Моделювання усного діалогу між людиною і технічною системою

Ніна Васильєва, Микола Сажок, Ольга Сухоручкіна, Валентина Яценко

Міжнародний науково-навчальний центр інформаційних технологій та систем
40 просп. Академіка Глушкова, Київ 03680
{n.vassilleva ; sazhok ; sukhoru ; yatsenko.valya} @gmail.com

Abstract

This paper describes the formation of data and knowledge base for a spoken interpreter system based on the generative model of speech signal understanding within subject areas. The proposed way to progress the dialog consists in step-by-step meaning type completing in agreement with the technical system state. The proposed structure allows for operating the subject areas, languages, meaning types and sentence types.

1. Вступ

Діалогові системи – новий тип систем, які дають змогу людині отримувати необхідну інформацію, користуючись властивими їй засобами спілкування (мовлення, жести, міміка тощо). Ці системи призначені для вирішення задач доступу до інформації автоматично, тобто без присутності людини-експерта [1].

Спектр застосування діалогових систем неймовірно широкий. Ними послуговуються для автоматичної перевірки знань, використовують як автоматичні служби підтримки користувачів, як розважальні програми. Наприклад, існують діалогові системи типу "людина—комп'ютер", що дають змогу отримувати інформацію про розклад руху транспорту, поточний стан банківського рахунку та бронювати квитки на потяг тощо.

Інший тип систем – "людина–комп'ютер–людина". Це системи вищого інтелектуального рівня, оскільки вони не тільки розпізнають мовлення користувача, а й забезпечують взаємодію з ним за допомогою голосу. Одним із прикладів такої системи є переклад з однієї мови на іншу шляхом "мовлення–мовлення" з метою автоматизації функцій паперового розмовника-перекладача [2], [3]. Такі автоматизовані розмовники на основі технології інтерпретації мовленнєвого сигналу є актуальними в наш час. Спілкуючись з іноземцем, користувач отримує переклад фрази іншою мовою у візуальній або звуковій формі, що значно спрощує спілкування.

Поширення і поглиблення використання різних інформаційних систем приводить до необхідності надання користувачу максимальних зручностей при роботі з технічною системою у режимі діалогу.

В останні роки у світі значну увагу приділяють розробці зручного інтерфейсу, що має на увазі спрощення взаємодії користувача з технічною системою. Звичним стандартом стали віконні системи, забезпечені візуальними засобами керування відповідно до принципів *GUI* (*Graphical Users Interface*). Керування інформаційними системами більше не вимагає пошуку потрібної клавіші на клавіатурі. Усе здійснюється наочно, і користувач бачить результати своїх дій на моніторі комп'ютера.

Все ж найбільш природним способом спілкування для людини є усна мова. І в цьому сенсі існує величезна прогалина між тим, як людина взаємодіє з собі подібними, і взаємодією людини і комп'ютера на сучасному етапі розвитку науки.

Що ж таке діалог? Це двосторонній обмін інформацією (розмова, спілкування) між двома сторонами (у нашому випадку – між людиною та технічною системою) у вигляді запитань/спонукань і відповідей.

Пропонуємо розвинути технологію розуміння мовленнєвого сигналу з метою її використання для діалогу з технічними системами. У попередніх роботах при створенні систем смислової інтерпретації мовлення для флективних мов з відносно вільним порядком слідування слів використовується багаторівнева структура [1]–[3]. Розглядаються предметні області (ПО), які містять обмежену кількість типів смислів (ТС), що у свою чергу містять типи речень (ТР). За типами речень генеруються еталонні тексти та відповідні сигнали речень, які порівнюються з пред'явленим сигналом. У роботах, зокрема, приділялась увага компактній специфікації типів речень, моделюванню аграматизмів, елементів спонтанності та відновленню типу смислу інтерпретації у випадку помилки відповіді розпізнавання.

Результат розпізнавання залежить від багатьох факторів, у тому числі й від того, який рівень деталізації слова використовувати (фонеми, склади, морфеми). Підвищення рівня складності елементу розпізнавання призводить до покращення надійності та коректності [4]. Похибки розпізнавання, які можуть виникати, мають бути контрольовані таким чином, щоб отримати правильний підсумковий результат послівного розпізнавання та/або смислової інтерпретації мовлення.

У системі усного діалогу з технічною системою важливо не лише розпізнати вимовлену фразу, зрозуміти її, а й адекватно відреагувати. Це може бути початок виконання сформульованого завдання або уточнення параметрів завдання в залежності від стану технічної системи та контрольованого нею середовища.

Після базової постановки задачі розглянуто модель взаємодії людини та технічної системи на прикладі робота. Розділ 4 присвячений базі даних і знань. У п'ятому розділі аналізуються акустичній аспекти.

2. Базова постановка задачі

В основі діалогу лежить діалогічна єдність: вираження думок і їх сприйняття, реакція на них, що знаходить відображення у структурі цього акту мовлення. Діалог складається з взаємопов'язаних реплік співрозмовників.

Розглянемо, яким чином можна досягнути взаєморозуміння людини та технічної системи.

Коли людина дає команди або ставить питання, технічна система повинна вміти виконувати ці завдання,

ставити уточнюючі питання, реагувати та відповідати в процесі спілкування. Для такої взаємодії задачі розпізнавання та розуміння мовленнєвого сигналу повинні виконуватися в єдиному взаємопов'язаному процесі [3].

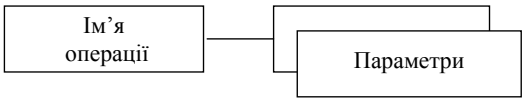
Кінцевою метою створення інтелектуальної системи усного діалогу, реалізованої на основі задач розпізнавання та змістовної інтерпретації злитого мовлення, які виконуються в єдиному взаємопов'язаному процесі "мовлення–мовлення", є виявлення змісту повідомлення, його інтерпретація та відповідна реакція з урахуванням стану технічної системи.

3. Модель взаємодії людини та технічної системи

Одним з прикладів сучасних технічних систем, розробкою яких займаються все більше наукових та виробничих колективів, є мобільні роботи сервісного призначення. Найбільший інтерес для сучасного ринку високотехнологічних систем спрямований на сервісні роботи для непрофесійного користувача. При розробці саме таких роботів суттєву увагу приділяють реалізації людино-машинного інтерфейсу, у тому числі з використанням інформаційних технологій розпізнавання та синтезу мовлення. Схему спілкування користувача з роботом як об'єктом управління (ОУ) показано на рис. 1.

Рисунок 1: Схема спілкування користувача з роботом

Залежно від рівня автономних можливостей робототехнічної системи, тобто від типу її керувальної системи, участь людини-користувача у процесі виконання роботом певних дій може суттєво відрізнятися [5]. У системах так званого командного типу людина-оператор повинна кожну елементарну дію технічної системи активізувати власною командою, як правило, у вигляді речення з досить простою структурою. Набір команд і їх можливих параметрів складає словник предметної мови інтерфейсу конкретного робота з управлінням командного типу (рис. 2).



< Переміститись > *< вперед, 1 метр >*

Рисунок 2: Типова структура речення інтерфейсу систем командного типу

Після розпізнавання мовленнєвого сигналу користувача для підтримки людино-машинного діалогу командного типу потрібна лише перевірка наявності саме таких елементів мови у словнику та активізація системою керування відповідних дій робота.

Від сучасних інтелектуальних роботів очікується можливість автономно виконувати завдання користувача, які сформульовані у вигляді досить узагальнених змістовних висловлень. Підсистема автоматичного ведення діалогу інтерфейсу таких технічних систем повинна виконувати перевірки відповідності цільових станів робота, які автоматично визначаються саме з комп'ютерного аналізу висловлювання користувача, можливостям управляючої системи робота привести його із поточного стану до заданого цільового. Тобто, крім лінгвістичної моделі діалогової системи обов'язково потрібна семантична модель мови саме керувальної системи робота та засоби її автоматичного аналізу. Іншими словами, для всіх можливих висловлювань користувача при формулюванні завдань автономному роботу, модуль розпізнавання та аналізу мовлення відповідає за коректну, з точки зору керувальної системи робота, постановку йому завдання, що може бути досягнуто автоматичним формуванням послідовності уточнювальних запитань для мінімізації розходжень отриманого роботом завдання з параметрами цільового стану робота, що однозначно сприймаються, і можливостями керувальної системи (рис. 3).

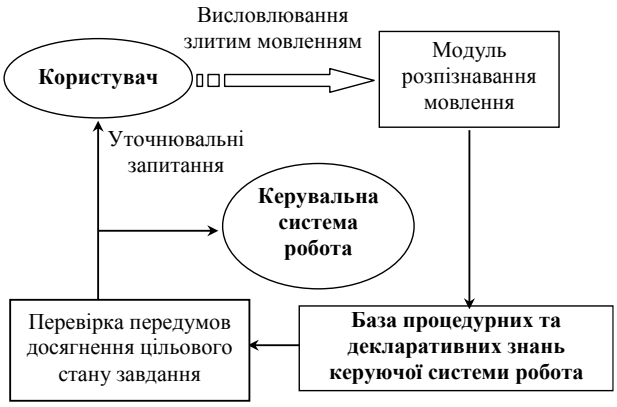


Рисунок 3. Структура підсистеми ведення діалогу

4. Структура БД і З

У системі усного діалогу типу "людина–технічна система–людина" взаємодія сторін відбувається природним шляхом – голосом. Усний діалог між людиною і технічною системою розглядається в рамках певної предметної області. Представлена інформація повинна бути зв'язною та структурованою. Саме така інформація формує базу даних, з якою може працювати експерт-лінгвіст. База даних є інформаційною моделлю предметної області [6].

Розглянемо загальну структуру системи усного діалогу людини і технічної системи (рис. 4). Ми пропонуємо розглядати діалог як заповнення один за одним типів речень одразу або поступово, шляхом озвучення уточнень однієї сторони до іншої. Уточнення у вигляді канонічних форм передаються на генератор природномовних текстів, що далі озвучуються. Характерною відмінністю усного діалогу від голосового керування пристроєм є те, що технічна система сама

формує уточнення з урахуванням змін контрольованого нею середовища, наприклад, тривимірної сцени.

Наведемо приклад, що пояснює запропоновану структуру. Нехай людина говорить у мікрофон таку фразу у рамках заданої ПО, як “Перейди вперед”. Мовленнєвий сигнал надходить в модуль розпізнавання та інтерпретації мовлення, в результаті чого отримуємо відповідь розпізнавання та інтерпретації у вигляді імені типу смислу MID (*Meaning Type Identification*) та розпізнаного тексту. У нашому випадку тип смислу набуває значення “*navigate*” (MID = “*navigate*”).

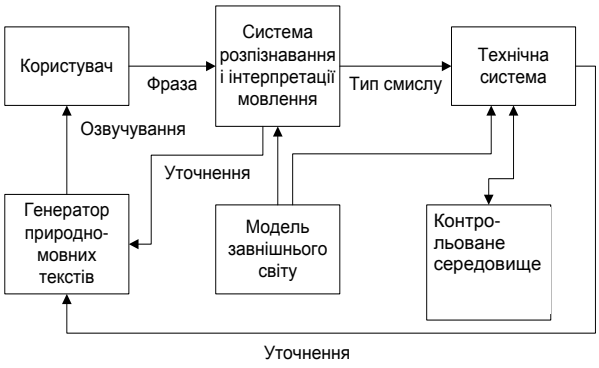


Рисунок 4. Загальна структура системи усного діалогу між людиною та технічною системою

Система звертається до бази даних і знань, де в таблиці типів смислу знаходить відповідне представлення типу речення: “Перейди $*where* [$*valid*]” (таблиця 1). Цей вираз містить метаслова тобто параметри команди для технічної системи, подані у вигляді змінних $*where* та $*valid*. Наявність метаслів у типі речення вимагає уточнення у тому випадку, якщо фраза не була вимовлена або розпізнана повною мірою. Слід зауважити, що метаслова, подані у квадратних дужках, не є обов’язковими при формуванні типу речення, тому уточнення не вимагають.

Таблиця 1. Приклад специфікації типів смислу

MID	MT ua	MT eng	MT ru
take_obj	візьми [об’єкт] $mobj [$wobj] [$direction]	…	…
come2obj	підійди до [об’єкту] $obj [$wobj] [$direction]		
	…		
navigate	перейди $where [$valid]		
	…		

Система, знайшовши метаслово $*where*, звертається до таблиці питань до метаслів та, в разі відсутності необхідних метаслів, ставить уточнююче питання відповідно до таблиці 2. Уявімо, що користувач не вказав напрямок пересування. Тоді, згідно таблиці, система усного діалогу повинна запитати “куди?” або “в якому напрямку?”. Користувач має дати коректну відповідь на це питання, наприклад, сказати: “вперед” або “до стільця”. Тоді змінна набуде значення $*where* = “вперед”. В залежності від стану технічної системи, вона може ініціювати «необов’язкові» уточнення, наприклад: “до якого стільця?”.

Імена та значення змінних визначаються таблицею 3. Змінні, які передаються діалоговій системі, можуть рекурсивно визначатися через інші змінні. У цьому випадку змінна $*where* визначається через змінні $*position*, $*direction* і $*range*.

Таблиця 2. Приклад специфікації запитань до метаслів у контексті предметних областей

ID	MID	Q:ua
$where	*	куди?
	*	в якому напрямку?
…	*	
$obj	$which_obj	об’єкт ще не знайдено, уточніть
	*	який об’єкт?
	*	назвіть об’єкт
	*	що за об’єкт?
	*	де знаходиться об’єкт?
...		

Таблиця 3. Приклад рекурсивного визначення змінних

ID	$val	Ua	Eng	Ru
$where		$position^{V}direction^{V}range		
$position	pos01	в задане положення	…	…
	pos02	в початкове положення		
	pos03	в стартове положення		
	...			
$direction	dir01	вперед		
	dir03	до об’єкта		
	dir04	назад		
	...			
$range		$linear^{V}$angular		
	…			

Система усного діалогу в кожний момент отримання *i*-ї відповіді розпізнавання та інтерпретації зберігає передісторію:

$$\text{History\_cash}_i = \{(\text{MID}_j, \text{params}_j) \mid 0 < j < i\},$$

де MID_j та params_j є, відповідно, іменем типу смислу та іменами параметрів, отриманих внаслідок *j*-ї відповіді розпізнавання та інтерпретації. У цій пам’яті зберігається інформація, що гіпотетично дає змогу зменшити кількість уточнень.

У момент, коли технічна повною мірою визначається тип змісту, відбувається реакція технічної системи у вигляді мовленнєвого повідомлення або дії на об’єкти оперативного середовища.

5. Вибір акустичної моделі розпізнавання мовлення

Слід розглянути питання, яку акустичну модель розпізнавання мовлення обрати для найефективнішої роботи технічної системи.

Параметри акустичної моделі оцінюються на основі мовленнєвого корпусу, що складається зі структурованої множини мовленнєвих фрагментів, текстового опису цих фрагментів, а також інструментарію для оперування всією множиною даних корпусу.

Одним зі способів формування мовленнєвого корпусу є запис диктора, який зачитує певний текст, у якому представлене все фонетичне розмаїття українського мовлення, описаний в [7]. Це дає змогу уникнути етапу ручного транскрибування та сегментування, а також одночасно створювати текстовий корпус у відповідній предметній області. У його основі – електронні тексти, що знаходяться у вільному доступі в Інтернеті. Інший спосіб – записати ефірне мовлення та транскрибувати його. Таким чином створено багатодикторний корпус мовлення, описаний в [8]. Обидва із запропонованих корпусів мають свої переваги та недоліки.

Для досліджень, описаних у [9], елементарною одиницею, використаною при побудові запропонованих акустичних моделей, була фонема. Було проаналізовано такі варіанти акустичної моделі розпізнавання: модель, побудована тільки на злитому мовлені; модель, яка об'єднує злите мовлення та ізольовані слова; модель, яка не враховує або враховує лише частково наголошеність голосних.

В експериментальних дослідженнях оцінювалися показники фонемної помилки (PER – *Phoneme Error Rate*), що відображає відношення між різницею правильно розпізнаних фонем і помилкових вставок до загальної кількості фонем.

При розпізнаванні була використана граматика вільного порядку слідування суб-слівних елементів: фонем, складів тощо. Ряд експериментів проведено накладанням обмежень на послідовності елементів із застосуванням біграмної лінгвістичної моделі (ЛМ) на фонемно-морфемному рівні.

Таблиця 4. Показники фонемної помилки PER (%) розпізнавання акустичної моделі, побудованої на одно- та багатодикторному мовленнєвих корпусах для КВ злитого мовлення на основі різних мовленнєвих образів.

Назва КВ	Фонема	Відкритий склад	Склад за правилами складоподілу	Фонема	Відкритий склад	Склад за правилами складоподілу
	Однодикторний корпус			Багатодикторний корпус		
Випадкова	24,8	27,5	24,8	47,3	45,9	37,5
Частотна	27,7	24,2	22,0	50,9	38,5	45,2
Вікіпедія	28,2	31,8	28,2	49,5	40,9	40,2

Було здійснено попередні експерименти на багатодикторній акустичній моделі, елементарною одиницею якої є фонема-трифон. Використана в дослідах ЛМ побудована на текстовому корпусі, з якого обиралися тексти для однодикторної мовленнєвої бази та тексти контрольних вибірок (КВ). У таблиці 4 наведені результати розпізнавання акустичної моделі, побудованої на одно- та багатодикторному мовленнєвих корпусах для КВ злитого мовлення на основі різних мовленнєвих образів, застосовуючи біграмні ЛМ.

Порівнюючи результати розпізнавання акустичних моделей, побудованих на одно- та багатодикторному мовленнєвих корпусах, можемо зробити висновок, що від використаної ЛМ також залежить ефективність розпізнавання. У рамках однієї предметної області результати розпізнавання, навіть для біграмної ЛМ, обнадійливі. Таким чином, можна припустити, що створені ЛМ способом, описаним в [10], відкриють перспективи використання акустичної моделі багатодикторного корпусу злитого мовлення, елементарною частиною для навчання буде не тільки фонема, а й контекстно залежна фонема-трифон.

6. Висновки

Показано, що підхід до смислової інтерпретації мовлення в рамках генеративної моделі [1] може служити основою для створення технологій та систем усного діалогу природною мовою між людиною та технічною системою.

Запропонований спосіб формування бази даних і знань систем усного діалогу дає змогу ефективно створювати такі системи для різних предметних областей із залученням експертів-лінгвістів.

Надійність розпізнавання мовлення, а відповідно і коректність інтерпретації залежить від вибраного мовленнєвого образу (фонеми, склади, морфеми). Покращення результатів розпізнавання відбувається як за рахунок ускладнення лінгвістичної частини, так і за рахунок використання найбільш відповідних акустичних моделей.

Планується дослідити вплив лінгвістичної та акустичної складових технології розпізнавання мовлення на покращення результатів змістовної інтерпретації висловлювань у системі усного діалогу.

Література

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

[2] В.В. Яценко. Параметризація типів речень предметної області для системи усного фразника-перекладача. // Науково-теоретичний журнал Штучний Інтелект (ШІ'2011), №4, стор. 134-142.

[3] Микола Сажок, Валентина Яценко. Система усного перекладу на основі інтерпретації мовленнєвого сигналу в межах предметних областей. // Праці Міжнар. конференції УкрОбраз'2010 – Київ, 2010, С.103-106.

[4] Васильєва Н.Б. Використання граматик вільного порядку слідування фонем і складів для пофонемного розпізнавання злитого мовлення. // НТЖ Штучний Інтелект (ШІ'2011), №4. стор. 80-86

[5] Попов Э.В. Общение с ЭВМ на естественном языке. — М.: Наука. Главная редакция физико-математической литературы, 1982. — 360 с.

[6] Валентина Яценко. Автоматизовані засоби формування лінгвістичної бази даних і знань для системи усного перекладу. // Науково-теоретичний журнал Штучний Інтелект (ШІ'2012), №4, стор. 211-220.

[7] Васильєва Н.Б. Дослідження невідповідності шкали акустичної та лінгвістичної моделей розпізнавання злитого українського мовлення. // НТЖ Штучний Інтелект (ШІ'2012), №3, стор. 118-125.

[8] Valeriy Pylypenko, Valentyna Robeiko, Mykola Sazhok, Oleksandr Radoutsky, Nina Vasylieva. Ukrainian Broadcast Speech Corpus Development. Міжнародна конференція "Speech and Computer – SpeCom".

[9] Васильєва Н.Б. Федорин Д.Я. Проблеми створення систем розпізнавання мовлення для різних комп'ютерних платформ. Штучний Інтелект (ШІ'2013), №4,. С. 158-167

[10] T. Vintsiuk, M. Sazhok, V. Yatsenko. Interpretation of continuous pronunciation for spoken dictionary-interpreter. In Proc. Of 12th Int. Conf. "Speech and Computer", Moscow, RF, 2007, pp. 170-175.