УДК 004.93

Козоріз Н.В., студент
Кабан В.О., студент
Бабак А.П., студент
Войцех В.О., студент

Гібридна система машинного перекладу технічних текстів на флективні мови

Київський національний університет імені Тараса Шевченка, 03680, м. Київ, пр-т. Глушкова 4д,
e-mail: kozoriz.nataliia@gmail.com
e-mail: vitalkaban@yahoo.com
e-mail: babak.anton@yandex.ua
e-mail: vlad.voitsekh@yandex.ua

Kozoriz N.V., student
Kaban V.O., student
Babak A.P., student
Voitsekh V.O., student

Hybrid system of machine translation of technical texts in flexional languages

Taras Shevchenko National University of Kyiv, 03680, Kyiv, Glushkova av., 4d,
e-mail: kozoriz.nataliia@gmail.com
e-mail: vitalkaban@yahoo.com
e-mail: babak.anton@yandex.ua
e-mail: vlad.voitsekh@yandex.ua

Описано структуру перекладача гібридного типу. Наведені етапи перекладу. Описано результат роботи парсера. Вказано будову правил та порядок їх застосування. Описано структуру словника. Наведено приклад обробки речення.

Ключові слова: перекладач, парсер, правила, словник.

A general structure of a hybrid-type translator is described. Stages of translation are listed. A parser builds a dependency tree for each sentence and binds each word with a tag indicating its part of speech and some grammar attributes. A structure of rules is described. In the left part of a rule, there is assigned a sub-tree, which it should apply to. In the right part, there is described the translation of the sentence's part that corresponds to the sub-tree. Words are listed in the order they should appear in the Russian sentence; part of speech of each sentence should be indicated, as well as grammar attributes, which a word has in the sentence structure specified. Rules are applied to the tree by decreasing number of nodes it encompasses. After the tree is folded, there will be left only one node containing all information about the structure of the sentence translated. Each word is transformed into its canonical form and sent to the dictionary, which finds a correct Russian equivalence based on the grammar attributes specified. Dictionary is a database containing words, their parts of speech and English translation, as well as tables corresponding to inflexional word groups. An example of sentence processing is provided.

Key words: translator, parser, rules, dictionary.

Статтю представив д.ф.-м.н., проф. Анісімов А.В.

Вступ

На сьогоднішній день основними підходами до машинного перекладу є такі: на основі правил (RBMT), статистичний (SMT) та гібридний, який поєднує в собі особливості двох попередніх. Жоден з підходів не є універсальним. Вибір способу машинного перекладу залежить від таких факторів, як тип тексту, мовна пара, а іноді й напрямок перекладу в мовній парі, наявність двомовних текстів для навчання системи тощо.

У статті [1]-[2] показано, що з перекладом на російську мову краще справляються системи на основі правил або гібридні системи завдяки своїй здатності пристосовуватися до специфічної морфологічної структури цієї мови.

Ми представимо розробку машинного перекладача, яка здійснюється на кафедрі математичної інформатики факультету кібернетики КНУ ім. Т. Шевченка. Систему призначено для перекладу технічних текстів. Вона має модуль перекладу на основі правил, а оскільки наявна достатня кількість двомовних текстів відповідної тематики для навчання статистичної складової, то було обрано гібридний тип перекладача.

На сьогодні не існує безкоштовних гібридних систем для російської чи української мов. Такі перекладачі, як Google Translate, є суто статистичними, що не дає достатньо якісного результату в разі перекладу на флективні мови, наприклад слов'янські.

Структура системи

Перекладач, що розробляється, має гібридну природу: текст перекладається за допомогою правил, а потім коректується на основі статистичних даних. Переклад здійснюється з англійської мови на російську.

Етапи перекладу:

- поділ тексту на речення;
- побудова дерев граматичних структур речень;
- знаходження канонічних форм слів;
- обробка дерев на основі правил, знаходження граматичних ознак слів та правильної послідовності слів;
- переклад слів на основі їх граматичних ознак за допомогою словника;
- корекція і виправлення помилок.

Перекладач зчитує текст і розбиває на речення. На наступних етапах кожне речення обробляється окремо.

Для знаходження граматичної структури англійського дерева та побудови дерева використовується Стенфордський парсер [3]. Кожному слово та розділовому знаку парсер ставить у відповідність його порядковий номер у реченні, нумерація починається з одиниці.

На основі даних, отриманих з парсера,

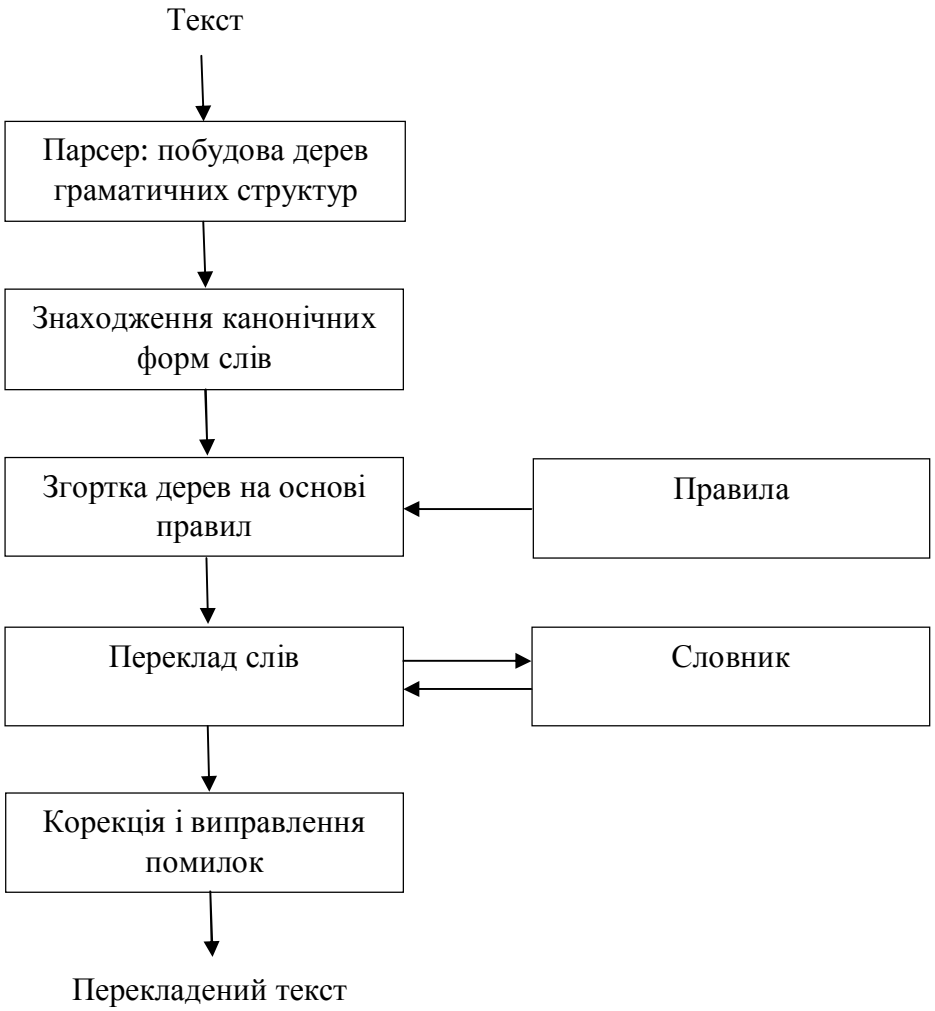

Схема 1. Структура перекладача

програма будує дерево залежностей. Кожна вершина дерева містить слово і його граматику – інформацію з тегу. Сполучаються вершини зв'язками залежностей. Коренем дерева є вершина ROOT, яка має порядковий номер 0 і вказує на головне слово речення.

Далі програма поступово об'єднує піддерева в одну вершину за допомогою правил. При цьому граматика слів доповнюється новими ознаками. Під час цього процесу вершина може містити не одне слово, а частину речення, в якій виділяється головне слово. Після закінчення об'єднання, залишається одна вершина, яка містить всі слова у тій послідовності, у якій вони мають бути перекладені, а також граматика цих слів.

Програма відправляє англійське слово з його граматичними ознаками словнику, який знаходить російське слово, що відповідає йому. Перекладене слово ставиться у потрібну форму на основі отриманої граматики і повертається програмі.

Отриманий переклад речення редагується за допомогою статистичних даних, що зберігається у базі даних програми.

Парсер

Стенфордський Парсер [3] працює на основі статистичних даних.

Парсер присвоює кожному слову тег, який визначає частину мови та граматичні ознаки цього слова, такі як множина, ступінь порівняння - тобто ті, які змінюють форму англійського слова.

Стенфордське представлення зв'язків залежності [4] забезпечує простий опис граматичних зв'язків у реченні. А саме: замість представлення фразової структури, воно зображує усі зв'язки речення як зв'язки залежностей. Всі зв'язки є бінарними, мають головне слово і залежне. Кожне слово, крім ROOT, має лише одне головне слово.

Максимальна кількість випадків, коли парсер неправильно визначає структуру речення, повинні виправляться при обробці дерева парсером.

Правила

Синтаксис правил

<rule>::=<left side>#<right side>

<left side>::=<PN>[.<w>](<Lnk>.<subtree>{,<Lnk>.<subtree>})

<subtree>::=<PN>[.<w>]|<left side>

<PN>::=<Pos><N>

Pos – part of speech, N – число, Lnk – dependency, w - word

<left side> правила репрезентує деяке піддерево dependency tree, де <PN> відповідає одному вузлу, а <Lnk> – дузі, що з нього виходить. У лівій частині правила можуть також зазначатися самі англійські слова. Усі Pos у лівій частині нумеруються. Піддерева перелічуються в лівій частині зліва направо. Правило може охоплювати не все піддерево, а довільну його частину.

<right side>::=<right rule>{$<right rule>}[(<new root Pos>)]

<right rule>::=<right part>{,<right part>}

<right part>::=<TPN>[.<Grammar>{&<Grammar>}]|<w>

<TPN>::=<Pos><N>

<Grammar>::=<gr attribute>:<value>

<gr attribute>::=Рід|Число|Відмінок|Час|Особа

<value>::=жін|чол|сер|наз|род|дав|…|одн|мн|…

new root Pos – це Pos вузла, яким замінюється гілка дерева після її обробки (якщо цей параметр не вказано, то та Pos, яка була в 1-му вузлі гілки, і залишається).

Права частина складається з одного або кількох правил. Переклад компонується згідно з одним із цих правил шляхом послідовної підстановки конкретних слів/словосполучень (w) або взятих зі словника TPN, які є перекладами PN з відповідними номерами N. Оскільки одному сорсовому слову можуть відповідати різні таргетові частини мови, то яку саме з них вибирати визначається параметром Pos у правій частині. У правилах правої частини повинні використовуватися ті самі номери Pos, що й у лівій частині, але порядок зазначення цих номерів та самі Pos можуть бути іншими. Можливо, деякі номери з лівої частини у правилі правої частини будуть опущені (але нових номерів, тобто тих, яких немає зліва, справа виникати не повинно).

Граматична форма таргетової лексеми визначається списком атрибутів Grammar (рід, число, відмінок тощо), значенням кожного з яких є граматична категорія (жін, чол, одн, мн і т.д.).

Якщо певна граматична ознака одного слова впливає на таку ж ознаку іншого слова (наприклад, рід іменника впливає на рід прикметника), але вона береться саме з перекладеного слова, тоді у правилі ми пишемо велику англійську літеру замість цієї ознаки у головному слові, і маленьку літеру – у

залежному слові (наприклад: С1.чис:Х,Г2.чис:х).

У лівих частинах правил Pos можуть замінятися на більш загальні. Для іменників NN – є найбільш загальним, NNS, NNP та NNPS – більш специфічні. Форма дієслова VBZ узагальнюється до VBP, яка в свою чергу разом з VBD, VBG, VBN може бути замінена на VB. Якщо для даних Pos не знайдено правила, шукається правило для більш загальних Pos.

До тегів, які розрізняє Стенфордський парсер додано VBC, яке позначає дієслово умовного способу.

За замовчуванням дієслова мають наказовий спосіб.

Якщо у граматику певного слова було занесено різні значення одного й того ж атрибуту, обирається останнє входження, тобто останній варіант покриває всі попередні.

Правила будуть застосовуватись у порядку спадання кількості вершин, які вони охоплюють. Також першим обробляється те піддерево, вершини якого знаходяться найближче до кореня піддерева (тобто різниця номера слова кореня піддерева і слів у піддереві мінімальна).

Знаходження канонічної форми слова

При приведенні англійського слова до його початкової форми всі його граматичні ознаки запам’ятовуються. Під час формування речення будуть грати роль тільки граматичні ознаки слова, російський переклад слова в канонічній формі, а також статистичні дані (наприклад, частота вживання деякого слова в деякому словосполученні з деяким іншим словом).

Таку ознаку, як число, вирішено отримувати із англійського слова. Модуль *Evo Inflector* реалізує алгоритм визначення числа, перетворення в множину чи однину слів. Алгоритм базується на роботі Даміана Конвея "*An Algorithmic Approach to English Pluralization*" [5]. У даній роботі розглядається алгоритм плюралізації, який повинен впоратися з трьома категоріями формами множини: універсальних (за замовчуванням), загальних правил на суфіксній основі слів, і конкретних виняткових випадках. Цей алгоритм, реалізовано в модулі *Evo Inflector*, що дає нам змогу легко отримати потрібну нам ознаку - число слова. Оскільки правила англійської мови щодо утворення однини становлять скінченну кількість, то доцільно використовувати готову базу цих правил.

Словник

Для знаходження перекладу слова за його початковою формою використовується база даних, оскільки пошук по ній достатньо швидкий. У створеній базі містяться слова, їх частини мови, англійський переклад, а також таблиці, що відповідають флективним групам слів.

Для побудови бази флективних груп потрібні таблиці з усіма можливими основами, з усіма флективними групами, і відповідно з усіма закінченнями. Кожне закінчення має свій код (дві букви російського алфавіту), яким відповідають певні ознаки (рід, число і відмінок). Пошуком по базі за основою досліджуваного слова можна отримати рід цього слова.

База флективних груп включає такі таблиці:

- Таблиця основ – це список всіх можливих основ російських слів.
- Таблиця флективних груп – номери відповідних флективних груп.
- Таблиця закінчень – всі можливі закінчення російських слів (включає нульове).
- Таблиця індексів закінчень – коди закінчень та ознаки (рід, число, відмінок), що їм відповідають.

Для заповнення таблиць флективних груп використовувались файли з заздалегідь записаною туди інформацією про основи, та номер флективної групи , що їй відповідає. Список флективних груп складає 3176 груп. В цих групах міститься різна кількість пар двобуквений індекс – закінчення, між якими є певні роздільники (символи, які допомагають при занесенні в базу). Окремо також зчитувався файл з розшифруваннями індексів , тобто вказаними наборами індекс та граматичні ознаки слів, наприклад, рід, число, відмінок, вид, стан тощо.

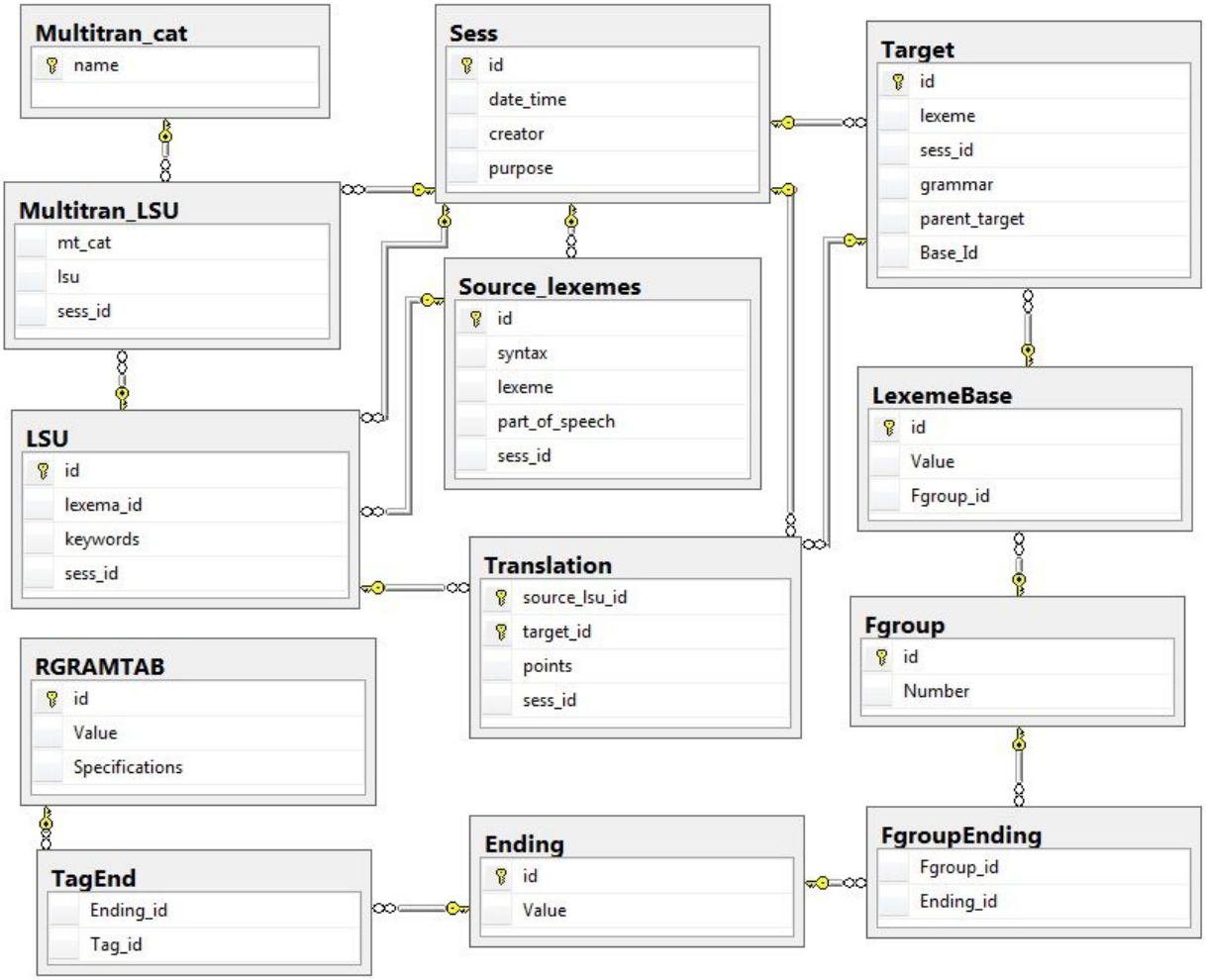

Схема 2. Структура словника

Приклад обробки речення

A database is an organized collection of data.

Теги

A/DT database/NN is/VBZ an/DT organized/VBN collection/NN of/IN data/NNS ./.

Зв'язки

det(database-2, A-1)
nsubj(collection-6, database-2)
cop(collection-6, is-3)
det(collection-6, an-4)
amod(collection-6, organized-5)
root(ROOT-0, collection-6)
case(data-8, of-7)
nmod(collection-6, data-8)

При обробці речення парсер надає кожному слову і розділовому знаку тег і порядковий номер, а також будує зв'язки залежностей.

У даному реченні:

- a, an – DT, артиклі
- database, collection – NN, іменники в однині, data – NNS, іменник у множині. Зводимо його до канонічної форми, а у граматичні ознаки заносимо число.
- is – VBZ, дієслово теперішнього часу 3 особи однини. Зводимо до інфінітиву – форми «be» – та зберігаємо граматичні ознаки.
- organized – VBN, дієприкметник минулого часу
- of – IN, прийменник

Правила

У російській мові прийменників немає, тому правила їх видаляють з дерева.

NN2(det.DT1.a)#C2

NN2(det.DT1.an)#C2

Іменник-підмет впливає на граматичні ознаки частини складного присудка – частки «to be», тому правило має охоплювати всі три слова. Так як речення в теперішньому часі, то у російському варіанті «бути» випадає. Після обробки піддерева вершина набуває тегу VB – дієслова в початковій формі.

NN3(nsubj.NN1,cop.VBZ2.be)#C1.пад:им&чис:Y&род:Z,C3.пад:им&чис:y&род:z(VB)

Дієприкметник граматичними ознаками залежить від головного для нього слова.

NN2(amod.VBN1)#ПРИЧ1.пад:x&чис:y&род:z,C2.пад:X&чис:Y&род:Z

Граматичні ознаки російських слів у структурі іменник-прийменник-іменник залежать від прийменника, тому для кожного з них будується окреме правило. У даному реченні прийменник «of» не перекладається, а залежний іменник ставиться у родовий відмінок.

NN1(nmod.NN3(case.IN2.of))#C1,C3.пад:род

У дереві залишається всього одна вершина, яка містить всі слова у правильному порядку та граматику цих слів:

database – С.пад:им&чис:Y&род:Z
organized – ПРИЧ.пад:x&чис:y&род:z
collection – С.пад:им&чис:y&род:z
data – С.чис:мн&пад:род

Число і рід слова «*database*» беремо з російського перекладу. Це жіночий рід, однина. Ці граматичні ознаки передаємо словам «*organized*» та «*collection*».

Російський переклад:

База данных – упорядоченный набор данных.

Список літератури

1. *Лори Тике* Технолого-агностический машинный перевод. Часть первая. / Лори Тике // Профессиональный перевод и управление информацией – 2015. – №4. – с. 16-20.
2. *Лори Тике* Технологически независимый машинный перевод: часть вторая. / Лори Тике // профессиональный перевод и управление информацией – 2015. – №5. – с. 34-38.
3. The Stanford Parser. http://nlp.stanford.edu/software/lex-parser.shtml http://nlp.stanford.edu:8080/parser/
4. *Marie-Catherine de Marneffe* Stanford typed dependencies manual. / *Marie-Catherine de Marneffe, Christopher D. Manning. //* http://nlp.stanford.edu/software/dependencies_manual.pdf
5. *Damian Conway* An Algorithmic Approach to English Pluralization. http://www.csse.monash.edu.au/~damian/papers/HTML/Plurals.html

References

1. LORI THICKE (2014) Technology agnostic machine translation: Part one, профессиональный перевод и управление информацией. No 4. pp. 16-20.
2. LORI THICKE (2014) Technology agnostic machine translation: Part two, профессиональный перевод и управление информацией. No 5. pp. 34-38.
3. The Stanford Parser. http://nlp.stanford.edu/software/lex-parser.shtml http://nlp.stanford.edu:8080/parser/
4. *Marie-Catherine de Marneffe* Stanford typed dependencies manual. / *Marie-Catherine de Marneffe, Christopher D. Manning. //* http://nlp.stanford.edu/software/dependencies_manual.pdf
5. *Damian Conway* An Algorithmic Approach to English Pluralization. http://www.csse.monash.edu.au/~damian/papers/HTML/Plurals.htm

Надійшла до редколегії 06.11.2015