Система усного перекладу на основі інтерпретації мовленнєвого сигналу в межах предметних областей

Микола Сажок, Валентина Яценко,
Міжнародний науково-навчальний центр інформаційних технологій та систем
40 просп. Академіка Глушкова, Київ 03680

Abstract

Spoken translation system based on speech understanding in subject area. The automatic speech understanding system within a subject area is considered. The system firstly recognizes the input sentence as a sequence of words, than it makes decision about the meaning type of the recognized sentence. Recognized words that missed in a target meaning type are caught by means of minimal edition distance estimation procedure. Strict, free and phonetic word based speech recognition word grammars for speech recognition are analysed. Results of their application for highly inflective languages with relatively free word order like Ukrainian are discussed and the respective translation system is presented.

1.Вступ

Задача розпізнавання усномовного сигналу мови передбачає не лише автоматичне відновлення текстів, що вимовляються людиною у вигляді слів, фраз, речень на природній мові. Не менш важливо розуміти, що сказано диктором. З одного боку, це дає змогу формувати більш адекватну відповідь розпізнавання. З іншого боку, маючи змогу оперувати смислами, легко перейти до того самого смислу, але вже іншою мовою, отримуючи таким чином систему усного перекладу.

В цій роботі представлено спроби оперувати смислами, але за певних обмежень. Ці обмеження не стосуються способу та манери вимовляння диктора. Він може формулювати думку без огляду на суворі граматичні або стилістичні канони. Тобто, слова в реченнях, які вимовляє диктор, можуть слідувати у настільки вільному порядку, допоки це не втрачає смисл. Слов'янські мови в цьому сенсі поводяться найбільш вільно, допускаючи не лише відносно вільно слідувати словам у реченнях, а і опускати певні слова, і при цьому не порушувати смисл. Крім порядку слідування, велика кількість словоформ дає змогу висловлювати одну і ту ж думку більш розмаїто.

Зауважимо, що перелічені особливості слов'янських мов значною мірою ускладнюють розв'язування задач розпізнавання та змістовної інтерпретації мовленнєвих сигналів.

Умови, які накладають обмеження в оперуванні смислами стосуються головним чином предметної області

Серед важливих практичних задач, пов'язаних з розпізнаванням мови, до яких відносяться система диктування текстів, довідкова система, система мовленнєвого керування різними пристроями, системи мовленнєвого діалогу (наприклад, по телефону) тощо, ми виокремимо систему усного перекладу. Її актуальність полягає, зокрема, в автоматизації розмовника, відомого всім у вигляді паперової книжки. Ця книжка розбита на теми, має безліч сторінок. Користувач вимушений шукати потрібну фразу та ще й «озвучувати» переклад іншою мовою своїм голосом. Натомість пропонується користувачеві лише вимовити фразу рідною мовою в обраній темі. Далі система все робить сама.

Такі системи є актуальними з огляду на використовування їх при розмові з іншомовною особою. Користувач не лише отримує переклад фрази іншою мовою, а ще й озвучення цієї фрази, що полегшить спілкування в іншомовному середовищі.

При побудові систем усного перекладу в межах предметних областей виникає ряд проблем, спільних з проблемами задачі розуміння мовленнєвого сигналу. Необхідно побудувати моделі всіх можливих речень мови діалогу, що виражають один і той самий зміст, генерації та пошуку найбільш правдоподібних сигналів. Хоч і не так явно, але у слов'янських мовах все ж існують обмеження на порядок слідування слів, які слід враховувати, використовуючи лінгвістичні знання.

Для дослідження та специфікації обмежень на допустимі послідовності слів у фразах використовувалися LISP-структури [1, 2]. На основі цих структур генерується величезна кількість речень, що мають один і той самий зміст. Втім, існує ряд обмежень на застосування цієї технології, пов'язаних як з суб'єктивним чинником при побудові LISP-структур, так і зі збільшенням обчислень, викликаних значним ускладненням графа розпізнавання.

В якості альтернативи до LISP-структур пропонувався спосіб оцінювання належності послідовності слів типам речень, що характеризують смисл [3]. Цей підхід потребує розвинення, зокрема, з метою врахування помилок розпізнавання.

Для моделювання обмежень на порядок слідування слів застосовувалися граматичні знання [3], що і надалі використовується в дослідженнях.

У розділі 2 ми дамо загальну структуру системи усного перекладу, розділ 3 присвячено заданню (специфікації) речень з урахуванням обмежень на допустимі послідовності слів, у розділі 4 проводиться вдосконалення способу оцінювання належності послідовності слів до типу речень, у розділі 5 описуються експериментальні дослідження.

2. Загальна структура системи усного перекладу в межах предметних областей

Розпізнавання та змістовна інтерпретація злитого мовлення виконується в єдиному взаємопов'язаному процесі. Кінцевою метою цього процесу є зміст

повідомлення, який передається послідовністю слів, та його переклад на іншу мову.

Розглянемо, в чому полягають і як взаємозв'язані задачі розпізнавання та інтерпретації злитого мовлення [1, 2]. Розпізнавання мови – це процес автоматичної обробки сигналу з метою визначення послідовності слів, які передаються цим сигналом. Змістовна інтерпретація мови – це процес автоматичної обробки мовленнєвого сигналу з метою виявлення змісту, що передається сигналом, та представлення цього змісту в певній канонічній формі, зручній для подальшого використання в системі усного перекладу.

Очевидно, що змістовна інтерпретація мови є більш високим ступенем узагальнення інформації, ніж розпізнавання, оскільки одну і ту саму думку можна виразити різними послідовностями слів. Оскільки кожну думку можна висловити різними реченнями в мові діалогу, але при цьому зміст не зміниться, то слід визначити певні обмеження на допустимі послідовності слідування слів у реченнях. Тому, при інтерпретації змісту мови різні речення, що передають одну і ту саму думку, повинні відображатися в один і той же результат, тобто відповідь розпізнавання (послідовність слів) не повинна суперечити синтаксису, семантиці та прагматиці предметної області.

Зважаючи на це, пропонується розглянути структуру системи усного перекладу в межах предметних областей (Рис.1). Задача змістовної інтерпретації злитого мовлення з метою подальшого перекладу ґрунтується на тому, що спочатку користувач має задати предметну область (ПО), з якою він бажає працювати. Для цього потрібно назвати цю ПО. Взагалі розглядається 15 ПО, з якими може працювати користувач. Активатор вибирає названу з 15 ПО і завантажує підсловник ПО з відповідними до цієї області ТР та граматику, за якою моделюються допустимі обмеження на послідовності слів у реченнях.

Рис.1. Структура системи усного перекладу в межах предметних областей

Диктор вимовляє мовою 1 деяке речення, що розпізнається з урахуванням акустичної моделі та побудованої згідно словника відповідної ПО та граматики лінгвістичної моделі (LM). Потім обирається *n* кращих послідовностей слів і порівнюється з нагенерованими моделями речень, які можуть задавати відповідний ТР. Використовуючи імовірнісне оцінювання, приймається рішення про належність розпізнаної послідовності слів до ТР. За цим ТР визначається ТС і інтерпретатор знаходить відповідний ТС іншою мовою. На виході ми маємо отримати текст мовою 2, який до того ж озвучується відповідною системою озвучення текстів мовою 2.

В описаній структурі перекладу залишається досить складною задача інтерпретації розпізнаного сигналу. Тому, перш за все слід навчитися економно задавати всі можливі допустимі речення в мові діалогу. Для вирішення цього питання є декілька шляхів. Один з них – побудова LISP-подібних структур та визначення за їх допомогою обмежень на допустимі послідовності слів. Цей спосіб було розглянуто в попередніх роботах.

В якості альтернативи до LISP-структур було розглянуто спосіб оцінювання належності послідовності слів типам речень, що характеризують смисл.

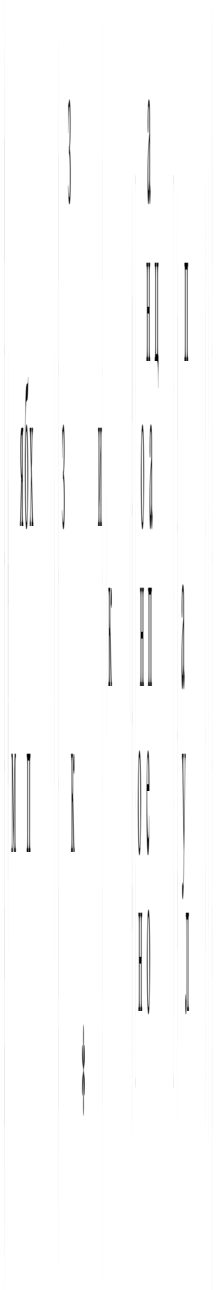
Таким чином, автоматичний переклад фрази вимовленої мовою 1 на мову 2 з озвученням результату, за допомогою пропонованої структури усного перекладу в рамках предметних областей буде полягати в тому, щоб спочатку для сигналу, що задається (вимовляється диктором), знайти найбільш правдоподібний тип речення серед всіх типів речень, які задають тип смислу, а потім визначити сам тип смислу того змістовного висловлювання, що було вимовлене, та знайти для нього відповідний тип смислу в мові 2. Насамкінець, речення, отримане мовою 2, озвучується.

Далі буде розглянуто способи задання всіх можливих речень мови діалогу, що виражають один і той самий зміст, генерації та пошуку найбільш правдоподібних типів речень та розроблення обмежень на допустимі послідовності слів згідно структур, якими можна представити речення.

3. Моделювання обмежень на допустимі послідовності слів

Оскільки структура перекладу повинна працювати в межах предметних областей, то пропонується розглянути певну ієрархію розташування речень. Мається на увазі, що всі мислимі речення мови діалогу розіб'ємо на предметні області (ПО) по типу паперового розмовника для іноземних мов. Кожна ПО складається із скінченої множини типів смислів (ТС). Наприклад, стосовно ПО «подорож» відвідання ресторану, типи смислів виражаються питаннями про бронювання столика, меню, замовлення тощо. Кожній ПО відповідає скінчена множина типів смислів. В кожен ТС входить множина еквівалентно змістовних типів речень (ТР), які описуються LISP-структурами [1–3]. Тип речення – це конструкція, що економно задає множину речень, отриманих з одного речення незалежними допустимою заміною та допустимою перестановкою чи випадінням слів та словосполучень.

Розглянемо приклад ТР для ПО «Подорож», що стосується замовлення квитків (ТС – замовлення квитка).



В дужках () вказані підсловники, які можна переставляти місцями, а в [] – які не можна переставляти. Переставляти підсловники можна лише всередині старших дужок. Символ * означає порожнє слово.

Неважко переконатися, що наведений тип речення задає 4!·1·4·1·3·3=864 різних речень, допустимих в мові діалогу та таких, що виражають один і той самий зміст щодо замовлення квитка. Серед цих речень є, наприклад, і такі:

Квиток я б хотів замовити на цей потяг.

На перший автобус мені потрібно забронювати квиток.

Тобто ми бачимо, що в даний ТР включено багато синтаксично допустимих речень розмовної мови. Але враховуючи вільний порядок слів, серед цих речень будуть утворені речення, які не є типовими для розмовної мови. Тому, щоб відкинути нетипові речення було запропоновано ввести певні обмеження, які вказують порядок слідування слів.

Всі речення мови діалогу можна задавати за допомогою ТС і відповідних їм ТР, використовуючи структуру, наведену у прикладі. За допомогою LISP-структур генерується величезна кількість речень, що мають один і той самий зміст. Оскільки побудова LISP-структур є досить громіздкою, потребує багато ручної роботи, то було розроблено автоматизований специфікатор предметних областей.

Для побудови всіх можливих речень мови усного діалогу можна використовувати так звану орієнтовану семантичну мережу (ОСМ) [1,2]. Ця мережа одночасно задає обмежену граматику порядку слідування слів, яку можна використовувати при розпізнаванні.

Альтернативною до цієї граматики є граматика вільного поряду слідування слів. Між цими протилежними за суттю граматиками може бути побудовано безліч інших відносно вільних або відносно обмежених граматик. Ми пропонуємо дещо обмежити вільну граматику за рахунок лінгвістичного поняття про фонетичне слово.

Під фонетичним словом розуміємо слово з невіддільними від нього супутніми словами. Наприклад, невіддільними є прийменник від іменника або прикметника, часка «не» спереду дієслова і частка «б» позаду нього. Пропонована нами відносно вільна граматика матиме такий вигляд:

(pau <[проклітик] [прийменник | проклітик] нейтральне [енклітик]> pau),

де вміст кутових дужок може повторюватися, а вміст квадратних дужок може бути опущений, pau – слово-пауза на початку та в кінці фрази.

Втім, за такої граматики прийняття рішення щодо смислу речення не є очевидним. Це будемо розглядати в наступному розділі.

4. Статистичне оцінювання належності послідовності слів до типу речень

При розпізнаванні в умовах граматики, що не задає строгих обмежень на послідовності слів, очевидно можуть бути отримані відповіді розпізнавання, що не входять у множину речень, які згенеровані певним типом речень. Це може бути зумовлено як помилками при розпізнаванні, так і при формуванні типів речень експертом. Крім того, сам користувач може вимовити речення з різного роду відхиленнями або аграматизмами, наприклад повторити деяке слово двічі.

Тому пропонується оцінювати ймовірність типу речення *ST* з ОСМ за умови розпізнаної послідовності слів $(w_1, w_2, ..., w_n)$ та оголошувати відповіддю інтерпретації той тип речень – *ST**, для якого ця ймовірність є найбільшою:

$$ST^* = \arg\max_{ST} P(ST / w_1, w_2, ..., w_n) \quad (1)$$

Імовірність у лівій частині (1) може бути переписана за формулою Байєса в такому вигляді:

$$P(ST / w_1, w_2, ..., w_n) = \frac{P(ST)}{P(w_1, w_2, ..., w_n)} P($$

(2)

Розглядаючи послідовність $(w_1, w_2, ..., w_n)$ як марківський процес, подаємо кожний із множників умовної імовірності у правій частині (2) у вигляді:

$$P(w_1, w_2, ..., w_n / ST) = \prod_{k=1}^{n} P(w_k / ST, w_{k-m},$$

, (3)

$$P(w_1, w_2, ..., w_n) = \prod_{k=1}^{n} P(w_k / w_{k-m}, ..., w_{k-1}$$

, (4)

де $m \geq 0$ – порядок процесу.

Оцінювання кожного з множників з правої частини виразів (3) і (4) може виконуватися різними способами в залежності від обраного порядку процесу.

Ми розглядали найпростіший випадок, коли $m = 0$. Тоді, враховуючи формулою Байєса, вираз (2) подаємо у вигляді:

$$P(ST / w_1, w_2, ..., w_n) = P(ST) \prod_{k=1}^{n} P(ST / w_k)$$

, (5)

Тут логічно зробити припущення щодо рівноімовірності всіх типів речень. Хоч насправді, деякі смисли трапляються частіше за інші, і це залежить від попереднього змісту. Залишається обчислити вираз вигляду $P(ST_k / w)$. Для цього розглянемо $ST(w_k)$ – множину типів речень, в яких зустрічається слово w_k. Тоді

$$P(ST/w_k) = \begin{cases} |ST(w_k)|^{-1}, & \text{якщо } ST(w_k) \neq \varnothing, \\ \alpha(ST, w_k), & \text{в іншому випадку.} \end{cases}$$

(6)

Вираз $\alpha(ST, w_k)$ має зміст імовірності того, що слово w_k розпізнано помилково замість деякого слова w: $ST(w) \neq \varnothing$. Цю ймовірність можна оцінити на основі деякої міри мінімальної редакторської правки $d(w_k, w)$, наприклад, відстані Левенштейна. При обчисленні цієї міри штрафуються вставки, видалення та заміни символів фонемного тексту порівнюваних слів. Отже, вираз $\alpha(ST, w_k)$ пропонується оцінювати таким чином:

$$\alpha(ST, w_k) = \left(1 - \min_{ST(w) \neq \varnothing} \frac{d(w_k, w)}{L(w)}\right) \times \times ST^{-1}\left(\arg\min_{ST(w) \neq \varnothing} d(w_k, w)\right), \quad (7)$$

де $L(w)$ – кількість фонем у слові w.

Рішення щодо приналежності розпізнаної послідовності слів певному типу речень приймається на основі (1)–(7).

Для апробації цього способу було проведено серію експериментальних досліджень.

5. Експериментальні результати

В якості експериментальних даних було розглянуто англійсько-український розмовник, що розробляється врамка. Розмовник складається з 3800 речень, які назвемо базовими фразами. Ці фрази розділені на 15 предметних областей, кожна з яких має свої типи смислів та типи речень. Для прикладу було розглянуто одну з 15 ПО, а саме «Повсякденні фрази». Ця ПО містить 47 ТЗ та 201 базових речень, в середньому по 5 базових речень на тип змісту. Щоб задати всю множину речень, яка породжується базовим реченням, кожне базове речення було розмічено у відповідності до описаних LISP-структур. Таким чином, для кожного базового речення було побудовано тип речення у вигляді LISP-структури.

Так, наприклад, для типу змісту прохання про допомогу, побудовані такі ТР:

((будь ласка | *) (допоможіть) (мені | *) ((вирішити | розв'язати) (цю проблему)))

((будь ласка | *) (допоможіть) (мені | *) (у [цій | *] (справі)))

У наведеному прикладі підсловники містять здебільшого по одному слову, але загалом потужність окремо взятого підсловника може бути більшою в залежності від кількості синонімів.

Було розроблено програмне забезпечення, яке з одного заданого таким чином ТР дає змогу будувати множину всіх речень шляхом відповідних перестановок чи заміни слів та словосполучень. В результаті застосування цієї програми до згаданих вище 201 ТР, було отримано 1045 фраз, не враховуючи змінні параметри. Якщо врахувати змінні, то фраз буде 4337. У словнику нараховується 290 слів.

Для ТР, взятого з наведеного прикладу, – ([чи | *] (([не | *] [допоможете]) (Ви мені) (вирішити | розв’язати) (цю проблему))) було згенеровано 24 фрази, без урахування змінних, серед них:

$p289 = $w11 Ви мені $w24 допоможете цю проблему $w19;

$p295 = $w11 цю проблему $w24 допоможете $w19 Ви мені ;

Тут змінними параметрами є $w11=(чи | *); $w19 = (вирішити | розв'язати); $w24 = (не | *). Враховуючи, що кожна змінна може приймати 2 значення, кожна фраза буде мати 8 варіантів. Отже, для даного ТР отримаємо 8·24=192 речення, з урахуванням змінних.

Розпізнавання фраз (речень) проводилося на основі пофонемного розпізнавача [4] за умов обмеженої (на основі LISP-структур), вільної та відносно вільної (на основі фонетичних слів) послівних граматик.

Акустичні моделі для розпізнавача створено на основі мовленнєвого корпусу окремо вимовлених слів, в якій брало участь 60 дикторів. Засобами [5] проведено навчання 55 прихованих марківських моделей фонем. Максимальна кількість нормальних законів у суміші – 20.

Для експерименту довільним чином було вибрано 200 фраз серед згенерованих 4337, до яких було застосовано алгоритми пофонемного розпізнавання

мовленнєвих сигналів в умовах обмеженої та вільної граматик відносно слів. На відміну від попередніх експериментів [3] при оцінюванні акустичних параметрів фонем використовувався корпус кооперативу дикторів. Прийняття рішень щодо смислової інтерпретації здійснювалося на основі (1)–(5).

Таблиця 1. Результати розпізнавання та смислової інтерпретації 200 речень з двох предметних областей.

Тип граматики	Надійність розпізнавання (%)		
	послівна	по реченнях	по типах смислу
Обмежена	97,3	93,5	97,0
Вільна послівна	52,1	2,5	83,0
Відносно вільна	77,9	21,0	95,5

За умов обмеженої граматики швидкість розпізнавання в 10 разів перевищувала реальний час, за умов вільної та відносно вільної граматик розпізнавання відбувалося швидше реального часу на ресурсах нетбука.

На основі проведених досліджень розроблено демонстраційне програмне забезпечення для перекладу речень, вимовлених українською мовою, на англійську мову (Рис. 2). При цьому слідування слів в українському реченні може бути будь-яким із допустимих. Реченню, вимовленому українською мовою, ставиться у відповідність англомовний тип змісту або речення, а перше речення цього типу змісту оголошується результатом перекладу.

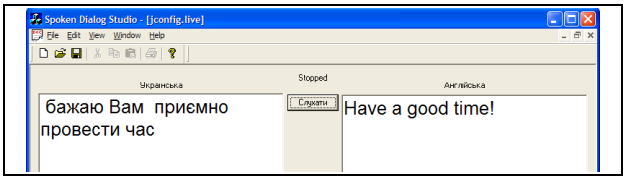


Рис. 2. Демонстраційне програмне забезпечення моделі усного фразника-перекладача.

6. Висновки

Розглянута в роботі система усного перекладу є електронним аналогом паперового розмовника, взаємодія з яким відбувається найбільш природними способом – голосом. При розпізнаванні вимовленої користувачем фрази використовуються лінгвістичні та семантичні знання щодо обраної предметної області. Введені при цьому «м'які» обмеження на порядок слідування слів дають змогу підвищити надійність розпізнавання, не підвищуючи вимог до обчислювальних ресурсів.

Розвинений спосіб оцінювання ймовірності приналежності послідовності слів деякому типу речень дає змогу використати інформацію, що міститься у помилково розпізнаних словах. Теоретично, навіть не маючи жодного правильно розпізнаного слова, залишається шанс отримати правильну смислову інтерпретацію.

Експериментальні дослідження підтвердили, що імовірнісний підхід при смисловій інтерпретації показує досить високі результати для обмеженої та відносно вільної граматик слідування слів.

На основі експериментальної моделі розроблено програмну модель усного словника-перекладача для перекладу з української мови на англійську в межах предметної області.

Одні і ті ж самі фрази з різною інтонацією можуть виражати як питальне речення, так і розповідне. Отже, в подальшій роботі слід дослідити можливість розпізнавання інтонації та ритму (просодики) з метою автоматичного розставляння розділових знаків у розпізнаних фразах.

Надалі також планується ставити у відповідність україномовній фразі більш точний англомовний відповідник серед типів речень з типу смислу.

Література

1. Т.К. Винцюк. Анализ, распознавание и смысловая интерпретация речевых сигналов. – Киев. Наукова думка, 1987.

2. Т.К. Винцюк. Учет синтаксиса языка при распознавании слитной речи. – Киев. Институт кибернетики, 1975.

3. T. Vintsiuk, M. Sazhok, V. Yatsenko. Comparison of word grammars for spoken translation in subject area // Proceedings of the 13th International Conference on Speech and Computer – SpeCom'2009. – P. 494-497

4. Lee, T. Kawahara and K. Shikano, "Julius – an open source real-time large vocabulary recognition engine." In Proc. European Conference on Speech Communication and Technology (EUROSPEECH), 2001, pp. 1691–1694

5. Young S.J. et al., HTK Book, version 3.1, Cambridge University, 2002.

6. T. Vintsiuk, M. Sazhok. Multi-Level Multi-Decision Models in ASR. In Proc. Of 10th Int. Conf. "Speech and Computer", Patras, Greece, 2005, pp. 69-76.