International Telecommunication Union

ITU-T F.746.5

(12/2017)

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

SERIES F: NON-TELEPHONE TELECOMMUNICATION SERVICES

Multimedia services

Framework for a language learning system based on speech and natural language processing (NLP) technology

Recommendation ITU-T F.746.5

ITU-T F-SERIES RECOMMENDATIONS

NON-TELEPHONE TELECOMMUNICATION SERVICES

TELEGRAPH SERVICE	
Operating methods for the international public telegram service	F.1–F.19
The gentex network	F.20–F.29
Message switching	F.30–F.39
The international telemessage service	F.40–F.58
The international telex service	F.59–F.89
Statistics and publications on international telegraph services	F.90–F.99
Scheduled and leased communication services	F.100–F.104
Phototelegraph service	F.105–F.109
MOBILE SERVICE	
Mobile services and multidestination satellite services	F.110–F.159
TELEMATIC SERVICES	
Public facsimile service	F.160–F.199
Teletex service	F.200–F.299
Videotex service	F.300–F.349
General provisions for telematic services	F.350–F.399
MESSAGE HANDLING SERVICES	F.400–F.499
DIRECTORY SERVICES	F.500–F.549
DOCUMENT COMMUNICATION	
Document communication	F.550–F.579
Programming communication interfaces	F.580–F.599
DATA TRANSMISSION SERVICES	F.600–F.699
MULTIMEDIA SERVICES	**F.700–F.799**
ISDN SERVICES	F.800–F.849
UNIVERSAL PERSONAL TELECOMMUNICATION	F.850–F.899
ACCESSIBILITY AND HUMAN FACTORS	F.900–F.999

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T F.746.5

Framework for a language learning system based on speech and natural language processing (NLP) technology

Summary

Recommendation ITU-T F.746.5 describes functional requirements and detailed functions for the framework of a language learning system based on speech and natural language processing technology. It provides a framework for a language learning system that will serve as a reference framework for language learning systems to be developed and used as low cost tools in many educational situations. This Recommendation defines the features, general requirements and functionality to support the language learning system based on speech and natural language processing (NLP) technology. The scope covers a high-level description of architecture, terminals, servers, interface and clients. .

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T F.746.5	2017-12-14	16	11.1002/1000/13427

Keywords

Language learning, natural language processing, speech interface, speech recognition.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T F.746.5

Framework for a language learning system based on speech and natural language processing (NLP) technology

1 Scope

This Recommendation presents an overview of the framework for a language learning system based on speech and natural language processing (NLP) technology. It describes the features, general requirements and functionality, which is a framework to support language-learning systems. The scope covers a high-level description of architecture, devices, servers and clients.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this document. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this document are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T F.746.3] Recommendation ITU-T F.746.3 (2015), *Intelligent question answering service framework*.

[ITU-T H.703] Recommendation ITU-T H.703 (2016), *Enhanced user interface framework for IPTV terminal devices*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 co-reference resolution [ITU-T F.746.3]: A function that detects the preceding referents of the pronouns which replace the noun phrases of the input sentences.

3.1.2 named entity recognition [ITU-T F.746.3]: A function that recognizes named entities such as PLO which are people, locations and organizations from the sentences. The PLO can be decomposed into more specific named entities depending on the applications.

3.1.3 natural language processing [ITU-T F.746.3]: A method that analyses text in natural languages through several processes such as part-of- speech recognition, syntactic analysis and semantic analysis.

3.1.4 part-of-speech recognition [ITU-T F.746.3]: A function that recognizes parts of speech (POS) in the sentences and assigns relevant POS tags considering contextual meaning of the target sentences.

3.1.5 semantic analysis [ITU-T F.746.3]: A function that recognizes the semantic relations among the words around predicates that exist in the same sentence. The semantic analysis function then generates a semantic predicate-argument structure (PAS).

3.1.6 speech [ITU-T H.703]: Speech is the vocalized form of human communication.

3.1.7 speech recognition [ITU-T H.703]: A kind of user interface for translation of spoken words into text.

3.1.8 syntactic analysis [ITU-T F.746.3]: A function that analyses sentence structures and generates dependency relation among words based on dependency grammars.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 dialogue-based speech interface: An interface based on speech, especially dialogues between the user and the device or system.

3.2.2 dialogue act: The user's intention or purpose of the utterances in a dialogue. Example: request for information, command for action, agreement.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

DA Dialogue Act

DNN Deep Neural Network

HCI Human Computer Interaction

ICT Information and Communication Technology

IT Information Technology

LLS Language Learning System Based on Speech/NLP Technology

NE Named Entity

NLP Natural Language Processing

PC Personal Computer

POS Part of Speech

SMS Short Message Service

SVM Support Vector Machines

TTS Text to Speech

5 Conventions

The following conventions are used in this Recommendation:

– The keywords "is required to" indicate a requirement which must be strictly followed and from which no deviation is permitted, if conformance to this Recommendation is to be claimed.

– The keywords "is prohibited from" indicate a requirement which must be strictly prohibited, if conformance to this Recommendation is to be claimed.

– The keywords "is recommended" indicate a requirement which is recommended but which is not absolutely required. Thus, this requirement need not be present to claim conformance.

– The keywords "is not recommended" indicate a requirement which is not recommended but which is not specifically prohibited. Thus, conformance with this Recommendation can still be claimed even if this requirement is present.

– The keywords "can optionally" indicate an optional requirement which is permissible, without implying any sense of being recommended. This term is not intended to imply that the vendor's implementation must provide the option and the feature can be optionally enabled by the network operator/service provider. Rather, it means the vendor may optionally provide the feature and still claim conformance with this Recommendation.

6 Introduction

Language learning requires a lot of time and cost for individual learners who wish to learn foreign languages, especially when personalized education is needed for each learner with a different level of learning capabilities. The ratio of students to teachers should be minimized to have an effective learning situation for individualized learning. Advances in information and communication technology (ICT), especially those in speech/language areas have made the language learning experience less expensive and more effective for individualized learning.

Speech interface and natural language processing (NLP) technology is an advanced technology that allows a computer to understand what a person says, and facilitates exchange of information in a smooth conversation with the person. Speech interface/NLP technology will likely be combined with other technologies in many application areas such as national defence, medical services, etc. Moreover, speech interface/NLP technology can also be applied to language learning systems for conversation training.

NLP is a core technology for human computer interaction (HCI) and a basis for knowledge and information services. It combines syntactic analysis and semantic analysis to understand human languages. NLP technology is used for understanding a user's speech in dialogue practices in various language learning scenarios.

More advanced dialogue-based speech interface technology is also being developed. The basic dialogue processing flow is shown in Figure 3. The technology recognizes and understands users' speech and generates appropriate responses in limited dialogue situations. The core dialogue processing technology is applied to foreign language learning systems that simulate one-on-one conversation training.

This Recommendation provides requirements, architecture and functions for a language learning system based on speech/NLP technology.

6.1 Basic concept of speech interface and NLP technology

Speech interface and NLP technology are the next generation interface that allows a computer to understand what a person says and also facilitates the exchange of information in a smooth conversation with the person. Figure 1 shows a basic speech interface and Figure 2 shows some basic NLP modules. Recently, speech interface has become one of the essential elements in the information technology (IT) industry such as intelligent robots, telematics, the next generation personal computer (PC), and digital-home. Speech interface/NLP technology will likely be combined with other technologies in many application areas such as national defence, medical services, etc. Henceforth, speech interface/NLP technology will grow into a multimodal interface with a united input device such as voice, pen, mouse and gestures. A high-level interactive speech interface will soon be attainable by incorporating circumstantial information and the speaker's intention. Moreover, speech interface/NLP technology plays the role of a core technology for mobile web information services, and can also be applied to language learning systems for conversation training.

NLP is a core technology for HCI and the basis for knowledge and information services. It combines syntactic analysis and semantic analysis to understand human languages. NLP technology is used for understanding a user's speech in dialogue practices in various language learning scenarios.

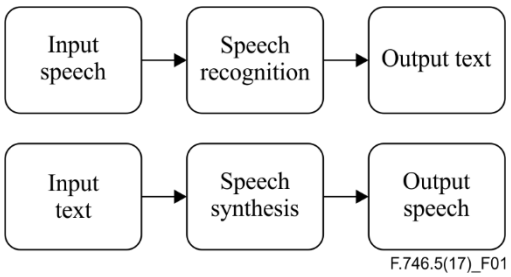

Figure 1 – Basic speech interface

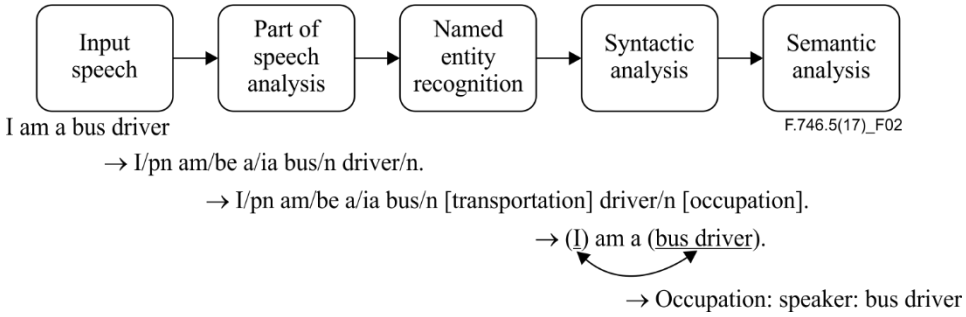

Figure 2 – Basic NLP flows with processing modules

6.2 Advanced technology for dialogue-based speech interface

More advanced technology for dialogue-based speech interface has also been developed. Figure 3 shows the basic dialogue processing flow. The technology recognizes and understands users' speech and generates appropriate responses in limited dialogue situations. The core technology is applicable to foreign language learning systems that simulate one-on-one conversation training. The core technology will also be used for natural speech interface of information services on mobile devices (smartphones, navigation terminals, etc.) or wearable devices such as smart glasses and smart watches. Using this technology, smartphones will accelerate various information services in mobile environments such as voice search, voice short message service (SMS), speech enabled virtual agents, etc. For example, the conversational speech processing technology can be applied to a high quality language learning system, and the resulting system can assess the pronunciation of students and leads the conversation to help upgrade speaking skills.

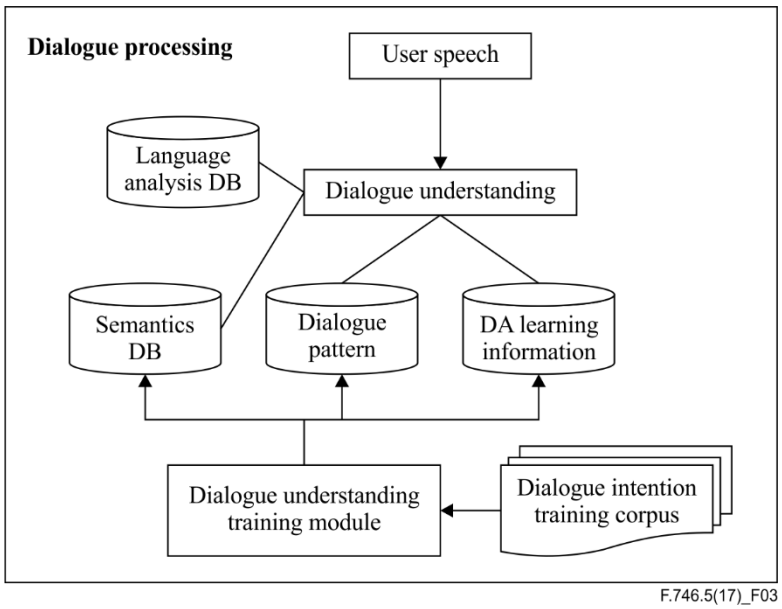

Figure 3 – Dialogue processing flow

7 Requirements for language learning system

The general requirements for a language learning system (LLS) based on speech/NLP technology are:

– LLS is required to provide speech recognition function for spontaneous speech.

– LLS is recommended to provide dialogue processing function for spontaneous utterances.

– LLS is recommended to provide evaluation function for spontaneous speech.

– LLS is recommended to provide error correction function for target language grammars.

– LLS is recommended to provide pronunciation modelling function for target language.

– LLS is recommended to provide expression learning function for target language.

– LLS is recommended to provide pronunciation fluency feature extraction function for target language.

– LLS is recommended to provide automatic evaluation function for pronunciation fluency for target language.

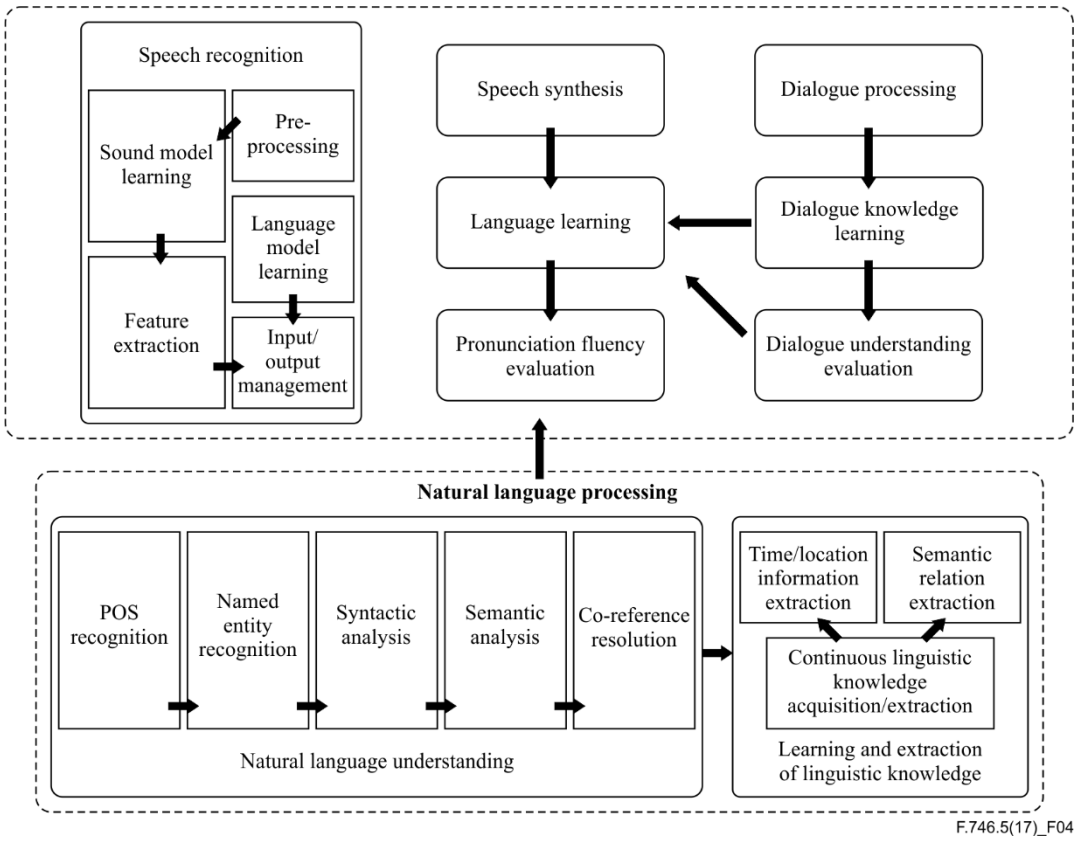

Figure 4 – Functional architecture of language learning system

8 Functional components and interfaces of a language learning system

Figure 4 shows the functional components of a language learning system. The essential modules of the functional components are as follows:

– Speech recognition module: supports speech recognition function for the user's input.

– Natural language processing module: supports natural language processing on the language learning system.

– Speech synthesis module: supports speech synthesis on the terminal for the system output.

– Pronunciation fluency evaluation module: supports pronunciation fluency evaluation function for the language learning system.

– Dialogue processing module: supports dialogue management and generation functions for the language learning system.
– Dialogue understanding evaluation module: supports evaluation of the user's understanding of the dialogue.
– Dialogue knowledge learning module: supports incremental knowledge learning on the dialogue.
– Language learning core module: supports core language learning functions such as grammar learning and dialogue expression learning.

8.1 Speech recognition module

The function of the speech recognition module is to generate output text based on the user's speech signals. The speech recognition module is composed of the following sub-functions: pre-processing, sound model learning, feature extraction, input/output management and language model learning sub-functions.

8.1.1 Pre-processing function

The pre-processing function extracts a feature vector and removes noise from the noisy speech input. In this module, distributed noises with a high energy level and active noises with low energy level are removed to improve the quality of sound.

8.1.2 Sound model learning function

The sound model learning functional entity produces a speaker independent acoustic model from a large speech database to recognize the general user's speech based on a deep neural network (DNN). This function also includes the process of building a speech database, and adaptation training for the specific speech environment, to make the speaker independent acoustic model optimal in the application environment. The sound model learning functional entity consists of a memory and a processor that generates sound-model state sets corresponding to multiple set training speech data. It then generates a multi-set state cluster from the sound-model state sets, and sets the multi-set training speech data as an input node, and the multi-set state cluster as an output node in order to learn a DNN structured parameter. When a user's speech and characteristic information are received by a user interface, the processor recognizes the user's speech on the basis of the learned DNN structured parameter by setting a sound-model state set corresponding to the characteristic information of the user's speech as an output node input/output management function.

8.1.3 Input/output management function

The input/output management function provides the audio input interface function which processes the speech data input from the microphone, and the output interface function which recognizes speech data and sends back the recognition result to the user.

8.1.4 Language model learning function

Language model learning functional entity performs training of word N-gram language model that defines the search network for continuous speech recognition. For language model training, text corpus of the recognition target should be prepared. The text corpus goes through a cleaning and refinement process first, and then language model training is performed on the clean text corpus. The language model is trained on the limited word list and the utterance dictionary is produced for the target word list. The language model database that stores a language model is used to convert the extracted feature vector into a text with a decoder.

8.2 Natural language processing module

The NLP module consists of a natural language understanding function and a linguistic knowledge learning/extraction function. The natural language understanding function is illustrated in Figure 5.

It analyses the user's speech using the technologies of semantic level analysis such as semantic entity recognition and co-reference resolution, in addition to the well-known part-of-speech recognition, named entity (NE) recognition and syntactic analysis for sentence structures. The linguistic knowledge learning/extraction function extracts semantic relations among words and time/location information in the documents, and provides the framework to continuously acquire linguistic knowledge and to extend it to be able to deal with new knowledge. Natural language understanding function consists of a part-of-speech recognition sub-function, named entity recognition sub-function, syntactic analysis sub-function, semantic entity recognition sub-function and co-reference resolution sub-function. Learning and extraction of linguistic knowledge function is composed of semantic relation extraction sub-function, time/location information extraction sub-function and continuous linguistic knowledge acquisition and extension sub-function.

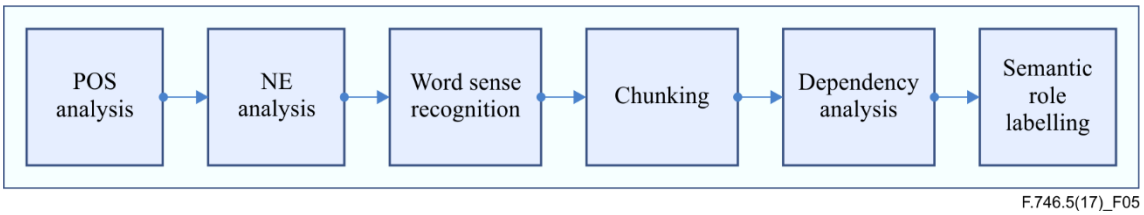

Figure 5 – Natural language understanding function

8.3 Speech synthesis module

Also known as text-to-speech (TTS) module, the speech synthesis module generates speech signals from the text input by text normalization and morphological analysis, intonation target (i.e., pitch contour) generation and speech signal generation using language analysis and intonation prediction information for the input sentences.

8.4 Pronunciation fluency evaluation module

The pronunciation fluency evaluation module evaluates the user's spontaneous speech input using the fluency feature model. The fluency evaluation consists of the analytic evaluation and the holistic evaluation for each evaluation item. The total scores of the analytic evaluation is the addition of each evaluation item score with different weights. The holistic evaluation is based on the impression of the fluency of the user by the evaluators. The final pronunciation fluency evaluation result is decided by selecting one of two evaluation methods.

The pronunciation fluency evaluation is performed with the speech recognition module whereby the utterance verification function verifies the user's speech based on multi-event detection information and a predefined utterance verification model. To verify the user's utterance in a speech recognition system the following processes are used:

– a noise processor configured to process the noise of an input speech signal;
– a feature extractor configured to extract features of speech data obtained through the noise processing;
– an event detector configured to detect events of the plurality of speech features occurring in the speech data using the noise-processed data and the data of the extracted features;
– a decoder configured to perform speech recognition using a plurality of pre-set speech recognition models for the extracted feature data; and
– an utterance verifier configured to calculate confidence measurement values in units of words and sentences using the information on the plurality of events detected by the event detector and a pre-set utterance verification model, and perform utterance verification according to the calculated confidence measurement values.

The event detector includes a detector configured to detect features such as a noise level, a breathing sound, an interjection, a repetition of a phrase, hesitation, an exclamation, sound stretching, laughing, speech rate, and an incorrect pronunciation from the noise-processed data and the extracted feature data.

The utterance verification model includes a word-level utterance verifier and sentence-level utterance verifier.

The word-level utterance verifier calculates word-specific confidence scores of words of sections in which the events have occurred and words for which no event has occurred using the specific anti-model and the filler model pre-set in a database.

The sentence-level event application utterance verifier calculates the confidence scores of the sentences using the statistical method, such as support vector machines (SVM) model, which has been pre-set in a database and trained by the statistical model such as SVM with training data reflecting a feature of a natural language. For sentence-level utterance verifier, the sentence structure and meaning analyser analyses the sentence structure and the meaning of the speech recognition result sentences using a morpheme analyser such as POS tagger and named entity analyser.

8.5 Dialogue processing module

The **dialogue processing module** performs dialogue act understanding function to express the dialogue intention/act of the utterances by analysing user's utterance strings and understanding their meaning.

The **dialogue act training function** automatically trains dialogue patterns and statistical classification training information based on the domain dialogue corpus annotated with the dialogue acts to understand the participating user's intention.

The **dialogue management function** analyses the n-best dialogue acts and meaning expressions and decides the best dialogue act, which is used to generate relevant system dialogue and manage the dialogue situations.

The **dialogue generation function** selects the correct system response templet for the user's intention and decides meaningful information values to fill in the slots of the selected templet. Finally, it generates the appropriate system dialogue sentences.

Dialogue management and generation function provides the following sub-functions:

– **User intention analysis function**, to understand the user's intention correctly for the current topic

– **Auxiliary topic handling function**, to provide responses if the topic is outside the main topic and guide the dialogue towards the appropriate topic flow

– **Topic management knowledge search function**, to find relevant response patterns, slots or task knowledge to achieve the objective of the current topic

– **Dialogue history management function**, to store changing tasks, slots and intentions and monitor the intention flows of the dialogue

– **Next utterance recommendation function**, to suggest the next intention and utterances of the user according to the ongoing dialogue flow

– **System reply generation function**, to provide responses according to the user's intention

– **Dialogue management log record function**, to record the dialogue flow, the history change or search conditions. This log information is used to check the errors in the system and dialogue knowledge

8.6 Dialogue understanding evaluation module

The dialogue understanding evaluation module illustrated in Figure 6 evaluates the user's understanding of the ongoing topic by comparing user's utterances to the dialogue model of the system. It uses the dialogue understanding function and dialogue management function to search the dialogue knowledge. This function is part of the language learning module as an evaluation function. The system provides feedback, such as a better example to the user while evaluating the dialogue understanding to correct his responses.

The response from the system differs according to the user's utterance determining if the topic is relevant or not. If the user's response contains grammatical errors or wrong expressions, the system provides the correct information for language learning purpose.

The system guides the dialogue if the user responds with a different topic by restoring the ongoing topic of the dialogue. It can also change the topic to achieve the final goal of the dialogue and for the language learning purpose.

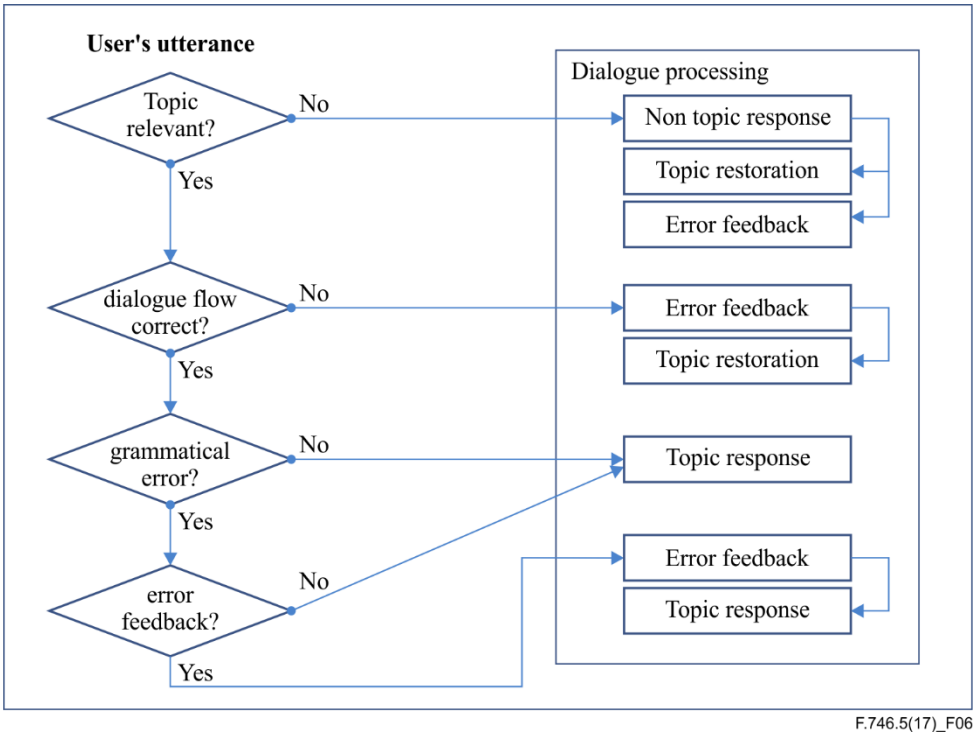

Figure 6 – Language learning by dialogue processing function

8.7 Dialogue knowledge learning module

The dialogue knowledge-learning module adds new dialogue knowledge in the dialogue management database by pseudo situation simulation and incremental learning of situation states. The sub-functions are automatic extraction function of conversation scenarios and incremental dialogue learning sub-functions.

Automatic extraction of conversation scenarios sub-function: serves to collect big data dialogues among people from the scenarios of movies and dramas, and performs the speaker boundary recognition, relevance assessment and topic classification.

Incremental dialogue learning sub-function: identifies the topic of the dialogue scenarios, augments with the slot information, intention and subjects according to the topic, and produces automatically or by human validation the dialogue knowledge that can be used for the dialogue system.

8.8 Language learning module

The language learning module, illustrated in Figure 7, takes the user's speech as input and automatically detects grammatical errors, corrects them and then sends the corrected sentence to the user as a feedback to improve his/her competence of foreign languages. The language-learning module consists of the grammatical error correction function and the expression learning function. The grammatical error correction function detects errors included in the user's spoken sentences and corrects them. On the other hand, the expression learning function provides feedback of the grammatical error and suggests better expressions to the user.

– Grammatical error detection sub-function: detects grammatical errors in the user's spoken sentences. The grammatical errors are limited to those defined in the system. Multiple errors can be detected in one sentence. Various error detection models are used for the grammatical error detection.

– Grammatical error correction sub-function: provides the correction information for detected grammatical errors. The error correction function produces grammatical error correction information by using the POS of the error words and context information around the words.

– Error selection and correction sub-function: selects the best error candidate among those from different error detection models and generates error correction information on the selected candidate error.

– Error feedback providing sub-function: informs the user of the grammatical errors and provides better expressions. The explicit/implicit error feedbacks are used accordingly depending on the situation of the error occurrences.

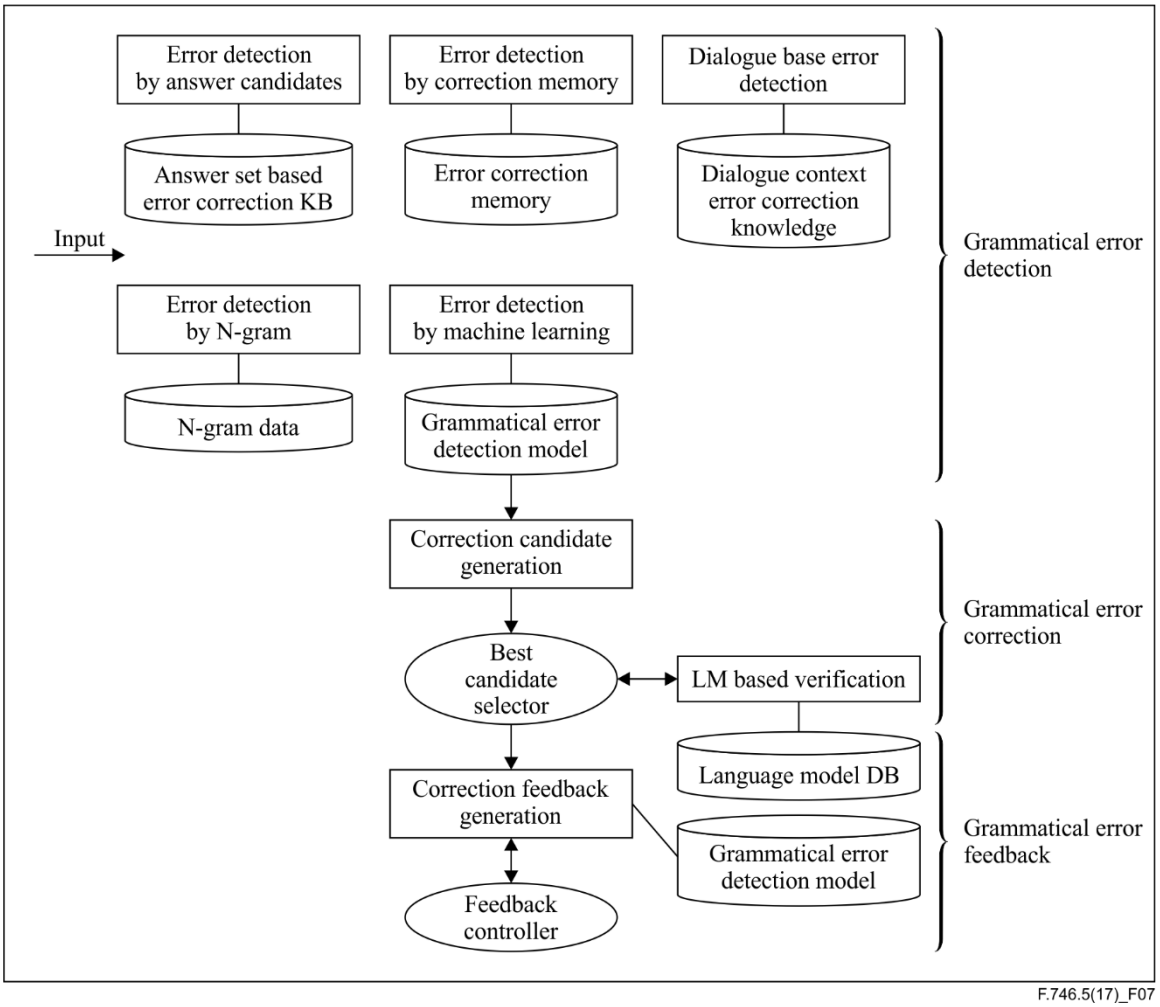

Figure 7 – Language learning module

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	**Non-telephone telecommunication services**
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2018