Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Implementation of a new language into a rule-based Spoken Dialogue System

MASTER

NATURAL LANGUAGE PROCESSING

Speciality :

Multilingual Engineering

by

Jielei LI

Thesis Director :

Damien NOUVEL

Supervisor :

David HOUSSIN

2017/2018

Contents

List of figures

List of tables

Abstract

The purpose of this thesis is to provide a solution to support the implementation of a new language into a rule-based Spoken Dialogue System. Our approach is based on the hypothesis that machine translation can help to solve the problem of language portability for Spoken Dialogue System. Therefore, we translated a dialogue program with machine translation from English to Mandarin and evaluated its performance. The results show that this approach is efficient to build a multilingual dialogue system.

Keywords : Spoken Dialogue System, chatbot, localization, machine translation, language portability

Le but de cette étude est de fournir une solution pour implémenter une nouvelle langue dans un système de dialogue vocal basé sur des règles. Notre approche repose sur l'hypothèse selon laquelle la traduction automatique peut aider à résoudre le problème de la portabilité vers une nouvelle langue. Nous procédons à la traduction automatique du système de dialogue de l'anglais vers le mandarin et évaluons ses performances. Les résultats indiquent que cette approche est efficace pour concevoir un système de dialogue multilingue.

Mots-clés : dialogue vocal, chatbot, localisation, traduction automatique, portabilité linguistique

Acknowledgements

I would like to express my appreciation and thanks to some professors and professionals who led me through the writing of this thesis and my internship period.

My sincere thanks to my tutor, Damien Nouvel, for his patient guidance, encouragement and precious advice he has provided throughout my time as his student.

Especially, I would like to thank Sophie Rosset who kindly helped me collecting papers for the state of the art.

Heartfelt thanks go to all the professors in INALCO, from whose devoted teaching I have benefited a lot and academically prepared for the thesis.

My research would have been impossible without the aid and support of my supervisor, David Houssin who took time out to hear, guide and keep me on the correct path.

Last but not least, I would also like to thank my colleagues in the Dialog Team of Softbank Robotics, Yufo Fukuda for his careful professional guidance in every phase of my internship, Jessica Marthe-Rose and Ilmo Gourdin for their warm-heart encouragement and technical support.

INTRODUCTION

Dialogue system is a branch of Natural Language Processing intended to converse with a person with a coherent structure. Recently, it has generated huge interest due to the tremendous investment in applications handling with personal assistant services and company services such as customer support or automatized FAQS.

A Spoken Dialogue System is distinguished from a written text dialogue system by adding two major components : a speech recognizer and a text-to-speech module. There are mainly three types : rule-based models, retrieval-based models and generative-based models.

The research was performed in the Dialog team of Softbank Robotics, which is composed of computational linguists and software engineers who design and manage the conversation component of the robots.

This work aims to provide a solution for implementing a new language into a rule-based Spoken Dialogue System. Our approach is based on the hypothesis that machine translation can help to address the problem of language portability.

This thesis is organized as follows : An overview of the company and their Spoken Dialogue System is presented in chapter1. Chapter2 is devoted to study previous works and an existing tool of machine translation. Our implementation and evaluation will be introduced in chapter3 and chapter4. Base on the results of evaluation and the challenges occurred during our experiment, we will take a step back in chapter5 and discuss what issues could arise in future work.

I

Context

CHAPITRE 1

SPOKEN DIALOGUE SYSTEM OF A HUMANOID ROBOT

Contents

1.1 Introduction

This chapter is devoted to present the context of this work. We will start with a broad description of the company and the robots, and then a classification of Spoken Dialogue Systems in order to identify the class of robots' system, some basic properties of spoken dialogue will also be developed from a linguistic perspective.

An initial overview of robots' dialogue system is will be presented in three aspects : the architecture, syntax and linguistic resource, this overview will serve as a background for presenting the process of implementing a new language.

1.2 SoftBank Robotics

SoftBank Robotics (formerly Aldebaran Robotics), as the leader in the humanoid robotics market, provides interactive humanoid robots who assist professionals in the fields of education, research, health, distribution and tourism, as well as help families and private individuals.

Since Aldebaran Robotics was acquired in 2012 by the company SoftBank, a major mobile operator in Japan, Aldebaran Robotics has become the SoftBank Robotics which believes in a future where humanoid robots will assist humans in daily lives. SoftBank Robotics has developed three humanoid robots, Nao, Romeo and Pepper.

1.2.1 Robot products

Nao

Nao is the first autonomous, programmable humanoid robot developed by SoftBank Robotics. Approximately 9,000 Naos are currently in use worldwide with particular success in the fields of research and education. Nao is also utilized as an assistant by companies and health-care centers to welcome, inform and entertain visitors.

58cm in height, Nao is a robot with pleasantly rounded features : two 2D cameras to recognize shapes, objects and even people, 4 directional microphones and speakers to interact with humans, 7 touch sensors, sonars as well as an inertial unit to perceive his environment and locate himself in space.

FIGURE 1.1 – Nao

Pepper

Pepper is the first social humanoid robot with the ability to recognize faces and basic human emotions in the world. Pepper was optimized for human interaction and can interact with people via conversation and his touch screen.

Pepper is available today for businesses and schools. Over 2,000 companies worldwide have adopted Pepper as an assistant to welcome, inform and guide visitors in an innovative way.

120cm in height, Pepper is equipped with perception modules to interact with humans via touch sensors, LEDs and microphones, infrared sensors, bumpers and an inertial unit.

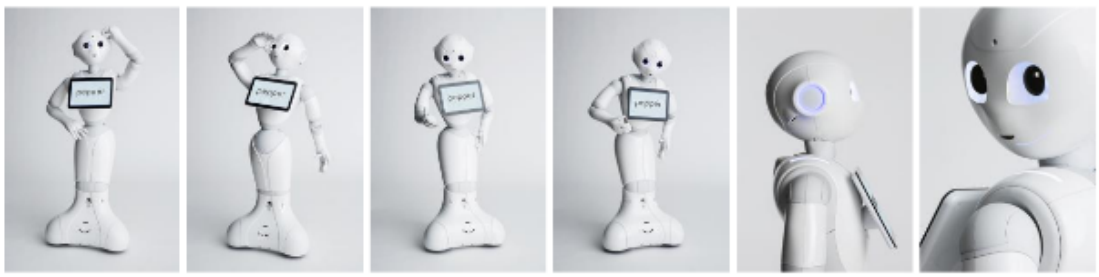

FIGURE 1.2 – Pepper

Romeo

Romeo is a research platform for personal assistance with the goal to intensify researches on assistance for people facing a loss of autonomy. The Romeo project is still at the research stage.

FIGURE 1.3 – Romeo

Robots use case

Their partner ecosystem enables the company to cover a variety of business relationships, with the aim to deliver the solutions that could best meet the specific objectives for customers. SoftBank Robotics designs and provides robots and solutions for :

• Sphere of business-to-business B2B services, which together refers to the activities aimed at business customers. SoftBank Robotics provides products or services for the commercial activities of companies such as Sephora, Renault, Carrefour...

• Sphere of B2C services, also called "Business to Customer", which refers to the activities that directly link the business directly to the end consumers. Pepper occupies for example a number of Japanese homes and Nao has been sold to many individuals.

• Sphere of B2E services or Business to Education which designates all the relations between the company and the educational administrations.

• Sphere of B2Dev or B2D services, Business to Developers, which together refers to the activities aimed at the clientele of developers, we will explore tools for developers in the next part.

SoftBank Robotics tools

To facilitate the interaction with robots, the company has developed its own operating system called Naoqi OS and Qisdk as well as several advanced programming software. These tools make it possible to create complicated behaviors, access data acquired by the robots' sensors, control robots and so on. Developers are able to choose the tools on the online store called the SoftBank Robotics Store including Choregraphe, Python SDK, C ++, JavaScript.

1.2.2 Interaction with a humanoid robot

The body shape of a humanoid robot is designed to resemble the human body for a range of functional purposes, such as facilitating interaction with both human and environments. Its physical presence could distinguish a humanoid robot from a basic conversational agent. As the humanoid robot is far more than a chatbot with a human body, this interaction human-robot will deliver a totally different experience.

Pepper has been designed in this direction, which has been demonstrated by the humanoid robot with lifelike eyes and a series of expressive gestures including speaking, moving, and interaction with people. Pepper features a variety of human-like characteristics such as recognition of emotions, understanding of the tones, and reactions to non-verbal signals such as smiles and frowns.

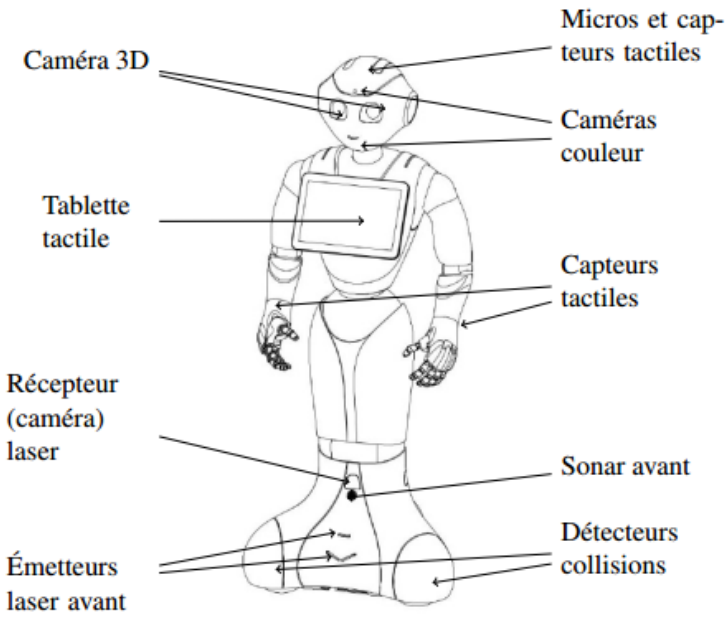

FIGURE 1.4 – Metric sensors of Pepper

Pepper is equipped with metric sensors as shown in the Figure 1.4, which is designed to detect obstacles around and trigger a move. In addition, there are more options in designing the conversation with interesting interactions.

Multiple tools are available to animate the interaction during a conversation :

• The animation library : a list of predefined animations usable in an application.

• Animated speech : this feature allows Pepper to automatically animate when speaking.

• Basic awareness : this feature makes Pepper aware of the surrounding environment.

1.3 Spoken Dialogue System of SoftBank Robotics

This part describes some of the basic attributes of spoken dialogue along with the techniques and issues associated with the development. We start with a classification of Spoken Dialogue Systems in order to narrow down the class of system that is referred to in this thesis.

1.3.1 Classification

Spoken dialog systems vary in their complexity. It was once categorized into two different parts : "simple" command-based systems associated with commercial systems and "complex" ones for the academic purposes which allow the users flexibility to some extent.

[Pieraccini and Huerta, 2005] pointed out that this distinction is partial as a system developed within the industry can also be complex. It must be compliant with the availability requirements and deal with "real" users to a large extent based on technical constraints. By contrast, for academic researchers studying the dialogue system, the objective is to explore how the system can handle both flexible and natural use of a certain language. At the same time, conducting large-scale usability studies and getting feedbacks collection from real users could bring difficulties to the researchers.

[SKANTZE, 2007] has proposed a distinction between the two types of dialogue systems ; we may refer to them as conversational systems and command-based systems. Table 1.1 in [SKANTZE, 2007] summarizes the distinction between these two systems which are prototypical class, and a given dialogue system does not have to exhibit all attributes.

	Command-based	Conversational
Metaphor	Voice interface metaphor.	Human metaphor.
Language	Constrained command-language.	Unconstrained spontaneous language.
Utterance length	Short utterances.	Mixed.
Semantics	Simple semantics. Less context dependence.	Complex semantics. More context dependence.
Syntax	More predictable.	Less predictable.
Language models	Strict grammar, possibly large vocabulary.	Less strict grammar, possibly smaller vocabulary.
Language coverage challenge	How to get the user to understand what could be said.	How to model everything that people say in the domain.

TABLE 1.1 – Classification of Spoken Dialogue System

People talk with a conversational system as their conversational partner, on the other side, a command-based system is predominantly considered as a "voice interface" with some options offered to activate devices with voice commands.

The challenge for a conversational system is how to predict everything that a user may say and how to model the language. To model more of the less predictable input,

a conversational system commonly keeps a flexible syntax and a limited vocabulary. For utterance length, there are long utterances with complex semantics which are mixed with shorter and context-dependent utterances.

In command-based systems, users are required to learn at first what can be said with the language models being often more stringent. The challenge may instead be how to handle a very large vocabulary.

Identification of the type of Spoken Dialogue System for SoftBank Robotics

In practice, human-robot interaction designed at SoftBank Robotics involves these two separate aspects, firstly, the robots have the functionality of conversational partner and companionship, secondly, there are also simple commands of a command-based system, such as "lower the volume" or "raise your arm". For these two aspects, linguists attempt to predict as much as possible utterances in order to deliver the users a more intuitive and efficient interaction.

1.3.2 Syntax of Qichat

Qichat[qic,] represents a combination of the natural language engine with the dialog management system designed originally for creating chatbots. To properly manage an interactive conversation, it's crucial to have an understanding to the main concepts below.

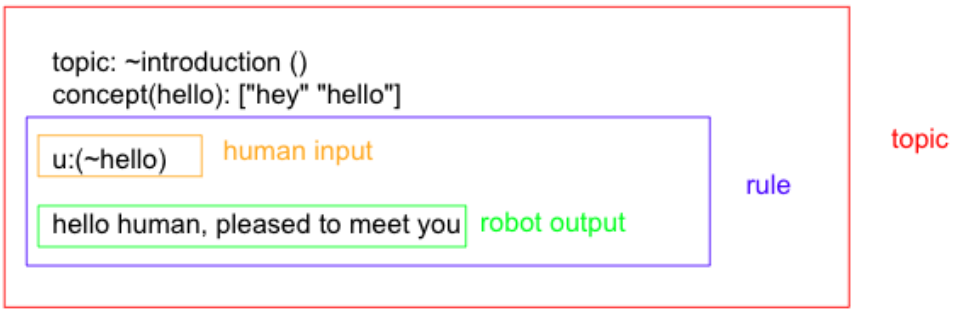

FIGURE 1.5 – Dialogue example

Topic

This keyword defines a topic for conversation. A topic is a script file in which rules are bundled in different categories, allowing for managing the dialogue system in a more efficient manner. The topic in the example of Figure 1.5 is called "introduction",

Rule

As the basic element of topic, a rule associates what a person says with a relevant answer of robot. Delimiters enable the creation of powerful rules and management in one line lots of different cases. The priority among different types of rules are detailed in chapter 1.3.3

Concept

A Concept is a list of words and/or phrases that refer to one idea. For example, a list of countries, a list of names, synonyms of a word.

Human input

Human input is viewed as part of a rule or a subrule delimited by parentheses and contains the message to be recognized by the robot.

When the human input matches, the message is recognized, which triggers the rule before the robot output being synthesized. In the mean time, the focus is set to the topic which contains the rules.

Delimiters

Syntax	Example
sentence delimiter : ""	"This is a sentence."
Choice :[ ]	[word1 word2 wordN]
Optional part :	{optionalpart}

TABLE 1.2 – Delimiters of Qichat

Delimiters allow to create a rich and dynamic conversation efficiently. Sentence delimiter "" allows to place a phrase instead of a single word in a choice [] or an optional part {}. For a human input, Choice [ ] creates one rule accepting variations, while in the case of a robot output where the rule is triggered several times, the words will be used sequentially in order to create varying responses. Optional part is a word or a sentence that could appear or not.

Dialogue example
u :([hello hi] how are you doing)
^ rand ["I am fine" "I am OK"]

Execution
> hi how are you doing
I am fine
> hello how are you
I am OK

1.3.3 Architecture of the Spoken Dialogue System

This section introduces the architecture of our Spoken Dialogue System. As presented by Figure 1.6, multiple components are consolidated in a single dialogue system, Speech-to-text, matching engine and Text-to-speech are indispensable components who need to work together for the system to function successfully.

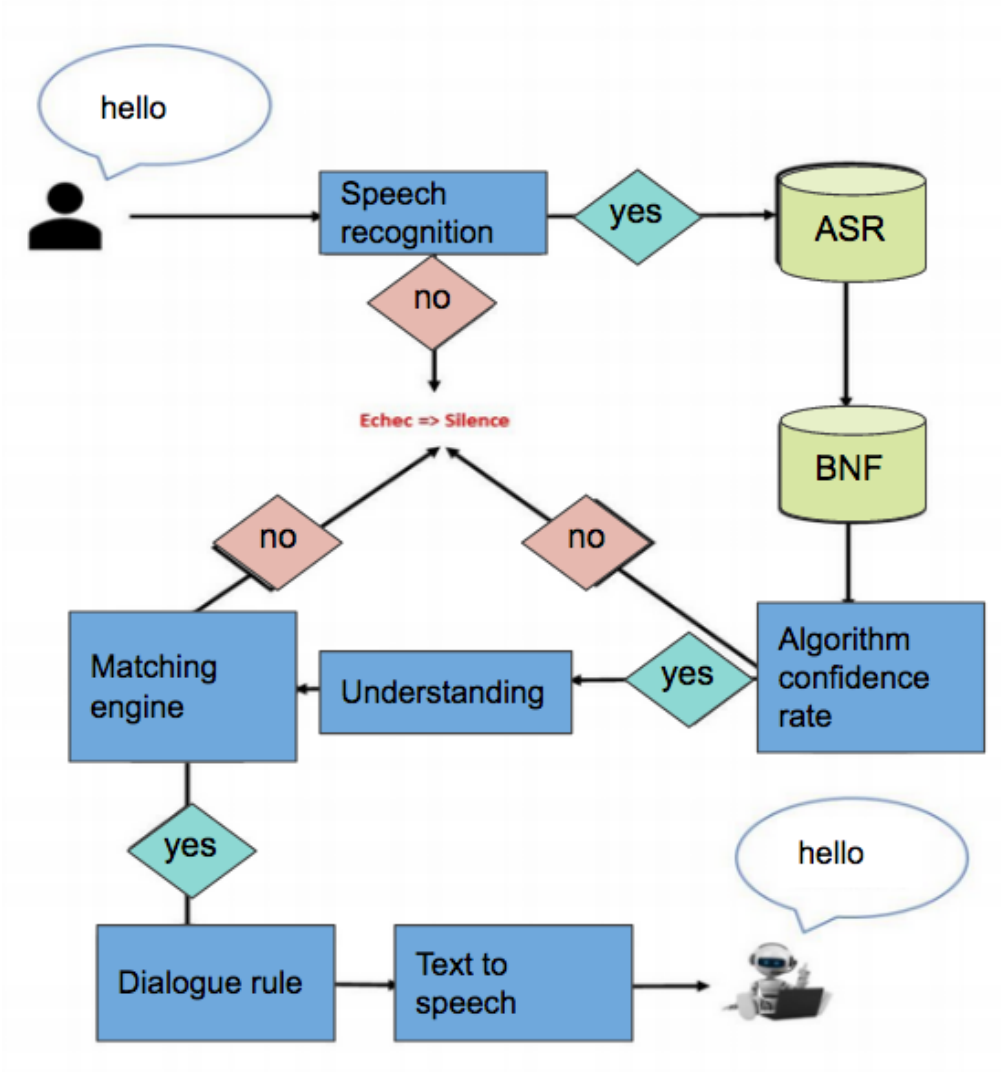

FIGURE 1.6 – Architecture of Spoken Dialogue System

Speech-to-text

First of all, a voice is processed by voice recognition to determine whether the voice has been heard or not. In case of failure, there will be no reaction, the robot remains silent. If the voice has been identified and processed by voice recognition, the signal will be transmitted to the ASR (Automatic Speech Recognition) before going through BNF (Backus–Naur form), a system that is employed to carry out checks on the syntax of the input with syntactic grammars.

Subsequently, confidence ratings will be established, where it scores higher than 50%, the input will be processed more smoothly to boost the rate of making response, otherwise the robot will remain silent, with no response being made.

We will discuss more specific difficulties in Mandarin in terms of ASR in Chapter 5.3.1.

Matching engine

A matching system applies to words from ASR and searches if the input corresponds to a certain predefined rule. Three principles are taken into account in this process.

Firstly, the difference of one or two tokens between predefined rule and a input is accepted by the matching engine which allows for a relative flexibility in creating the predefined rules. For instance, where "an apple" is defined as a rule, either "a red apple" or "a green apple" will be recognized as a match to the same predefined rule, so a response will be made with the same answer.

Secondly, when there are multiple rules that could match an input, the matching engine will automatically select the longest one.

For the same circumstance where input " a red apple" is recognized, in case that there are double rules of "a red apple" and "an apple", the input "a red apple" will be selected as a match to input "a red apple" instead of "an apple".

Thirdly, there is a priority among different types of rules, ranging from the highest priority level to the lowest we have subrule, standard rule, private rule and fallback rule.

Subrule

Subrule plays a crucial role in initiating multiple-round conversations. Users are allowed to make alteration to their intent on a continued basis during the conversation, with the context taken into account. In addition, Subrule enables maintaining a set of currently identified intents.

```
u :(How many songs can you sing?)
I can sing ten songs.
        u1 :(Which songs?)
        Do you really want me to list all the songs I know?

                u2 :(yes)
                ^ enumerate(~ songlist). That's all for now!

                u2 :(no)
                ^ rand["Alright then!" "Ok!"]

                u2 :(Just one example)
                For example Brother John
```

In this circumstance, "u1 :", "u2 :" are known as user subrules. The indentation underlines the relationship between a rule and its subrules. In case that when the first "u :" is matched by applying this rule, the subrule "u1 :" will be activated with the highest priority. The same principle still applies where the "u1 :" is matched, with "u2 :" having the highest priority.

Standard rule

A standard rule is of less priority than a subrule but prioritized over a private rule.

Private rule

A private rule is only activated at the time when its corresponding topic contains the last triggered rule.

topic : ~ film()
u :(Let's talk about films) Great!
u :^ private(What are we talking about) We are talking about films.

In this circumstance, there are two rules included in the same topic. When the input "Let's talk about films" is matched, the rule with "^ private" will be given the priority for matching.

However, if there has been no rules triggered in this topic, while at the same time there is a similar input "What are we talking about" being repeated in several topics, no priority will be given and the selection will be random.

Fallback rule

Fallback rules are rules contained in a topic with a "^ private" mark, they have the lowest priority in matching, and return usually a general response.

Text-to-speech

If a rule containing "Hello" as input exists, then the system of text-to-speech will be activated for the robot to make response with a predefined answer, for example "hello! ".

Pepper is available in three vocal styles : Neutral, Joyful, and Didactic. Neutral is selected for a majority of Latin languages, whereas the joyful style is preferred for Mandarin and Japanese to create a more natural voice.

Pepper features an unique voice that is supposed to be used in a consistent way as it voice is instantly recognizable. Therefore, its sound and tone are required to be consistent across all applications.

1.3.4 Linguistic resource analysis

Utterance

Before analyzing this linguistic resource, it is necessary to introduce some fundamental properties of oral conversation from a linguistic perspective.

The sentence is commonly used as a basic unit for analysis of written text. Sentences are delimited by punctuation marks, where each sentence could contain one or more propositions. In the opposite, spoken dialogue contains no punctuation marks and constitutes largely of fragmentary linguistic constructions.

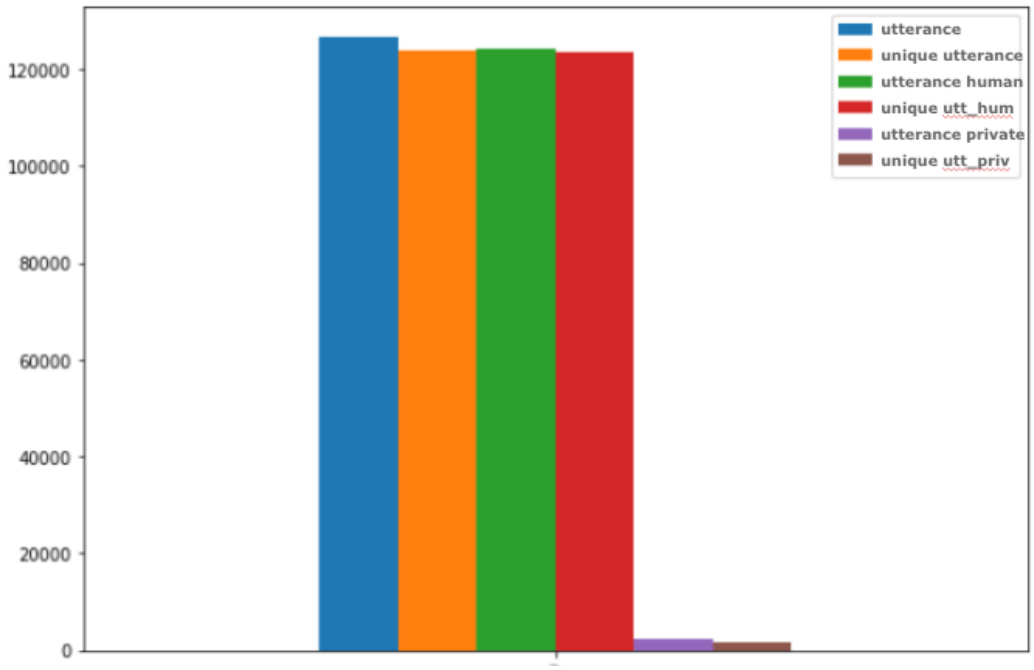

FIGURE 1.7 – Statistics quantity of utterances

In spoken language analysis, an utterance is the smallest unit of speech. As introduced by [Edlund et al., 2004], an utterance is a continuous piece of speech beginning and ending with a clear pause, it could also contain some para-linguistic features, such as facial expression, gesture, and posture.

In this thesis, the term utterance is utilized to refer to an uninterrupted sequence of speech from one human input or robot output.

Topics and rules

As illustrated in Figure 1.7 There are 92 topics containing 2,495 rules, with the overall number of utterances standing at 126,625, of which 123,943 utterances are unique. Among the 1,242,242 human input utterances, there are totally 123,943 being unique. In regard to private utterances, the number is 2,383, of which 1,532 considered unique. For human input, there are 2,459 standard input, 37 private input which is designed in a certain context, and that these rules are only active when its corresponding topic has the focus.

Distribution of utterances length

What noticeable from many of the dialogue blind tests is, with a robot being confronted, people tend to avoid using clauses with complex syntactic structures, and keep their utterances short and simple, which is understood to facilitate speech-to-text processing. As show in Figure 1.8 and Figure1.9, upon calculations of the utterances predefined as human input in English and in Mandarin, we found out that most sentences in English contain 4-7 tokens, whereas in Mandarin the figure is slightly smaller, coming to 3-7 tokens.

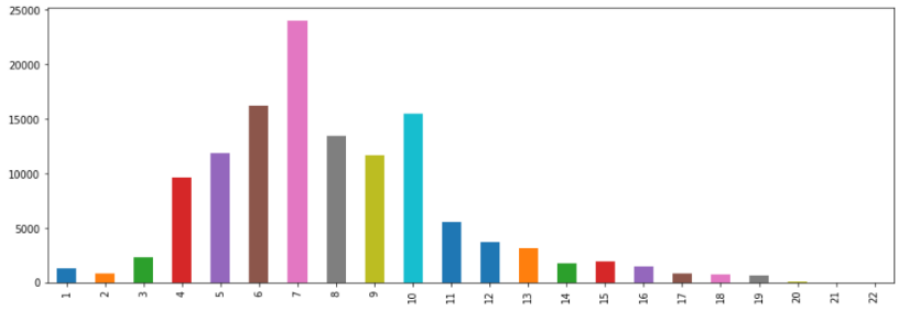

FIGURE 1.8 – Distribution of utterances length in English

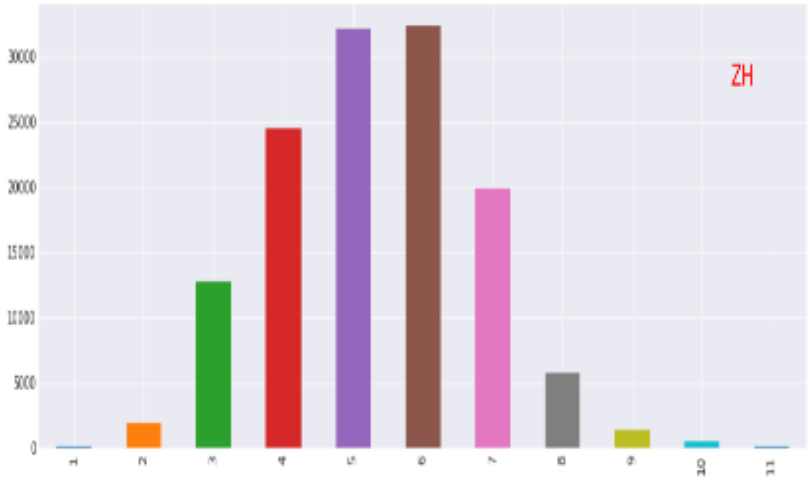

FIGURE 1.9 – Distribution of utterances length in Mandarin

Vocabulary

Vocabulary is an ultimate form of expression, which suggests that having an extensive set of vocabulary will enable a robot to express themselves explicitly to ensure a smooth communication. A linguistic vocabulary is identical to a thinking vocabulary, which indicates that we are able to have concise thoughts with precision. We have 8400 words in English vocabulary stored in robots, allowing them to speak in a clear, organized way on certain subjects. Nonetheless, it is still quite limited on the understanding of a wide range of long and complex texts, especially for some subtextual or stylistic nuances.

1.4 Conclusion

The context of this research has been presented in the first chapter. After a description of the company and the robots, we discussed some basic properties of spoken dialogue from a linguistic perspective. It is argued that two general types of dialogue systems may be distinguished, command-based and conversational. Furthermore our target system contains attributes of these two types.

We also explored a overview of robots' dialogue system in three aspects : the architecture, syntax and linguistic resource, this overview will help to understand the

process that need to be considered in the task of implementing a new language.

CHAPITRE 2

STATE-OF-THE-ART

Contents

2.1 Introduction

In this chapter, we will firstly study the previous works which address the problem of language portability in the related field. In the second section, we will explore the machine translation tool of Google, and investigate the gap between human translation and machine translation.

2.2 Related work

A massive challenge arising from building up a multilingual system is how to adapt an existing application system to a new language. Plenty of time and fund for setting up a spoken dialogue system are invested to create different rules which allow the system to process, comprehend and respond to the requests made by users.

As demonstrated in numerous researches like [Ravi and Artstein, 2016] [Wang and Seneff, 2006] [Suendermann et al., 2009a] [Jabaian et al., 2010] [Suendermann et al., 2009b] [Servan et al., 2010], the application of a machine translation system is beneficial in solving the problems associated with language portability by translating the existing source language to reduce the time and cost required for development.

These portability methods are generally developed based on statistical machine translation or automatic word alignment techniques. It has been revealed in the study [Suendermann et al., 2009a] that data can be translated automatically before a retrained stochastic grammar being applied to perform recognition. [Servan et al., 2010] proposes a solution of translating the entire training corpus into the target language via automatic translation.

Despite many of these studies being based on automated learning systems, they contribute to a reference for the rule-based dialogue systems, and demonstrate that

	PBMT	GNMT	Human
English → Spanish	4.885	5.428	5.504
English → French	4.932	5.295	5.496
English → Chinese	4.035	4.594	4.987
Spanish → English	4.872	5.187	5.372
French → English	5.046	5.343	5.404
Chinese → English	3.694	4.263	4.636

TABLE 2.1 – Mean of side-by-side scores on production data

when transferring a system from the original language to a new one, it is vitally important to maximize the linguistic resources acquired in the source language, for which machine translation is an efficient way of achieving dialogue system portability.

2.3 Gap between human and machine translation

Based on the conclusion of the last section, it would be interesting to apply the machine translation in terms of incorporating a new language, but it is necessary for us to conduct study on the machine translation performance.

Following the launch of Neural Machine Translation (GNMT) system in 2016, Google took a decision to test all the techniques that are critical to its accuracy, speed, and robustness. The data needed for the evaluation consisted of a total of 500 randomly sampled sentences from Wikipedia and news websites, with ratings being given with a score ranging from 0 to 6 for each translation.

All human raters were require to rate the translations by means of in a three-way side-by-side comparison. The three sides include the translations from the production phrase-based statistical translation system used by Google, the translations from their GNMT system, as well as the translations by individuals proficient in both languages.

According to the study [Wu et al., 2016], the mean value of side-by-side scores on production data results indicates that their model functions reasonably well regarding these major pairs of languages, especially from English to Spanish, where the rates of machine translation and human translation are very close : 5.428 and 5.504. Unfortunately, however, the program has also exhibited a high rate of translation errors with translation from English to Mandarin. Meanwhile, it's worth noting that this pair of language is also challenging for human beings to translate.

The Figure 2.1 in [Wu et al., 2016] shows that the distribution scores for 500 sentences translated from English to Spanish by Google, matched the conclusion of a earlier study [Aiken et al., 2009] in which compared the translations of German and Spanish text to English, they reported that people preferred the human translation for complex long sentences, but the machine translation could also have accurate results for short sentences.

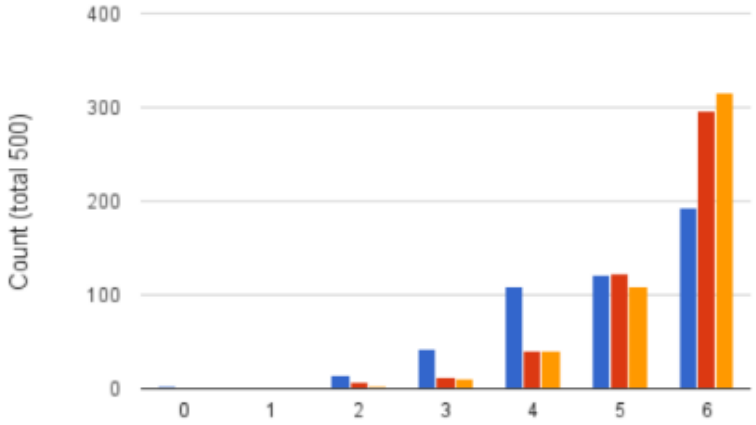

FIGURE 2.1 – Distribution scores (PBMT blue, GNMT red, Human orange)

The utterances for translation in this task tend to be relative short. As mentioned in 2.3 linguistic resource analysis, the average length is 4-7 words, meaning that complex understanding is not required here, and that the translation of these utterances could be appropriately dealt with by Google translation service.

2.4 Conclusion

In this chapter, we studied the previous works in the related field which demonstrated that machine translation could be an efficient approach for dialogue system portability. In the second section, an investigation of the gap between human translation and Google machine translation is presented, the result showed that the translation of simple utterances could be appropriately dealt with by Google translation service.

II

Experiments

CHAPITRE 3

METHODS

Contents

3.1 Introduction

This chapter details the implementation and tools required to progress the project. As indicated in the schema of Figure 3.1, our methodology is comprised of four different stages. At stage one, we put in place a clean text intended for machine translation, with all the sentences generated in the form of human input. Stage two and stage three focus on the use of Google translation API along with the tokenization and the normalization of the translated text. At the final stage, an algorithm is implemented on the translation of normalized text in order to establish usable and correct dialogue rules according to Qichat syntax in chapter1.3.2.

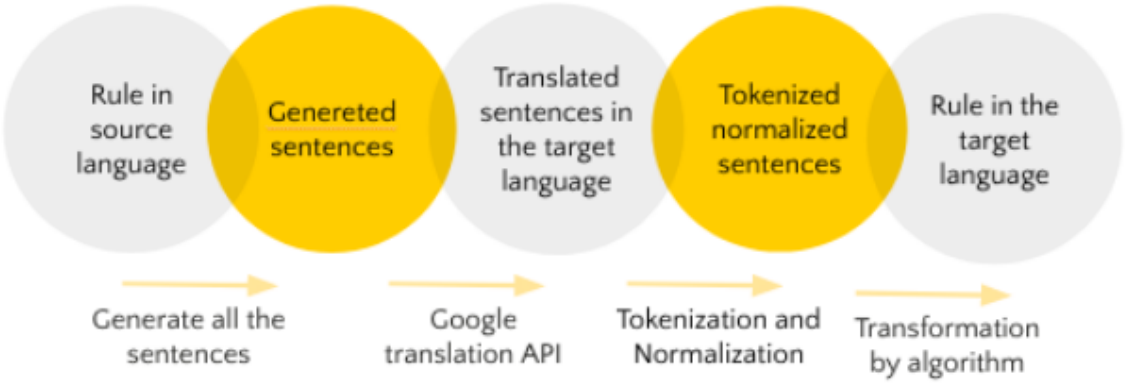

FIGURE 3.1 – Transformation shema

3.2 Stages of the implementation

3.2.1 Generation of all the possible utterances in the source language

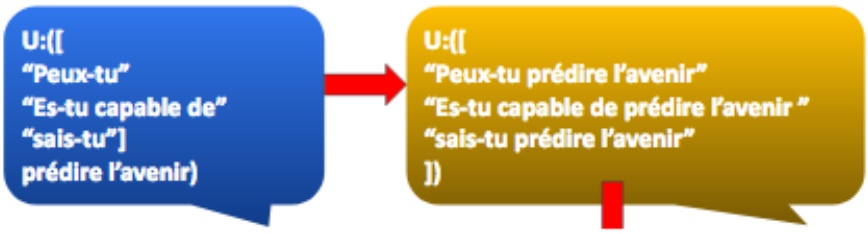

FIGURE 3.2 – Generation from rule to utterances

In this first stage, as illustrated in Figure 3.2, we generate all the possible combinations of human inputs with the application of Qisdk developed by SoftBank Robotics. Given the outputs of robots are normally presented in the form of complete sentences or words, most robot outputs don't need to be processed in this step.

This stage is indispensable for improving the performance on our methods, otherwise we will provide separated words for machine translation, indicating that a literal translation where source words are translated individually to target words.

3.2.2 Machine translation

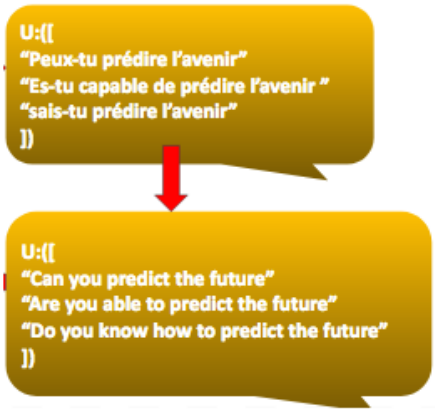

FIGURE 3.3 – Machine translation

API of Google translation is used in this step, as showed in Figure 3.3, with the part ["Peux-tu prédire l'avenir" "Es-tu capable de prédire l'avenir " "sais-tu prédire l'avenir"] being treated as the input, Google will factor in every signle utterance in "" to translate in a separate way, prior to displaying the result of translation in the target language with "".

3.2.3 Tokenization and normalization

The objective at this stage is to prepare a clean text for altering the rules, with the focus being placed on tokenization and normalization of the translated utterances.

In order to cut every utterance into an accurate word segmentation, we prefer to use some open sources tokenization tools : module NLTK selected for English and French, Jieba segmenter for Mandarin and Tiny segmenter for Japanese.

In the process of normalization, our primary intention is to normalize the punctuation marks. Because in the course of translation, problems are highly likely to arise from the different punctuation marks associated to the target language, for example some punctuation marks in Japanese and in Mandarin are not accepted by Qisdk, such as dot and comma.

3.2.4 Transformation into dialog rule

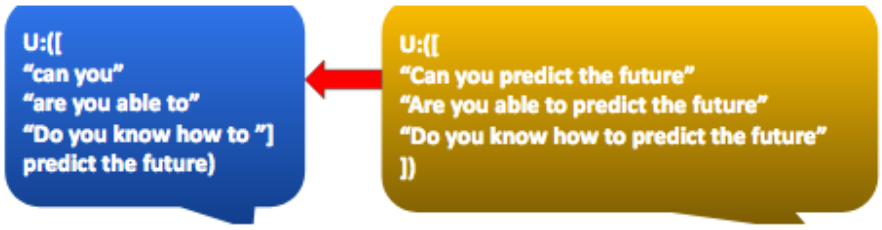

FIGURE 3.4 – Transformation from utterances to rules

This stage aims at transforming sentences into rules in the regular expressions of Qichat. As there are a total of 2000-3000 rules in the same system, we make effort to ensure that every rule is as concise as possible, otherwise the compilation will be long and a large amount of time is thus required to match every single sentence in each rule.

Challenges of the transformation

The main challenge of this transformation is how to satisfy at the same time three criterias :

— keep the rule as short as possible

— avoid to lose utterances in the transformation

— avoid to add other utterances in the transformation

These three criterias are conflicting. If more importance is attached to reducing the time for compilation, then we will have simpler shorter rules, but these rules accept more utterances than the translation result of the last step. Besides, the more we focus on the exactitude of as much information on these two forms, the more elements we will have to process in the dialogue rules, the longer these rules will be.

Transformation algorithm

An algorithm is developed for this transformation, five steps are proposed in this solution :

1. Identify the common parts shared by all the utterances.
2. Pick the longest "common" part which contains the biggest combination of tokens.
3. Split all those elements in three parts : "before", "common" and "after".
4. Keep the "common" part, and continue to parse the other two parts.
5. Repeat the step 4 until there is no "common" part.

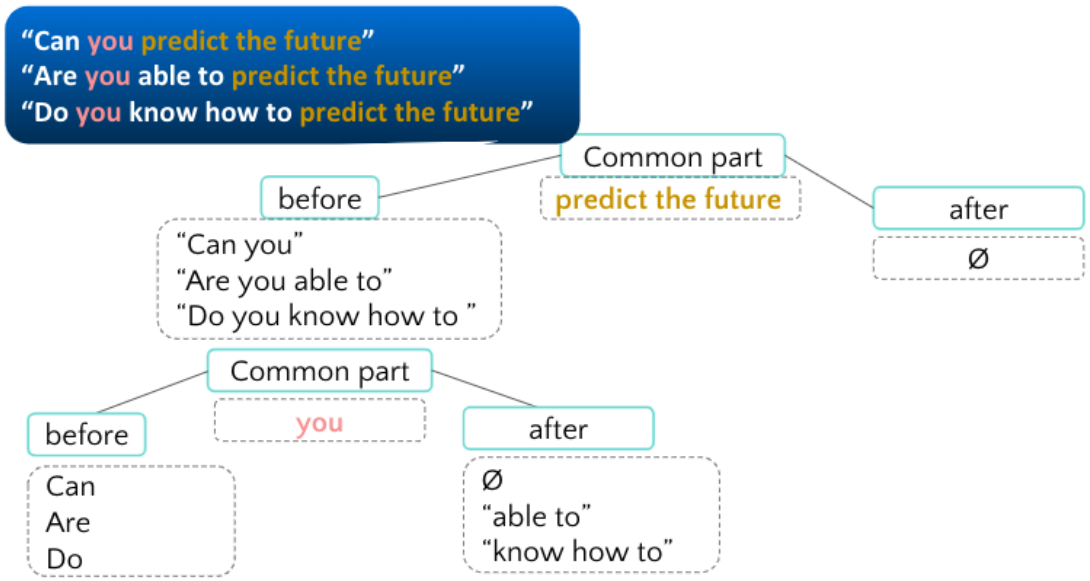

FIGURE 3.5 – Parsing example

For example, of the three utterances in Figure 3.5, we can identify two common parts "you" and "predict the future", as a result of which we pick "predict the future" for the longest as common part, before dividing into three parts : "before", "common" and "after". As far as this example is concerned, there is nothing in the "after" part, so we can carry on to parse "before" part, where we can still spot "you" as the "common" part. Subsequently, the same process is repeated by separating it into three different parts. Since there is no common part found in the "before" or "after" part, parsing is ended.

As showed in the last step of Figure 3.5, a "Ø" is written when nothing left in a utterance during the parsing. It can be discovered in this step that, we have three elements in the "before" part, with two in the "after" part, so we have a choice to make here.

Compilation time focused - Solution 1

With the focus put on the compilation time, we would prefer the solution 1 in Figure 3.6 which proposes to maintain the "after" part in an optional form, allowing us to have the shortest rule. In the mean time, it should be noted that there are some

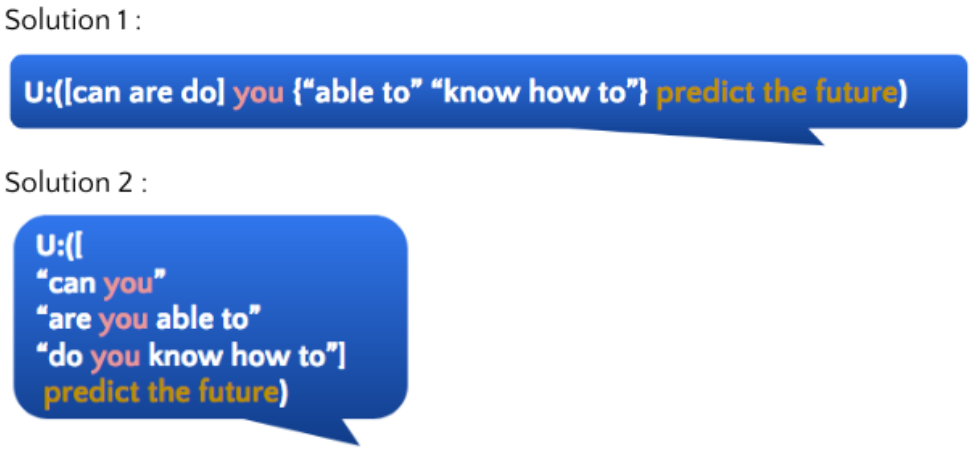

FIGURE 3.6 – Parsing solution 1

grammatically incorrect sentences generated, such as "can you able to predict the future", "are you know how to predict the future", which cause false positive without influencing true positive, and reduce the recall result.

Utterance results focused - Solution 2

On the flip side, when the situation arise where the two parts, "before" and "after" don't contain the same number of elements, parsing would be stopped, then we will switch to solution 2 in Figure 3.6, which ensures exactly the same result with the last step of translation. The downside of this is that the compilation time will be longer.

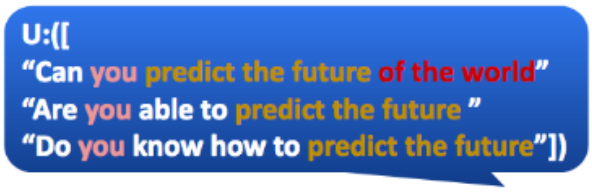

FIGURE 3.7 – Parsing solution2

Advanced Solution 2

What if the "after" part is not empty as the example of Figure3.7? Actually, the strategy of choosing the longest common part does not necessarily work well in each case for solution 2. We can see in the Figure 3.8, there are words in the "after" with the same problem of certain utterances having more word than others, it means that the program will return this rule without any transformation.

In this case, we propose an additional process for solution 2, abandoning the last token of the "common" part with more than one token. In contrast, if there is only one token, the parsing is stopped.

For the same example in the Figure 3.8, once there is a "Ø", we go to check the number of token of "common" at the same level. For the "Ø" at level2, since there is only one token "you" in "common", we go back to the level1 and keep the form "before"at level 1. For the "Ø" at level1, there are more than one word in the "common", so we move the last token of the "common" part "future" to the "after" part, the result of parsing is showed in Figure 3.9 and the rule in Figure 3.10.

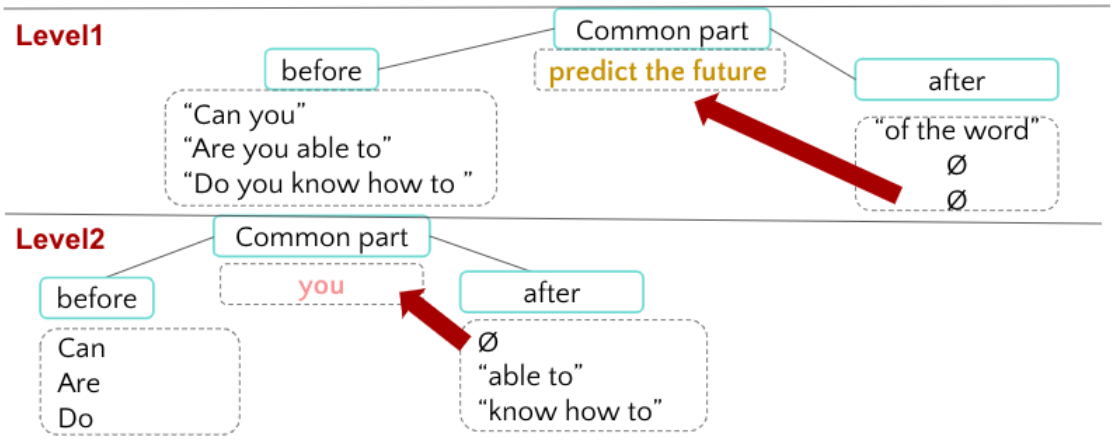

FIGURE 3.8 – Parsing of advanced solution 2

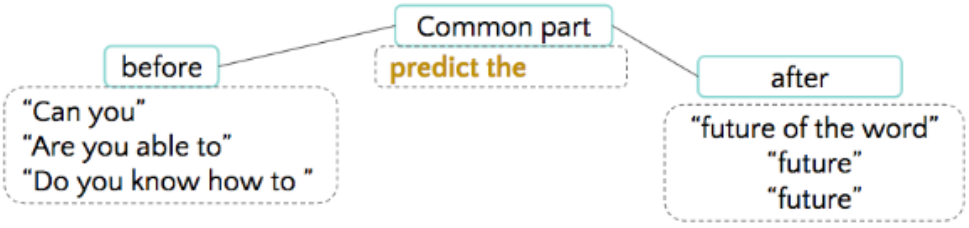

FIGURE 3.9 – Parsing result of advanced solution 2

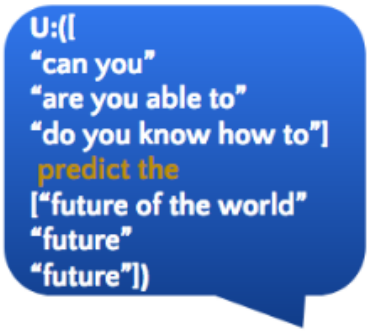

FIGURE 3.10 – Rule result of advanced solution 2

Conclusion of the transformation algorithm

In this part, an algorithm with different focus is presented for the transformation from a collection of utterances to a dialogue rule, the main challenge is how to maintain the rules as short as possible while keeping the same sentences without too much noise.

In practice, it is hard to be compliant with all the critarias, so we developed two solutions : the solution 1 focusing on a fast compilation and solution2 focus on a exactitude of content. The former allows to generate short rules, but it could also causes some false positives. The latter is more complex and accurate, in the same time rules generated in this solution could be longer.

3.3 Conclusion

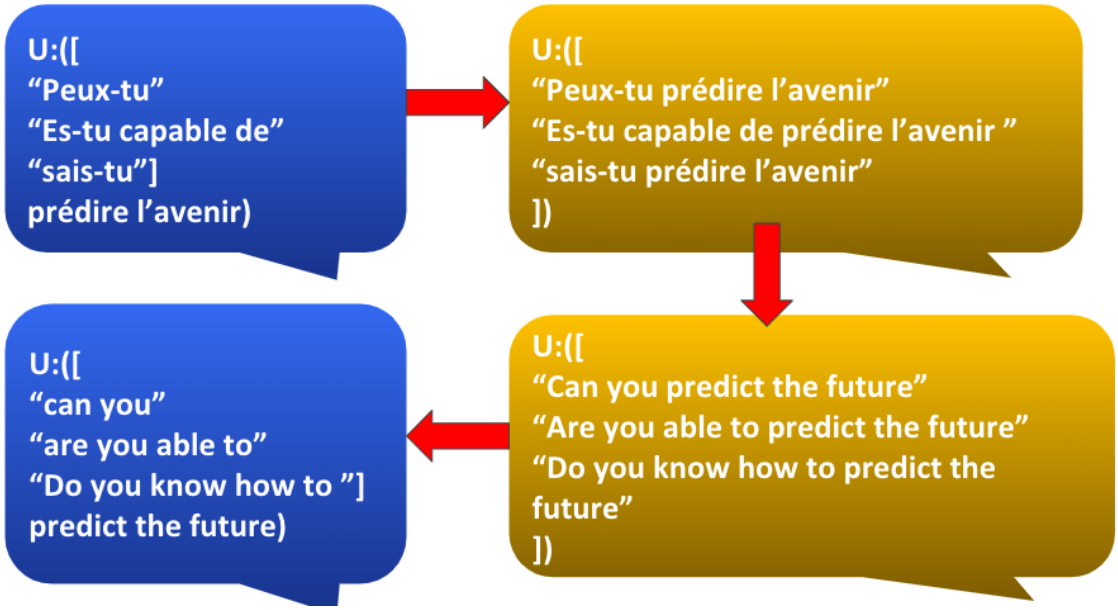

FIGURE 3.11 – Transformation example schema

This chapter presented the implementation and tools required to progress the project. The Figure 3.11 provide a overview with the example used in this chapter. Firstly, we put in place a clean text intended for machine translation, with all the sentences generated in the form of human input. Then we focus on the use of Google translation API along with the tokenization and the normalization of the translated text. At the final stage, an algorithm is implemented on the translation of normalized text in order to establish usable and correct dialogue rules.

CHAPITRE 4

EVALUATION

Contents

4.1 Introduction

A dialogue system can be evaluated on different aspects, such as performance, humanity, effects and accessibility, but our objective here is to evaluate the difference in implementing a new language into a dialogue system by manual and by automatic localization via the same matching engine, so our focus is on the consistency of these two results.

4.2 Corpus

The test program is processed with 100 rules containing 2,041 utterances as human input. These rules are translated directly by Google from English to Mandarin and refined according to Qichat syntax in 1.3.2 with solution2 in 3.2.4. It is worth stressing that no manual intervention for the content has been made after the translation for this test.

The request inputs consist of 84 utterances of 14 rules translated manually from English to Mandarin, these 14 rules are chosen from different linguistic perspectives, as illustrated in Table4.1, we picked one utterance of each rule to present the content.

4.3 Evaluation measure

Our evaluation task is a multi-class classification task, every rule of which can be considered as a class, and the request questions are to be assigned to these 14 rules.

Different evaluation measures are available when computing evaluation scores for classification. Recall, precision and F-score specified in [Van Rijsbergen, 1979] are often used for information retrieval purposes.

Given Table 4.2, we can calculate Recall(R) and Precision(P)

Rule	Focus of test
What is your name	Simple question
What is your current volume	Variety of sentence pattern "what is"
What time is it	Simple question
Can you open your left hand	Variety of sentence pattern "can you"
Speak softer	Variety of command pattern
Do you have a sense of humor	Variety of sentence pattern "do you"
Can you predict the future	Variety of sentence pattern "can you"
Who named Nao	Name entity
Do you know Isaac Asimov's laws of robotics	Foreign name entity
Is my girl friend going back to me	Future tense
I don't want to be your friends anymore	Negative sentence
Let's be friends	Imperative
How long does your battery last	Variety of sentence pattern "how long...last"
Talk about SoftBank	Organization entity

TABLE 4.1 – Rules and their test focus

	Predicted label X	Predicted not X
True label X	true positives (TP)	false negatives (FN)
True not X	false positives (FP)	true negatives (TN)

TABLE 4.2 – Performance table for instances labeled with a class label X

$$\text{Recall} = \frac{\text{true positives}}{\text{true Positive + false negative}} \tag{4.1}$$

$$\text{Precision} = \frac{\text{true positives}}{\text{True Positive + False Positive}} \tag{4.2}$$

$$\text{F-score} = \frac{(1+\beta^2) \times \text{ precision } \times \text{ recall}}{\beta^2 \times \text{ precision } + \text{ recall}} \tag{4.3}$$

However, recall, precision and F-score are initially designed to evaluate binary classification systems, such as search engines. Accuracy can be used to evaluate a set of predicted labels or performance of machine learning models, it is calculated as the portion of true labeled instances to total number of instances, but it is not the best choice when the dataset is unbalanced.

As pointed out by [Van Asch, 2013], when multiple class labels are to be retrieved, averaging the evaluation measures can provide a view on the general results. Two notions are available to refer to averaged results : micro-averaged and macro-averaged results.

$$\text{Micro-recall} = \frac{\sum_{k=1}^{n} \text{true positive}_k}{\sum_{k=1}^{n} \text{true positives}_k + \sum_{k=1}^{n} \text{false negative}_k} \tag{4.4}$$

$$\text{Micro-Precision} = \frac{\sum_{k=1}^{n} \text{true positives}_k}{\sum_{k=1}^{n} \text{true positives}_k + \sum_{k=1}^{n} \text{false positives}_k} \tag{4.5}$$

$$\text{Macro-recall} = \frac{\sum_{k=1}^{n}\left(\frac{\text{true positives}_k}{\text{true positives}_k + \text{false negative}_k}\right)}{n} \tag{4.6}$$

$$\text{Macro-precision} = \frac{\sum_{k=1}^{n}\left(\frac{\text{true positives}_k}{\text{true positives}_k + \text{false positives}_k}\right)}{n} \tag{4.7}$$

Micro average calculates sums of TP, FP, and FN across all classes, and F-score is based on these values. On the other hand, macro average calculates precision and recall for each class separately followed by the calculation of the mean precision and recall with F-score based on the mean values.

Compared with macro averaging, micro average takes the frequency of each label into consideration. There is a long debate among experts on choosing which method to use. However, it is commonly believed that macro-averaging can be a bad practice in cases that there is a considerable difference in number of samples of each class label.

4.4 Results

Rule	Predicted	Requested	TP	P	R
What is your name	5	5	5	1	1
What is your current volume	3	3	3	1	1
What time is it	5	5	5	1	1
Can you open your left hand	4	5	4	1	0,80
Speak softer	4	6	4	1	0,67
Do you have a sense of humor	3	3	3	1	1
Can you predict the future	5	8	4	0,80	0,50
Who named Nao	3	5	2	0,67	0,40
Do you know Isaac Asimov's laws of robotics	3	8	3	1	0,38
Is my girl friend going back to me	5	8	5	1	0,63
I don't want to be your friends anymore	5	8	4	0,80	0,50
Let's be friends	5	8	5	1	0,63
How long does your battery last	6	7	6	1	0,86
Talk about SoftBank	5	5	5	1	1

TABLE 4.3 – Experiment results

Among these 84 utterances, 58 matched correctly, 23 didn't trigger any rule, 3 are matched by another rule instead of the predefined rules. We have a macro precision of 0.95, macro recall of 0.74, macro f-score of 0.83, a micro precision of 0.95, micro recall of 0.69, micro f-score of 0.80.

Considering the presence of a class imbalance of our data structure, we prefer to choose micro-averaging as our evaluation measure.

4.5 Conclusion

In this chapter, we evaluated 84 utterances of 14 rules which are translated manually from English to Mandarin, these rules were chosen from different linguistic perspectives. We took this evaluation as a multi-class classification task, considering the presence of a class imbalance of our data structure, micro-averaging is chosen as the evaluation measure of which the micro f-score is 0.80.

Moreover, this experimental results of the rule "Do you know Isaac Asimov's laws of robotics" also shows that, the identification and translation of foreign name entity are difficult to dealt with in the context of Spoken Dialogue System, and it affects directly the result of the implementation, we will discuss more about the entity issues in chapter 5.3.2.

CHAPITRE 5

DISCUSSION

Contents

5.1 Introduction

Base on the results of evaluation in chapter 4 and the challenges occurred during our experiment, we will take a step back in this chapter and discuss what issues could arise in future work.

5.2 Quality check and localization

A high-quality content without misunderstandings or mistranslations must be ensured during the translation process, if we want to use this method in the industry. This step is very important as it will ensure that end-users receive a conceptually accurate local language version.

It is worth to emphasize that there is a difference between the localization and the translation. As defined in [Anastasiou and Schäler, 2010], "the localization is the provision of services and technologies for the management of multilingualism across the global information flow". As a process of modifying a product for a specific region, the aim of localization should be that people from a specified locale can use the product easily in their own culture and language. Therefore, the quality check can be considered as a process of localization, this step helps to make up the deficiencies of culture adaptation in machine translation.

However, is it possible that the quality check takes more time than pure manual localization? A test was thus carried out to answer this question. 100 rules of different lengths were localized respectively by two methods, with the first 50 rules purely manual by a linguist, and the second method being applied for the other 50 rules. After the automatic treatment, these rules have been checked and corrected by the same linguist who is familiar with this localization task. Finally, the result of shows that, for the localization of a dialogue rule, the manual method averagely takes 5 to 6 mins, while takes 3 mins for this automated process and the quality check.

Our experimental results of quality check are limited in the localization from English to Chinese, besides, the localized dialogue rules are relatively simple. We can use more complicated rules and try more languages in future experiments to verify this method in different perspectives.

5.3 Specific difficulties of localization into Mandarin

5.3.1 Automatic Speech Recognition

[Hwang et al., 2009] proved that core technologies developed for English ASR are applicable to a new language such as Mandarin. However, to achieve the best performance, language-dependent components are needed to be added.

For Mandarin, the challenge consist of the extraction of tone-related features. Four tones in Mandarin who can also change with different combination, for example, when there are two third tone together, the first becomes second tone.

The optimization of accents in different regions is also of great challenge, most Mandarin speakers learned Mandarin as the second language, and their pronunciations are strongly influenced by their native regional language. As a result, ASR systems implemented for processing standard mandarin, perform poorly for non-native accented speech, especially when there are multiple accents.

In order to differentiate between multiple accents, different sets of accent specific units in the futhur workshould be generated individually based on data-driven method.

5.3.2 Entity detection

The entity detection in Mandarin can be complicated for those linguistic features. Firstly, a person name consisting of a family name and a given name is of the pattern family name and given name, and both are one or two characters long. Secondly, there is no patterns for Location names. Usually being named entities, organization names are even more difficult to identify.

The most challenging part is the recognition of foreign names which are usually transliterated using Mandarin character strings whose sequential pronunciation mimics the source language pronunciation of the name. With effectively unlimited original pronunciation, they can be of any length.

The study of [Wang et al., 2000] shows that improving vocabulary coverage may require study of the transliteration process for foreign names. The translation of phonetic sequences in foreign languages is usually followed by patterns. For example, "ton" in "Houston", "Boston" and "Washington" are all translate as "dun4" with the same pronunciation. This method shared could be reused and help to define dialogue rules containing foreign entity.

Conclusion

This work attempted to address the problem of implementing a new language into a Spoken Dialogue System. We discussed some basic properties of our target system and the techniques and issues involved in translation of Spoken Dialogue Systems.

Firstly, we presented some basic properties of spoken dialogue from a linguistic perspective, It is argued that two general types of systems should be distinguished, command-based and conversational. Our target system contains attributes of these two systems. We explored an overview of robots' dialogue system in three aspects : the architecture, syntax and linguistic resource.

The previous works in the related fields demonstrated that machine translation could be an efficient approach for this purpose. Furthermore, an investigation showed that the translation of simple utterances could be appropriately dealt with by machine translation. This global view of the context serves as a background for our experiment. Based on the existing English system, we implemented Mandarin with the machine translation and the algorithm developed for this project.

This experiment shows that machine translation helps to solve language portability problem for a rule-based dialogue system with a micro f-score of 0.80, and that translated dialogue system can be an initial step towards tailoring a system to a new population.

References

[qic,] http://doc.aldebaran.com/2-5/naoqi/interaction/dialog/dialog-syntax_full.html. – Cité page 20.

[Aiken et al., 2009] Aiken, M., Ghosh, K., Wee, J., and Vanjani, M. (2009). An evaluation of the accuracy of online translation systems. *Communications of the IIMA*, 9(4):6. – Cité page 30.

[Anastasiou and Schäler, 2010] Anastasiou, D. and Schäler, R. (2010). Translating vital information: Localisation, internationalisation, and globalisation. *Syn-thèses Journal*, 3:11–25. – Cité page 47.

[Edlund et al., 2004] Edlund, J., Skantze, G., and Carlson, R. (2004). Higgins-a spoken dialogue system for investigating error handling techniques. In *Eighth International Conference on Spoken Language Processing*. – Cité page 25.

[Hwang et al., 2009] Hwang, M.-Y., Peng, G., Ostendorf, M., Wang, W., Faria, A., and Heidel, A. (2009). Building a highly accurate mandarin speech recognizer with language-independent technologies and language-dependent modules. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1253–1262. – Cité page 48.

[Jabaian et al., 2010] Jabaian, B., Besacier, L., and Lefèvre, F. (2010). Investigating multiple approaches for slu portability to a new language. In *Eleventh Annual Conference of the International Speech Communication Association*. – Cité page 29.

[Pieraccini and Huerta, 2005] Pieraccini, R. and Huerta, J. (2005). Where do we go from here? research and commercial spoken dialog systems. In *6th SIGdial Workshop on Discourse and Dialogue*. – Cité page 19.

[Ravi and Artstein, 2016] Ravi, S. and Artstein, R. (2016). Language portability for dialogue systems: Translating a question-answering system from english into tamil. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 111–116. – Cité page 29.

[Servan et al., 2010] Servan, C., Camelin, N., Raymond, C., Béchet, F., and De Mori, R. (2010). On the use of machine translation for spoken language understanding portability. In *ICASSP*, pages 5330–5333. – Cité page 29.

[SKANTZE, 2007] SKANTZE, G. (2007). Error handling in spoken dialogue systems. – Cité page 19.

[Suendermann et al., 2009a] Suendermann, D., Evanini, K., Liscombe, J., Hunter, P., Dayanidhi, K., and Pieraccini, R. (2009a). From rule-based to statistical grammars: Continuous improvement of large-scale spoken dialog systems. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4713–4716. IEEE. – Cité page 29.

[Suendermann et al., 2009b] Suendermann, D., Liscombe, J., Dayanidhi, K., and Pieraccini, R. (2009b). Localization of speech recognition in spoken dialog systems: How machine translation can make our lives easier. In *Tenth Annual Conference of the International Speech Communication Association*. – Cité page 29.

[Van Asch, 2013] Van Asch, V. (2013). Macro-and micro-averaged evaluation measures [[basic draft]]. *Belgium: CLiPS*. – Cité page 44.

[Van Rijsbergen, 1979] Van Rijsbergen, C. (1979). Information retrieval. dept. of computer science, university of glasgow. *URL: citeseer. ist. psu. edu/vanrijsbergen79information. html*, 14. – Cité page 43.

[Wang et al., 2000] Wang, C., Cyphers, D. S., Mou, X., Polifroni, J., Seneff, S., Yi, J., and Zue, V. (2000). Muxing: A telephone-access mandarin conversational system. In *Sixth International Conference on Spoken Language Processing*. – Cité page 48.

[Wang and Seneff, 2006] Wang, C. and Seneff, S. (2006). High-quality speech translation in the flight domain. In *Ninth International Conference on Spoken Language Processing*. – Cité page 29.

[Wu et al., 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*. – Cité page 30.