Створення багатомовної самонавчальної системи автоматизованого перекладу

Валентин Соломко

Інформаційна дирекція АТЗТ "САММІТ"

pere@slovnyk.org.ua

Abstract

The creation possibility of low-budget self-training machine translator accessible for ordinary users overall was investigated.

Анотація

Розглядається можливість створення малобюджетної самонавчальної системи автоматизованого перекладу доступної для масового пересічного користувача.

Вступ

Розвиток інформаційних технологій збільшив інформаційний обмін між окремими людьми та їх об'єднаннями. Зараз існує нагальна потреба у доступних пересічному користувачеві програмах-перекладачах, які б максимально відповідали поточному стану мов. Тобто йдеться про масову доступність і актуальність. Досить гостро проблема стоїть щодо перекладу українською мовою. Тут спродуковано три системи: "Плай$^{(tm)}$" [1] , "L-Master" [2] і "ПАРС/У" [3] . Вони забезпечують переклади українсько-російського та українсько-англійського напрямків. Інші напрями – "відкриті".

Опис проблеми

Існуючі системи машинного перекладу - це все комерційні продукти, і для їх створення та вдосконалення потрібно залучати відповідних висококваліфікованих мовників-фахівців (далі експерти). Іншими словами існуючі перекладачки є, за способом створення та відтворення, експерто-орієнтованими. Наразі для забезпечення перекладу в кожному новому напрямку потрібні досить серйозні кошти, які можна залучити тільки за наявності комерційної доцільності. Крім того існує проблема виявлення експертних знань, яка полягає у тому, що експерт змушений виконувати роботу, протилежну до своєї звичної (найчастіше: замість аналізу даних та закономірностей – їхній синтез), що створює серйозні методологічні й фінансові труднощі. Методологічні полягають у тому, що хтось повинен видобувати з експерта знання предметної області – робота для когнітолога; фінансові в тому, що потрібно оплачувати цю роботу.

При цьому експерт, як мінімум, повинен:

- знати внутрішні структури мов, між якими здійснюється переклад;
- мати ясне уявлення про культуру, історію, мораль, переважні типи мислення народів, що є носіями мови;
- володіти по можливості більшим словниковим запасом, більш-менш структурованим за областями застосування слів (спеціальна термінологія, діалекти, ідіоматика, сленг);
- мати явний чи інтуїтивний тезаурус слів обох мов, тобто по даному слову вміти запропонувати семантичні функції від нього, такі як синонім, антонім, конверсив, класичний атрибут (докладніше див. [4]), а також уміти запропонувати похідні частини мови від даного слова, якщо такі існують (добро – добрий – добріше – подобрів і т.п.).

Суттєве здешевлення розробки можливе за рахунок експертів (а, значить, і когнітологів). Особливо на початкових, найбільш коштоємних та ризикових етапах. Можливим розв'язком була б система, яка може самостійно створювати й удосконалювати мовну пару на основі існуючих двомовних текстів. Це особливо важливо щодо повсякденного неспеціалізованого спілкування. В ролі сирцевого матеріалу придатні, зокрема, звичайні газетно-журнальні публікації. Перспективність створення мінімальних перекладових баз підкріплюється успішним існуванням та використанням "базової англійської"(Basic English [5]). Результати роботи такої системи перекладу мають чернетково "закрити" відповідний мовний

напрямок і стати основою для створення експертно-орієнтованих систем перекладу. Крім того за допомогою таких програм вирішується завдання відстежування мовних змін, тобто проблема актуалізації наявних мовних пар.

Засадами для створення такої системи машинного перекладу стають статистичні закономірності розподілу частоти, довжини та структури елементів тексту (закон Ципфа) і деякі припущення, які випливають з цих закономірностей.

Теоретичні засади

Наприкінці 40-х років нашого століття Дж. Ципф [6], зібравши величезний статистичний матеріал, спробував показати, що розподіл слів природної мови підпорядковується одному простому закону, який можна сформулювати так. Якщо для якогось досить великого тексту скласти список всіх, які зустрілися в ньому слів, потім розмістити ці слова у порядку спаду частоти їх появи у тексті та пронумерувати в порядку від 1 (порядковий номер найчастішого слова) до *R*, то для будь-якого слова добуток його порядкового номера (рангу) в такому списку та частоти його появи у тексті буде величиною постійною, яке має приблизно однакове значення для будь-якого слова з цього списку. Аналітично закон Ципфа може бути виражений у вигляді

$$f \cdot r = c, \qquad (1)$$

де f – частота появи слова у тексті;
r – ранг (порядковий номер) слова у списку;
c – емпірична постійна величина.

Отримана залежність графічно виражається гіперболою. Дослідивши таким чином найрізноманітніші тексти та мови, у тому числі давні мови, Дж. Ципф для кожної з них побудував зазначені залежності, при цьому криві мали однакову форму – форму "гіперболічної драбини", тобто, при заміні одного тексту іншим загальний характер розподілу не змінювався.

Закон Ципфа було відкрито експериментально. Пізніше Б. Мандельброт запропонував його теоретичне обґрунтування. Він гадав, що можна порівнювати письмову мову з кодуванням, причому всі знаки повинні мати певну "вартість". Виходячи з вимог мінімальної вартості повідомлень, Б. Мандельброт математичним шляхом прийшов до аналогічної закону Ципфа залежності

$$f \cdot r^{\gamma} = c\,, \qquad (2)$$

де γ – величина (близька до одиниці), яка може змінюватися залежно від властивостей тексту.

Дж. Ципфом та іншими дослідниками встановлено, що такому розподілу підпорядковуються не лише всі природні мови світу, й інші явища соціального та біологічного характеру: розподілу вчених за кількістю опублікованих ними статей (А. Лотка, 1926 р.), міст США чисельністю населення (Дж. Ципф, 1949 р.), населення за розмірами доходу (У. Парето, 1897 р.), біологічних родів за чисельністю видів (Дж. Уїлліс, 1922 р.) та ін.

Впорядкованість задається ранжуванням (порядком розміщення) найменувань елементів за частотою їх появи в порядку її зменшення. Така впорядкована сукупність найменувань елементів називається ранговим розподілом. Розподіли, які свого часу вивчав Ципф, – це типові приклади рангових розподілів. Виявилося, що вигляд рангового розподілу, його будова характеризує ту сукупність документів, до якої відноситься цей ранговий розподіл. З'ясувалося, що при побудові рангові розподіли в більшості випадків мають форму закономірності Ципфа з поправкою Мандельброта (2).

При цьому коефіцієнт γ – величина змінна. Незмінність коефіцієнта γ зберігається тільки на середній ділянці графіка розподілу. Ця ділянка набуває форми прямої, якщо графік наведеної закономірності побудувати в логарифмічних координатах. Ділянка розподілу з $\gamma = const$ називається центральною зоною рангового розподілу (значення аргументу на зазначеній ділянці змінюється від Inr_1, до Inr_2). Значенням аргументу від 0 до Inr_1 відповідає зона ядра рангового розподілу, а значенням аргумент а від Inr_2 до Inr_3 – так звана зона утинання.

Який же сенс в існування трьох явно розрізнюваних зон рангових розподілів? Якщо останнє належить до термінів, які становлять якусь область знання, то ядерна зона, чи зона ядра рангового розподілу, містить найбільш загальновживані, загальнонаукові терміни. Центральна зона містить терміни, найхарактерніші для галузі знань, котрі у сукупності відображають її специфічність, відміну від інших наук, "охоплюють її основний

зміст". У зоні ж утинання зосереджені терміни, які порівняно рідко використовуються у даній області знань.

Таким чином, основа лексики якоїсь області знань зосереджена у центральній зоні рангового розподілу. За допомогою термінів ядерної зони ця область знань "стикується із більш загальними областями знань", а зона утинання відіграє роль авангарду, який начебто "намацує" зв'язки з іншими галузями науки. Так, якщо кілька років тому в ранговому розподілі термінів тематичної області "Обробка металів" зустрівся б термін "лазери", то через його низьку очікуваність він, напевно, потрапив би саме в зону утинання: зв'язки між лазерною технікою та обробкою металів ще тільки "намацувалися". Проте сьогодні цей термін, без сумніву, потрапив би в центральну зону, що відобразило б уже його досить високу зустрічальність і, отже, стійкий зв'язок лазерної техніки з обробкою металів.

Графік рангового розподілу наповнений глибоким змістом: адже за відносною величиною тієї чи іншої зони на графіку можна судити про характеристики всієї області знань. Графік з великою ядерної зоною й малою зоною утинання належить до досить широкої, скоріш за все консервативної області знань. Для динамічних галузей науки характерна збільшена зона утинання. Мала величина ядерної зони може говорити про оригінальність області знань, до якої відноситься побудований ранговий розподіл, і т.д. Так, на підставі аналізу рангового розподілу виявилося можливим дати якісні оцінки документальним інформаційним потокам відповідно до тих галузей науки, де вони формувалися.

Тобто основні висновки із закону Ципфа полягають у тому, що

Розподілу Ципфа підпорядковуються будь-які внутрішньо структуровані масиви даних. Осмислені тексти підпадають під дію цього закону.

У структурі цих масивів відображені процеси концентрації та розсіяння, що, в принципі дає можливість вичленувати структурні зв'язки у масиві.

Припущення:

1. В існуючих загальнолексичних текстах присутні практично всі слова та словосполучення (далі - елементи), потрібні для мінімального базового перекладу.

2. Згідно до закону Ципфа імовірність появи відповідних пар елементів різномовних текстів суттєво вища (особливо для найуживаніших елементів) за імовірність випадкових збігів.

3. Як наслідок припущення 2., найуживаніші простіші пари обрамлятимуть пари більш складних утворень (словосполучень та ін.).

4. Виходячи з закону Ципфа слід передбачити два протилежноспрямовані процеси концентрації та розсіювання.

5. Як наслідок припущення 4., навчання має відбуватися не за один прохід. Тобто слід передбачити ітеративний, періодичний процес, який згасав би, за умови замкненого середовища.

Реалізація

Власне система складається з трьох основних елементів: розбирач тексту, врядувач бази перекладів, який, заодно, виконує функцію перекладу та пошуковик для найчастіших і найкоротших пар у різномовних текстах, що сполучений з врядувачем зворотним зв'язком. Розбирач готує текст для подальшої обробки, розбиваючи його на елементарні відрізки. Пошуковик та врядувач забезпечують два протилежноспрямованих процеси у взаємодії яких “вилущуються” пари перекладних відповідників. Пошуковик здійснює, за певними критеріями, відбір найкращих та подовження, за необхідності, контекстного оточення, а врядувальник - видалення, за певними критеріями, найгірших та розламування механічних поєднань.

Результати

У результаті експериментів з'ясувалося, що елементарним відрізком тексту може бути як одна із літер, так і послідовність літер. Оптимальним, з точки зору часу обробки, є розбиття: пробіл, послідовність літер, послідовність інших знаків. На початку словник може бути й порожнім але краще, для скорочення часу навчання, наповнити його найуживанішими перекладами.

Найменший обсяг опрацьованого тексту – 500кб для кожної мови (приблизно 1-2 номери одного з тижневиків). Повна обробка такого масиву займає близько доби самостійної роботи

середньопотужного неспеціалізованого (побутового, офісного) комп'ютера: обробник – 850 МГц, обсяг пам'яті – 256 Мб, ОС Лінукс (АльтЛинукс Мастер v 2.2), мова програмування – Perl (v 5.8). Результатом роботи є діючий словник приблизно з 10 тис. перекладних пар слів і словосполучень, що цілком відповідає базовому мінімуму. Звичайно, що для реального використання цього буде замало.

Інтернет надає можливість для отримання необхідних обсягів двомовних українсько-російських текстів, достатніх для створення повноцінної перекладної бази.

Навчання відбувалося на двомовних текстах тижневика “Дзеркало тижня” та “Зеркало недели” (14743 статті).

В результаті обробки текстів отримано перекладні бази:

- Українсько-російська – 1190074 пари перекладів
- Російсько-українська – 1236479 пар перекладів

Якість перекладу за результатами ручної перевірки експертом 1000 перекладених речень складає:

- 61% – повністю перекладених
- 36% – частково перекладених (не всі слова, не у тому відмінку, тощо)
- 3% – не перекладених (результат перекладу не дає можливості зрозуміти зміст оригінала або спотворює його)

Зараз продовжується робота із поповнення існуючих українсько-англійських та російсько-англійських перекладних баз до рівня достатнього для практичного використання.

Після підготовки необхідного рушія інтернет-версія системи автоматичного перекладу була встановлена на майданчику http://pere.slovnyk.org там само (а також на http://pere.sourceforge.net) розміщені перекладні бази для різних пар мов та програмна реалізація системи прекладу.

Висновки

Експериментально доведено можливість створення малобюджетної самонавчальної перекладачки доступної для масового пересічного користувача. Планується проведення експериментів з парами мов, які є значно менш спорідненими або неспорідненими взагалі.

Джерела

Сирцеві тексти віднайдено та отримано за допомогою інтернету:

- Локалізація ОС Лінукс: http://www.gnome.org/i18n, http://i18n.kde.org, http://www.iro.umontreal.ca/contrib/po/trans – українською, англійською та іншими мовами;
- Тижневик “Дзеркало тижня”: http://www.zn.kiev.ua – українською, http://www.zerkalo-nedeli.com – російською та http://www.mirror-weekly.com – англійською;
- Тижневик “Галицькі контракти”: http://www.kontrakty.com.ua/ukr – українською, http://www.kontrakty.com.ua/rus – російською.

Література та майданчики

1. http://www.mtsoft.kiev.ua/plaj.htm
2. http://www.trident.com.ua
3. http://www.ling98.com/
4. Жолковский А. К., Мельчук И. А. О семантическом синтезе //Проблемы кибернетики. – Вып. 19. – 1967.– с. 177–238.
5. Basic English: International Second Language, Ogden, C. K., 1968.
   Authorized by the Orthological Institute and prepared by E.C. Graham.
   New York : Harcourt, Brace & World, 1968. - xii, 525p
6. Яблонский А.И. Математические модели в исследовании науки. Ответственный редактор Ю.Н. Гаврилец., М. Наука 1986г. 352 с.