описують Німецькі речення

І правила для трансформації зі структури Англійської мови в Німецьку.

Зазвичай процес перекладу виглядає так:

**№1:** getting basic part-of-speech information of each source word:

a = indef.article; girl = noun; eats = verb; an = indef.article; apple = noun

**№2:** getting syntactic information about the verb “to eat”:

NP-eat-NP; here: eat – Present Simple, 3rd Person Singular, Active Voice

**№3:** parsing the source sentence:

(NP einen Apfel) = the object of eat

Часто достатньо тільки часткового розбору тексту

**№4:** translate English words into German

a (category = indef.article) => ein (category = indef.article)

girl (category = noun) => Mädchen (category = noun)

eat (category = verb) => essen (category = verb)

an (category = indef. article) => ein (category = indef.article)

apple (category = noun) => Apfel (category = noun)

**№5:** Mapping dictionary entries into appropriate inflected forms (final **generation**):

A girl eats an apple. => Ein Mädchen isst einen Apfel.