IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 12, 2020, accepted August 31, 2020, date of publication September 8, 2020, date of current version September 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022674

Towards a Context-Free Machine Universal Grammar (CF-MUG) in Natural Language Processing

QUANYI HU[1], JIE YANG[1,2], PENG QIN[1], (Graduate Student Member, IEEE), AND SIMON FONG[1], (Member, IEEE)

[1]Faculty of Science and Technology, University of Macau, Macau 999078, China
[2]Chongqing Industry and Trade Polytechnic, Chongqing 408000, China

Corresponding author: Simon Fong (ccfong@um.edu.mo)

This work was supported in part by the Nature-Inspired Computing and Metaheuristics Algorithms for Optimizing Data Mining Performance, University of Macau, under Grant MYRG2016-00069-FST, in part by the Temporal Data Stream Mining by Using Incrementally Optimized Very Fast Decision Forest, University of Macau, under Grant MYRG2015-00128-FST, in part by the Scalable Data Stream Mining Methodology, Stream-Based Holistic Analytics and Reasoning in Parallel, FDCT, Macau Government, under Grant FDCT/126/2014/A3, in part by the Key Project of Chongqing Industry and Trade Polytechnic under Grant ZR201902 and Grant 190101, and in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant KJQN201903601 and Grant KJQN2020.

ABSTRACT In natural language processing, semantic document exchange ensures unambiguity and shares the same meaning for documents sender and receiver cross different natural languages (e.g., English to Chinese), this difference makes the translation between natural languages becomes complex and inaccurate. This paper proposed a novel framework of Context-Free Machine Universal Grammar which consists of local mode (sender and receiver) and mediation mode (Machine Universal Language) based on the concept of collaboration, the framework improves semantic unambiguity and accuracy in crossing language document, meanwhile makes document computer-readable through unique ID for each word or phrase. More importantly, inspired by grammatical case in linguistics, a novel Machine Universal Grammar provides a universal grammar that accepts all coming languages and improves semantic accuracy in natural language processing.

INDEX TERMS Semantic document exchange, natural language processing, universal grammar.

I. INTRODUCTION

In global document exchange, the accuracy of translation and disambiguation during semantic document exchange [1]–[3] is crucial crossing different languages and make sure the sender and receiver understand the exchange document's meaning. When a document is exchanging across unknown business parties (i.e., cross domains or cross contexts), the meaning interpretation between the document sender and the document receiver might be different and significantly impact on the correct execution of business transactions. Table 1 shows that a simple inquiry sheet.

If it is written by a document writer of company A (English users) send to company B (Chinese users). Unfortunately, human cognitions on the same document are often different, Company B in another context has no knowledge of this semantic relationship for the incoming document. It thus cannot correctly interpret. As interesting as these ideas maybe, they could remain restricted and void in the absence of a theoretical foundation to explain the situations in the following examples:

- If company A writes with a particular document template and sends it to a potential company B who never knows the company before, and if the document template is peculiar to company A, what shall the potential company B do?
- If Company A writes an inquiry sheet in English and sends to a potential company B who only knows Chinese, what shall the potential company do?

These reactions implicitly illustrate the non-autonomy of company B which is obliged to make remedies in securing

The associate editor coordinating the review of this manuscript and approving it for publication was Chang Choi.

TABLE 1. Inquiry sheet.

EMPLOYEE INQUIRY INFORMATION			
Date Requested:			
Employee Name:		Employee Number:	
E-Mail Address:		Phone:	
INQUIRY DETAILS			
Complete the form and give it to the receptionist at the Human Resources department in {ENTER LOCATION}. All inquiries will be addressed within 48 hours of receipt.			

a business chance. In crossing language business document exchange, it is not easily accessible to the majority of users because of language barriers that hamper the cross-lingual search. Problems arise due to ambiguity in language, existing techniques just understand the sheet template, for example, ''Date Requested'', ''Employee Name'', but company B does not know how to process user filled sentences in the inquiry sheet, for example, ''INQUIRY DETAILS'' part is required to fill the form of inquiry. A semantic document autonomously designed by the writer in one context is also consistently interpretable by a document reader situated in another context exactly as the document writer means. Therefore, based on this situation, one problem requires that the computer translates sentences in crossing languages, not just understands the template of business documents. In crossing language document exchange, it is not easily accessible to the majority of users because of language barriers that hamper the cross-lingual search. The problem rises to translation in natural language processing.

The lack of available resources and limitations have motivated many scholars to rely on hand-constructed linguistic rules. Inspired of language features, and drawback of solving in the sentence-based translation problem when business document exchange, this paper proposed a Context-free Machine Universal Grammar (CF-MUG) framework, CF-MUG mainly consists of three layers: local layer, mapping layer and mediation layer. The local layer is that human user 1 instructs the computer system to do something via a set of connected computers and the document should be human and computer readable. The mediation layer is human user 1 communicates with human user 2 by drafting and sending documents (computer readable) via a set of computers which recreate or infer a set of new documents for the recipient human 2. The mapping layer is the rule-based structural mapping between the source language and target language based on predefined grammatical rules. In the mediation layer, a novel Machine Universal Grammar (MUG) based on the grammatical case is proposed, which is a universal language structure that provides a mediation that accepts all languages. Meanwhile, to ensure document exchanging computer-readable and understandable, each term looked up from the CONDEX Dictionary [4] is tagged a unique IID. The characteristics of the proposed system framework are:

- Computer readable and understandable through unique ID for each term
- It adapts to any crossing languages business documents exchange
- It has a strict sense of well-formedness in mind for computer and human
- Reduce the complexity of handling crossing language processing.
- It allows the use of heuristics (such as a verb cannot be preceded by a preposition)
- It relies on hand-constructed rules that are to be acquired from language specialists rather than automatically trained from data.
- It is easy to incorporate domain knowledge into linguistic knowledge which provides highly accurate results.

This paper is organized as follows: Chapter II introduces the related works. Chapter III introduces our proposed framework -CFMUG. Chapter IV introduces Machine Universal Grammar. Chapter V illustrates the steps of sentence translation. Chapter VI introduces the definition of representation for one sentence or plaintext or any sentence-based documents. Chapter VII shows the evaluation. Finally, a conclusion of our works.

II. RELATED WORK

A. TRANSLATION

When people communicate with each other, their conversation relies on many basic, unspoken assumptions, and they often learn the basics behind these assumptions long before they can write at all, much less write the text found in corpora. Meanwhile, document exchanging is often a process of user participation to translate context correctly. Existing approaches that firstly the raw data are collected and processed to make them web consumable. Once the data are converted into a common format, they are then semantically enriched based on the knowledge of domain experts, the collected data are processed using rules to deal with the uncertainty aspects of the semantic model. The idea is to recognize activity and learn new rules that are governing an activity.

The machine Translation process can be broadly classified into the following approaches Machine Translation process can be broadly classified into Rule-based [5]–[7], Statistical-based [8]–[10] and Neural-based approaches [5], [11], [12]. For these approaches, the translation system is trained with a bilingual text corpus to get the desired output. A bilingual corpus is trained, and parameters are derived to reach the most likely translation. From a linguistic point of view, what is missing in this approach is an analysis of the internal structure of the source text, particularly the grammatical relationships between the constituents of the sentences. Rule-based Translation considers semantic, morphological and syntactic information and based on these rules the source language is transformed to the target language through an intermediate representation. For example, using translation rules translates

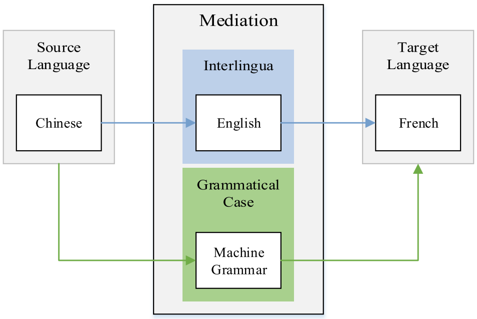


FIGURE 1. **Processing of rule-based approaches.**

the input language to the output language which is done in three phases. Firstly, the source language is converted into an intermediate representation which is subsequently converted into target language representation in the second phase. The third phase involves the generation of the final target language. Figure 1 shows approaches to rule-based natural language processing. There are two main approaches: the Interlingua approach uses the third common language as a mediation, such as English, however, the method increases transmission complexity as it passes through the third language and reduces the accuracy of the language. Therefore, this paper proposes a novel mediation which is a universal language based on the grammatical cases, in the following chapters will discuss in detail.

B. SEMANTIC DOCUMENT

There are three existing approaches to enabling semantic document exchange and representation [13] such as semantic web, machine translation [14], text processing [15], and human-computer interaction [16], which are approaches of standard-based, ontology-based and collaboration-based. They have different contexts and requirements for autonomy. The Standard-based approach refers to a kind of semantic document exchange methodology of developing standard terms and models for building semantic document and hence enabling semantic documents to be readable and interpretable among standard adopters and their document systems. Relevant to e-business documents, there are three types of standards, which are EDI-based [17], XML-based [18] and Web service-based. However, document standards only guarantee the semantic consistency for document templates and partial document terms within the scope of standard adoption. They do not consider the semantic conflicts out of standard scope. To improve the standard-based approach and enable a wide scope of semantic document exchange, the ontology-based approach has been developed [19], which refers to a kind of semantic document exchange methodology such that semantic documents are composed of ontological models. It is a formal description of the concepts and relationships approach to semantic document exchange is a methodology of semantically creating, use and exchanging documents. Existing popular ontology languages are RDF (Resource Description Framework) [20] and OWL (Web Ontology Language) [21]. Nevertheless, for e-business semantic document exchange, OWL-based ontology for document representation has an inherent problem such that ontology is domain-wide. The ''domain-wide'' means that when an e-business document is exchanging across unknown business parties (i.e., cross domains or cross contexts), the meaning interpretation between the document sender and the document receiver might be different and significantly impact on the correct execution of business transactions. This is because the document senders and document receivers may have different interpretations on the same ontology due to context heterogeneity. To eliminate the ''domain-wide'' problem of the ontology based approach, the collaboration-based approach has been developed [22]–[26], which refers to a kind of semantic document exchange methodology such that terms for composing semantic documents are collaboratively designed and implemented by collaboration tools and document models are generic for all domains and contexts. Through this approach, any semantic document can always achieve semantic consistency in document exchange across heterogeneous domains and contexts.

In summary, semantic document representation approaches of standardization, ontology modeling, collaborative templating and autonomous editing all have their respective merits and drawbacks. Their researches are in progress but still not resolve one of problem are document features of document types have not been extracted for document representation; Sentence-based semantic structure has not been researched. To resolve the above issues, this paper focuses on the research of how a user-generated document can directly semantically understandable and comprehensible by both other humans and computers across heterogeneous contexts. Fulfilling this task will tremendously save intermediate work used in document processing and increase the understandability and comprehensibility of both humans and computers. Therefore, this paper proposes Machine Universal Grammar (MUG) through the reconstruction of collaborative concepts based on sentence document, where the collaborative common dictionary is provided to reconstruct any semantic terms, phrases and sentences. The execution of this will lay a solid foundation on how to enable the understanding and comprehensibility of semantic documents by both humans and computers. The research result will have high theoretical values in information digitalization, semantic study and communication.

C. UNIVERSAL GRAMMAR

Construction of sentence structure plays an important role in the mediation model when semantic document exchanges. All human languages consist of sentences, but they vary in the sentence structure, as it shows the physical nature of the sentence and explains the elements from which the sentence is made up. We note that there is a difference between languages in the structure of sentences which adjusts the structure of sentences in the language, that is the language rules, and in light of this we conclude that languages share the components of words and differ in their structure and rules. Nevertheless, it is difficult to find general rules that represent all languages, that is, each language is unique by its rules and characteristics. The word order has to do with the arrangement of the grammatical structure of language, for human languages differ in the order of words, that is to say, the way sentences are structured of the language fundamental components. This is a feature that distinguishes a language from another as seen by linguists. One of the divisions of these scholars of languages is based on the way sentences are structured in the discourse of a particular human group. They divide languages into various types according to the succession of a sentence (Subject), (Verb) and (Object) as well as the (complements), which is regarded as a distinctive feature of a particular language. A sentence, any sentence, consists basically of a verb, a subject, and an object, with other additions. There are six patterns represent the word order in a language: they are added (SVO) subject, verb, object, (SOV) subject, object, verb, (VSO) verb, subject, object, (VOS) verb, object, subject, (OSV) object, subject, verb, and (OVS) object, verb, and subject. The overwhelming majority of the world's languages follow either SVO or SOV patterns. Some languages have a fixed word order, and others have a free unfixed word order. The word order in the human language is arranged on several structures that consist of the subject (S), the object (O), and the verb (V), and languages have been classified into six categories according to the word order structure that can be found in human languages. An example shows in Table 2.

Languages structures are similar it in general, but they are different, construction of sentence structure plays an important role in the sentence-based business document. To measure a language whether is a universal grammar structure [27]–[29], it depends on the language is compatible with various linguistic phenomena. From the perspective of linguistics, every language has sentences that include a Subject, Object and a verb, although some sentences do not have all three elements. Languages have been classified according to the basic or unmarked order in which these constituents occur in the language. The sentence structure of English and Chinese follows the Subject+Verb+Object (SVO) order. Unsurprisingly, there are many differences between the two languages. Sentence structure (syntax) is one of the areas in which great differences exist. For example, the positions of the modifier are different. In English sentences, the modifier can be placed either before or after the modified elements (subject, predicate or object), but if the modified element is a phrase or clause, the modifier is always placed after the modified elements. In Chinese sentence, the modifier is always placed before the modified elements. There are two examples:

1: *We* ① *wanted to create* ② *a watch* ③ *that enchants the consumer with the beauty of its movement.* ④

Structure: Subject ① Verb② Object③ Modifier (Adjectival Clause) ④

TABLE 2. An example of word order.

Example Rules:
subject –> I
verb-phrase –> <adverb> <verb> | <verb>
adverb –> always
verb –> like
object –> the <noun> | <noun>
noun –> game

Word Order	Rule	English equivalent	Example Language
SVO	**sentence –> <subject> <verb-phrase> <object>**	I always like the game	Latin, Japanese
SOV	**sentence –> <subject> <object> <verb-phrase>**	I the game always like	English, Mandarin
VSO	**sentence –> <verb-phrase> <subject> <object>**	Always like I the game	Irish, Filipino
VOS	**sentence –> <verb-phrase> <object> <subject>**	Always like the game I	Malagasy, Baure
OVS	**sentence –> <object> <verb-phrase> <subject>**	The game always like I	Apalaí, Hixkaryana?
OSV	**sentence –> <object> <subject> <verb-phrase>**	The game I always like	Warao

Direct Translation (Chinese): *我们* ① *希望制造出* ② *一个腕表* ③ *心醉消费者以它的运动之美* ④ (×)

The adjectival clause should be placed before the object/verb to form a {modifier + modified element} structure in Chinese sentence. And, the conjuction "that" before the adjectival clause can be ignored. Particular "的" (auxiliary word, particle; used before noun or noun phrase) should be added after adjectival clauses, before objects.

Corrected:
我们希望制造出以它的运动之美心醉消费者的
一个腕表. (√)

2. *Susana* ① *was attacked* ② *by a vagrant who usually spent the night near the LRT station.* ③

Structure: Subject ① Verb② Modifier (Adverbal clause) ③

Direct Translation (Chinese): *苏珊娜* ① *被攻击了* ② *一个无业游民, 通常过了一夜轻铁站附近*. ③ (×)

This is a "被" (by) sentence with an action performer (a vagrant). The verb clause structure is "被" + action performer + verb. The elements in the sentence structures are then reorganized to form a syntatically correct Chinese sentence with subject ① + "被" adverbal clause③ +verb ② order. And, the conjunction "who" can be ignored.

Corrected:
苏珊娜被 一个通常在轻铁站附近过了
一夜的无业游民攻击了. (√)

Based on the discussion above, the main difference in sentence structure between English and Chinese language is the position of clause modifier in the sentence. And, the different positions of clause modifier in English and Chinese sentences make difference to the two languages. This paper concludes some structure issues in Chinese and English:

1. The adjectival or adverbial clauses.
2. Conjunctions (who, which, that, etc.) and prepositions (with).
3. Particular Chinese character "的" (auxiliary word, particle; used before noun or noun phrase).
4. Passive voice sentences.
5. Sentences with clauses and sub clauses.
6. Verb Tense representation.

After studying both linguistic phenomena, in fact, ordering, increasing or decreasing any element constructions in different phenomena are performed through by the specific rules. More importantly, all languages can be constructed in the same grammatical pattern. It could be assumed that the sentence pattern of all languages is rather similar or same he pattern of each grammatical construction, phrase, clause and sentence is supposed to be the same. Most of all the meaning transfer from source to target language tends to be accurate because the identical surface structure usually underlies or refers to the same deep structure or pattern.

III. MEDIATION APPROACH TO DOCUMENT EXCHANGE FRAMEWORK

The proposed Context-free Machine Universal Grammar (CF-MUG) presents a rule base translation framework which is designed for global document exchange, CF-MUG mainly designs two modes: local user mode and mediation mode. Local user mode provides checks and maintains local language grammar between terms used for composing universal semantic documents by resolving semantic mapping conflicts. Then, it generates the computer-readable documents using CONEX Dictionary; Mediation mode: A human user 1 initiates a communication with a remote human user 2 by drafting and sending a business document via a set of connected computer which might mediate, recreate or infer a set of new documents for the document's recipient human user 2, the mode provides the exchange platform of different language documents and reorders the computer-readable semantic documents; Users are simply semantic document writers and semantic document readers who write and read exchangeable semantic documents. By these modes, Machine universal grammar approaches semantic document exchange crossing any languages in two layers. The one layer represents the computer-readable and understandable document. Another layer is the exchangeable semantic document model of the systems which represents semantic documents in an exchangeable manner. Figure 2 logically shows the framework of CF-MUG.

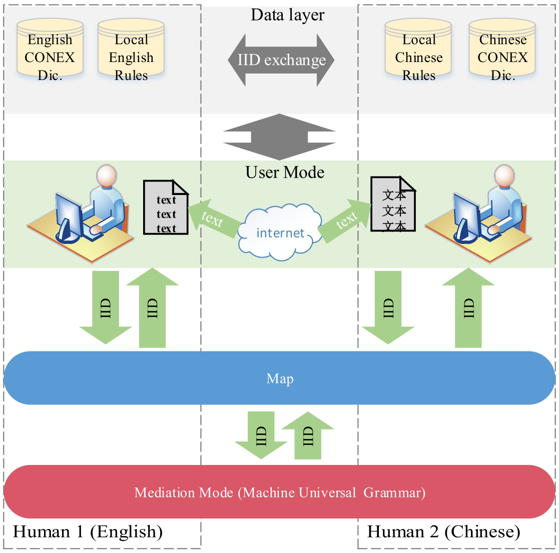


FIGURE 2. Mediation approach to semantic document exchange framework.

Step 1 (Local mode-1): Human user 1 inputs the sentences, the local mode has a Chinese CONEX Dictionary and Chinese grammar self-checking system. CONEX Dictionary transforms inputted documents into the computer-readable and understandable document which includes ID sequences term by term. To present a universal sentence for better mapping to Human user 2 language by one to one, we provide the function of local grammar self-checking by rule-based approach, the function removes or modifies local vocabulary (such as "的" DE, "吧" BA in Chinese) and grammar rules to approach universal grammar rules.

Step 2 (Mediation mode): During the mediation mode, the document exchange transforms into machine code exchange, which means the document is computer-readable and understandable. Mediation mode reorders and maps Chinese sentence to be English sentence structure order using context-free grammar based on defined structure rules and Chinese⇔English CONEX Dictionary. In step 1, preprocessed the sentence makes map words one by one. Reordering to target order ensures that semantic transfer accurately. Therefore, Human user 2 received a computer-readable universal document.

Step 3(Local mode-2): Local mode-2 makes the universal document to be near local English grammar, because in the mediation mode, mode changes to English sentence structure, such as SOV to SVO, so mapping in step 2 does not guarantee that the grammar is completely correct, only guarantee structure is correct. Based on English grammar self-checking, local mode modifies the incorrect sentence to conform English grammar, such as like change to likes because the sentence's subject is he, then generate a new English document for Human user 1. In Fig.3, A simple and general step shows how to transfer from a Chinese document to an English document.

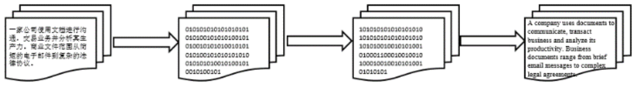


FIGURE 3. Semantic document exchange from Chinese to English.

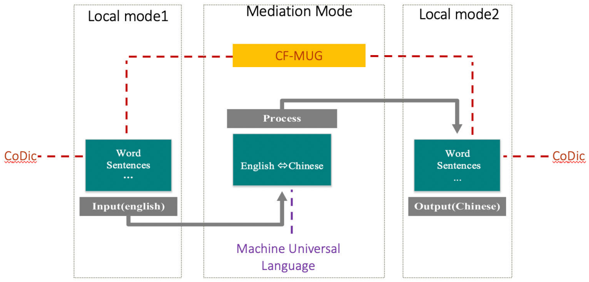


FIGURE 4. CF-MUG framework.

Meanwhile, Fig.4 shows the difference between traditional translation and our proposed method.

In summary, local mode: encoding a human inputted sentence to a human-machine readable sentence and decoding the human-machine readable sentence to a human readable. Mediation mode: generate machine-readable sentence from human-readable sentence in the sender location and decode machine-readable sentence to human-readable sentence (in the receiver location). It is based on the case grammar for incoming and outgoing sentences.

The novel collaborative framework through a mediation approach – CF-MUG has the characters:

1) Treat any message as a document, a document as a tree of sentences, and a sentence as a compound concept that is a sequence of atomic concepts for all human natural languages such as Chinese and English.
2) Develop a novel machine natural language (MNL) as a kind of mediation language.
3) Map all compound concepts of sentences from sender's sentences onto the compound concepts sentence of receiver's sentences through a set of grammar rules.

IV. MEDIATION – MACHINE UNIVERSAL GRAMMAR

Machine Universal Grammar (MUG) is a universal language and the heart of mediation mode, even in the whole framework of CF-MUG. The MUG refers to all languages written by human beings, as opposed to artificial languages, computers, mathematical or logical. In fact, the grammar does not directly concern the language but covers linguistics structure. MUG approach aims to achieve the translation task in two independent steps. First, meanings of the source language sentences are represented in a universal language-independent (Mediation) representation. Then, sentences of the target language are generated from those universal language representations. In this chapter, we explain why we proposed this method from a sentence structure perspective. We introduce components of MUG from basic elements – Term and Part-of-Speech (PoS), and then sentence elements - Grammatical Case (GC).

A. VACABULARY – TERMS AND POS

After a brief introduction of the framework, we go to the most basic element in machine universal grammar – vocabulary. A part of speech (POS) is a category of words that has similar grammatical properties and determines the definition of grammar rules. Therefore, tagging words plays an important role when generating a new document. CONEX Dictionary consists of generous words for a different language and each word includes a unique ID that presents the same meaning word in a different language to make the document universal and computer-readable. Generated Semantic document by CONEX Dictionary is a method of defining the meaning which enables any language document without meanings to be meaningfully signified for both human and computer to correctly interpret, that means human-readable and computer-readable. CONEX Dictionary has these elements, for example:

$W_{\text{lang}} = \{\text{id, term, POS, Description}\}$
$W_{\text{en}} = \{001, \text{“Apple”}, \text{Noun}, \text{“a kind of fruit”}\}$
$\Leftrightarrow W_{\text{cn}} = \{001, \text{“苹果”}, \text{Noun}, \text{“一种水果”}\}$

As an example, the English Dictionary must have the same ID in the same meaning with the Chinese Dictionary. During document exchange, we guarantee semantics disambiguation, which not only human users understand the same meaning, but also the computer also understands the same meaning, to maximize the accuracy of the semantics. POS depends on the dictionary to get possible tags for each word to be tagged and are used to identify the correct tag when a word has more than one possible tag. Disambiguation is done by analyzing the linguistic features of the word, its preceding word, its following word and other aspects. For example, if the preceding word is article then the word in question must be noun. This information is coded in the form of rules. Unique term ID is supposed as universal computer-readable context for mediating heterogeneous contexts. The unique term ID establishes the concept mapping between a human interpretation of the objects onto a computer. Thus, given heterogeneous contexts or domains such as local context A (e.g., a Chinese human-readable document), common text B (e.g., Chinese computer-readable document), common context C (e.g., English computer-readable document), and local context D (e.g., an English human-readable document), such that any heterogeneous term A'∈A, B'∈B, C'∈C and D'∈D with the same meaning can always be transformed between heterogeneous contexts along a concept supply chain of $A' = \text{Map}(A', B') \Rightarrow \text{Map}(B', C') \Rightarrow \text{Map}(C', D') = D'$. In our machine universal grammar framework, words are roughly classified in types enumerated based Chinese and English as follows:

Noun (*n*) : A noun is a word denoting an entity such as a person, a place or a thing. It is uninflected by its own in this defined MUG. There are four kinds of nouns, which are *common noun (ncm), person proper noun (npp), geographical proper (ngp) noun*, and *organizational proper noun (nop)*. The common nouns can all be defined in a dictionary while proper nouns are not possible to be fully enumerated and defined in a dictionary. Thus, in designing a machie universal language grammar that aims at being computationally processed by computers, how to dynamically recognize a proper noun and differentiate it from other parts of a sentence becomes a problem. In the MUG, a noun can take its own functional form as a noun or as other parts of speech based on the rules defined.

Verb (*v*) :

A verb is a word denoting an action, an event, or a state of being. There are two kinds of verbs: *common verb* (vc) and *linking verb* (vl). The former consists of three forms of *intransitive verb* (vi), *transitive verb* (vt) and *ditransitive verb* (vd). An intransitive verb cannot follow any object. A transitive verb must follow one object that is accusative. A ditransitive verb must follow two objects in which one is accusative and the other is dative. A linking verb plus its followed phrase defined by rules form a larger predicative phrase similarly as interpreted in English grammar.

Adjective (ad):

An adjective is a describing word denoting a state or an attribute that qualifies a noun or a noun phrase. It means that a noun possesses a property that the adjective is describing.

Adverb (av):

An adverb is an uninflected word that modifies a verb, an adjective, another adverb, clause or sentence. It typically describes a property of being modified such as manner, place, time, frequency, degree, or level of certainty or possibility. It answers questions such as how, in what way, when, where, and to what extent. This function of an adverb is called adverbial function and can be realized by a single word of adverb or by a multi-word expression (i.e., an adverbial phrase and an adverbial clause).

Numeral (nm):

A numeral is a word denoting a number that can be dynamically assembled following a construction a number rule. For example, twenty + one = twenty one. When a numeral is in front of a noun, it is an adjective. When it is independent, it is a noun.

Onomatopoeia (on):

Onomatopoeia is an uninflected word that phonetically imitates, resembles or suggests the sound that it describes. It belongs to a special sub-category of uncountable nouns. It can often be used as an independent sentence structure, as a component of a sentence, or as a modifier of a noun or a verb.

Particular (pa):

A particular is pair of words in the form of "beginning mark of an instance + data type + (a word in dictionary) + X + ending mark of an instance", in which beginning mark and ending mark signify a particular word. For example, when we want to express "16.53" of a concept for "length", we can write as "1+decimal+16.53&meter + (length)". It reads 16.53 meter in length. Here, "1" indicates a particular word; "decimal" indicates a data type; "&" is used to reference to a ready concept (often a word denoting a unit) in the MUG dictionary. The concept "length" is optional.

Preposition (pp):

A preposition is an uninflected function word that connecting with a noun or a noun phrase to express place, time, manner, and other aspect. It cannot be used independently in a sentence. The noun or noun phrase following a preposition is the object of the preposition. A preposition plus a noun or noun phrase is called prepositional phrase (PP). A prepositional phrase can take different cases depending on its context in a sentence.

Conjunction (cj):

A conjunction is an uninflected function word that serves to conjoin words or phrases or clauses or sentences. It denotes to follow a meaning group or connects two or more meaning groups. There two kinds of conjunctions: *beginning conjunction* or *intermediate conjunction*. A beginning conjunction introduces a phrase that constructs a meaning group (MG) that follows a preposition, while an intermediate conjunction links two phrases before and after it. The main purpose of conjunction is to build the structure of the meaning groups of a sentence. Based on the rule of conjunction processing, a sentence will be finally grammatically parsed as a tree of meaning groups in the form of MG (MG (...), ..., MG (...)).

Mark (mk):

A mark is an uninflected function word that denotes a beginning or an ending of a meaning group. There are three types of marks: in-sentence mark, sentence mark, and non-sentence mark. *In-sentence marks* are a kind of punctuation marks that are used to separate two different meaning groups. There are three in-sentence marks defined in MUG, which are comma, semicolon and colon.

(1) Comma (,) is to separate two different meaning groups.
(2) Semicolon (;) is to separate larger meaning group that comma does, or to separate two or more clauses without conjunctions.
(3) Colon (:) is to introduce an accusative object or appositive of the accusative object, which can be an independent phrase, a sentence or a paragraph.

A *sentence mark* is to signify an end of a sentence. In the MUG that we are designing, a specific sentence mark not only signifies the end of a sentence but also informs the mood of a sentence. For example, period "." signifies a positive sentence and its end, while the question mark "?" signifies an interrogative sentence and its end. Differently, a *non-sentence mark* is a pair of function words specially defined in our MUG dictionary that signifies the beginning of a sentence, a paragraph or a leveled section.

Interjection (it):

An interjection is a word or expression of exclamation and expresses a spontaneous feeling or reaction. It is often used as an independent structure.

Article (at):

An article is a word that is used with a noun to specify grammatical definiteness of the noun, and in some languages extending to volume or numerical scope. In the defined MUG, article is not used in any MUG-based sentence. However, while a sentence including articles is translated into the machine natural language from any other language, the articles remain as the part of speech in article so that they can be noted when they are translated back to a non-MUG language. For MUG, it interprets an article as an adjective to modify a noun. The approach to processing an article is particularly called *property consistency maintenance* of a part of speech, which means that the defined machine natural language maintains a placeholder for a grammatical property particular to certain natural languages by mapping the blank placeholder in MUG onto the peculiar property in other natural language. For example, a placeholder for the number or gender of a noun.

B. UNIVERSAL GRAMMAR

Basically, the sentence structure follows the Subject + Verb + Object (SVO). Unsurprisingly, there are many differences, such as singular and plural of noun, gender of noun, active and passive voice, different tenses, special words and sentence type. The MUG concentrates on a contrastive study of language feature in an attempt to explore the similarities, particularly to handle the dissimilarities among languages, under such circumstances, it is best for us to seek common ground while reserving differences crossing different languages.

A part of speech can be either finite (*f*) or non-finite (*nf*). In MUG, a *finite part of speech* is a word or phrase that only provides the functions defined in the original part of speech usually predefined in a dictionary. A non-finite part of speech is a word or phrase that not only provides its predefined functions but also functions as another part of speech. Meanwhile, *n*, *ad*, *v*, *nm, on* and *pa* can not only take their own functions but also the functions of other parts of speech. Adverb (*av*) can only take its own function. Preposition (*pp*), conjunction (*cj*) and mark (*mk*) cannot independently exist but can only exist by taking the functions of other parts of speech. The finiteness and non-finiteness of a part of speech indicates that clues of how a part of speech can be used can be found.

A case is a form of grammar in which the structural relations hip of a part of speech with itself or other parts of speech in a sentence is analyzed. This differentiates the cases into two kinds: *intrinsic case* and *extrinsic case*. An intrinsic case denotes that a part of speech has various forms to indicate different situational usages to express itself. An extrinsic case denotes the particular structural function of a part of speech while it aligns with other parts of speech. It informs an approach to the phrase combination when several parts of speech are given.

1) EXTRINSIC CASE

In the defined MUG, these are eight extrinsic cases. They are:

TABLE 3. Allowed extrinsic cases of parts of speech.

CASE pos	N	A	D	G	P	B	C	S
n	*f*	*f*	*f*	*nf*	*(nf)*	*(nf)*	×	*nf*
ad	*nf*	*nf*	×	*f*	*nf*	×	*nf*	*nf*
vi	*nf*	*nf*	×	*nf*	*f*	*nf*	*nf*	*nf*
vt	*nf*	*nf*	×	*nf*	*f*	*nf*	*nf*	*nf*
vd	*nf*	*nf*	×	*nf*	*f*	*nf*	*nf*	*nf*
vl	*nf*	×	×	×	*f*	*nf*	×	×
av	×	×	×	×	×	*f*	×	×
nm	*nf*	*nf*	×	*nf*	×	×	×	*nf*
pa	*nf*	*nf*	×	*nf*	×	×	×	*nf*
it	×	×	×	×	×	×	×	*f*
pp	*nf*	*nf*	×	*nf*	×	*nf*	×	×
cj	*nf*	*nf*	*nf*	*nf*	*nf*	*nf*	*nf*	*nf*
mk	×	×	×	×	×	×	×	×

- Nominative (N): denoting a subject of an action, an event or a state.
- Accusative (A): denoting a direct object of an action, an event or a state.
- Dative (D): denoting an indirect object of an action, an event or a state.
- Genitive (G): denoting an attribute of an entity.
- Predicative (P): denoting an action, an event, a state or an indication.
- Adverbial (B): denoting a property of an action, an event or a state.
- Complementary (C): denoting an expression that helps complete the meaning of an action, an event or a state (i.e., predicate).
- Clausal (S): denoting an introduction to a clausal sentence.

Table 3 provides the usage of finite and non-finite part of speech (reading as row functions as column), in which f stands for finite, nf for non-finite, (nf) for non-finite rarely used, and × for not applied.

Table 3 shows the allowed extrinsic cases of a part of speech to inform how a part of speech finitely functions as a case or non-finitely functions as a case. For example:

- He paints the wall red. (Nn Pvt An Cad, in which n, vt, n and ad represent parts of speech while N, P, A and C represent extrinsic cases.)

In the above examples, a phrase as a meaning group is separated by round brackets or comma. When a noun, an adjective or a preposition to be predicative, it can only be associated with linking verb not common verb. It means that linking verb forms predicate by associating with noun, adjective or prepositional phrase while common verb provides predicate function by itself. A rule is that the same adjacent cases composes the same case disregarding their underlying different parts of speech. This highly simplifies the sentence analysis.

2) INTRINSIC CASE

An intrinsic case denotes a situational usage of a part of speech, which refers to a special property of a part of speech. For example, the number and gender of a noun and an inflected form of a verb. In the defined MUG, only noun (and pronoun) and verb are inflected and have intrinsic cases as listed in Table 4.

Given the extrinsic cases and intrinsic cases, a part of speech is represented in the following form:

$$\text{POS} =: \text{root} + \text{intrinsic case} + \text{extrinsic case}$$

where the intrinsic case tells the actual form of a part of speech and the extrinsic case informs how a part of speech shall combine with other parts of speech to construct a meaning group. It should be noted that while a part of speech is given intrinsic and extrinsic cases, its original property born in a word has not been lost. For example, a ditransitive verb, no matter how intrinsic case and extrinsic case are given, it still requires two objects for the action to enforce upon.

3) CASE PHRASE CONSTRUCTION

In MUG, the case phrase construction is completely different from the existing popular phrase construction and analysis where a phrase is constructed based on the association relationship between parts of speech. A case phrase, in essence, is constructed based on the cases in which parts of speech are recognized. It is a kind of association of cases pertaining to the parts of speech. As described before, there are eight cases in MUG. Case phrases are constructed by associating these eight cases, where a case might have a finite or a non-finite part of speech. To make it clear, the case phrases in terminology are defined in Table 5.

In Table 5, conjunction ci in Nci, Aci, Dci, Gci, Pci, Bci and Cci are intermediate conjunctions that used within a sentence to associate several smaller meaning groups in parallel. The conjunction ci in Sci is the beginning conjunction of a clause to introduce a sentence. All modal verbs in MUG are regarded as adverbs.

C. NOMINATIVE PHRASE (NP)

A nominative phrase is a phrase consisting a word group that functions as the initiator of an action, an event or a state. It controls a verb that is followed. There are several forms of NP.

(1) N =: Nn | Nad | Nvi | Nvd | Nvl | Nnm | Npa

(2) NP =: N

(3) NP =: N, . . . Ncj N

(4) NP =: GP NP (GP is a genitive phrase or a list of genitive phrases)

(5) NP =: Npp [Pvl] (Nominative preposition Npp only appears before predicative linking verb Pvl.)

Rule (4) informs that MUG does not care how a particular genitive phrase is arranged in the list of genitive phrases as long as it is in the GP list. More examples can be found as follows:

TABLE 4. Allowed intrinsic cases.

Part of Speech	Intrinsic Case		Case Representation
	Number	*Gender*	
Noun (n)	Singular (s)	Genderless (a)	-sa
		Masculine (e)	-se
		Feminine (o)	-so
		Neuter (u)	-su
	Plural (p)	Genderless (a)	-pa
		Masculine (e)	-pe
		Feminine (o)	-po
		Neuter (u)	-pu
Verb (v)	*Initiative*	*Tense*	
	Active (a)	Present simple (ce)	-ace
		Past simple (pe)	-ape
		Future simple (fe)	-afe
		Past future simple (ve)	-aye
		Present progressive (co)	-aco
		Past progressive (po)	-apo
		Future progressive (fo)	-afo
		Past future progressive (vo)	-avo
		Present perfect (cu)	-acu
		Past perfect (pu)	-apu
		Future perfect (fu)	-afu
		Past future perfect (vu)	-avu
		Present perfect progressive (cy)	-acy
		Past perfect progressive (py)	-apy
		Future perfect progressive (fy)	-afy
		Past future Perfect progress (vy)	-avy
	Passive (i)	Present simple (ce)	-ice
		Past simple (pe)	-ipe
		Future simple (fe)	-ife
		Past future simple (ve)	-ive
		Present progressive (co)	-ico
		Past progressive (po)	-ipo
		Future progressive (fo)	-ifo
		Past future progressive (vo)	-ivo
		Present perfect (cu)	-icu
		Past perfect (pu)	-ipu
		Future perfect (fu)	-ifu
		Past future perfect (vu)	-ivu
		Present perfect progressive (cy)	-icy
		Past perfect progressive (py)	-ipy
		Future perfect progressive (fy)	-ify
		Past future Perfect progress (vy)	-ivy

TABLE 5. Allowed case phrase in MUG.

POS	N	A	D	G	P	B	C	S
n	f: Nn	f: An	f: Dn	nf: Gn	(nf): Pn	(nf): Bn	×	nf: Sn
ad	nf: Nad	nf: Aad	×	f: Gad	nf: Pad	×	nf: Cad	nf: Sad
vi	nf: Nvi	nf: Avi	×	nf: Gvi	f: Pvi	nf: Bvi	nf: Cvi	nf: Svi
vt	nf: Nvt	nf: Avt	×	nf: Gvt	f: Pvt	nf: Bvt	nf: Cvt	nf: Svt
vd	nf: Nvd	nf: Avd	×	nf: Gvd	f: Pvd	nf: Bvd	nf: Cvd	nf: Svd
vl	nf: Nvl	×	×	nf: Gvl	f: Pvl	nf: Bvl	×	×
av	×	×	×	×	×	f: Bav	×	×
nm	nf: Nnm	nf: Anm	×	nf: Gnm	×	×	×	nf: Snm
pa	nf: Npa	nf: Apa	nf: Dpa	nf: Gpa	×	×	×	nf: Spa
it	×	×	×	×	×	×	×	f: Sit
pp	nf: Npp	nf: App	×	nf: Gpp	×	nf: Bpp	×	×
cj	nf: Nci	nf: Acj	nf: Dcj	nf: Gci	nf: Pcj	nf: Bci	nf: Ccj	nf: S*cj
mk	×	×	×	×	×	×	×	×

- White and black cannot be told. (Nominative adjectives)
- Lying is a bad habit. (Nominative intransitive verb)

In MUG, nominative phrase is flexible for replacing one part of speech with another by following the rules. For example, "white and black" can be interpreted as "Nad Nci Nad" or "Nn Nci Nn" when white and black are nouns.

D. ACCUSATIVE PHRASE (AP)

An accusative phrase is a word group that functions as the direct recipient of an action, an event or a state. It is formed as follows:

(1) A =: An | Aad | Avi | Avt | Avd | Anm | Apa
(2) AP =: A
(3) AP =: A, . . . Acj A | A . . . A
(4) AP =: GP AP
(5) AP =: [Pvl] App (Accusative preposition only appear after a linking verb.)

For example:

- Cat catches mouse. (An)
- It turns red. (Aad)

E. DATIVE PHRASE (DP)

A dative phrase is a word group denoting an indirect recipient of an action.

(1) D =: Dn | Dpa
(2) DP =: D
(3) DP =: D, . . . Dcj D | D . . . D
(4) DP =: GP DP

For example:

- He gives me a book. (Dn)
- He gives John, Mary and I a book for each. (DP)
- He awards the brave John a medal. (GP DP)
- The program assigns X5T a value. (Dpa)

The example X5T informs a rule of MUG such that the nature of dative case of a noun will not be changed no matter how it is located in a sentence. This allows the flexible arrangement of phrase order without altering the meaning of the sentence. This is a key different between MUG and order-dependent natural languages (e.g., SOV sentence pattern for English).

F. GENITIVE PHRASE (GP)

A genitive phrase is a word group that describes the attributes of a noun and takes the function of a finite adjective in linguistics.

(1) G =: Gad | Gn | Gvi | Gvt | Gvd | Gvl | Gnm | Gpa | Gpp
(2) GP =: G
(3) GP =: G, . . . Gcj G
(4) GP =: BP GP

For example:

- Three smiling, brave and handsome men marked with XOU, having guns, and from the capital are walking along the street.

This example can be analyzed as "GP([Gnm]Three [GPad](smiling, brave and handsome)) men GP([Givt] (marked with XOU), [Gvt](having guns), and Gpp(from the capital)) are walking along the street." When it is written in MUG, it is transformed into:Three, marked with XOU, having guns, from the capital, smiling, brave and handsome men are walking along the street.

This form can be translated into any other forms in order as necessary in other natural languages.

G. PREDICATIVE PHRASE (PP)

A predicative phrase is a word group denoting an action, an event or a state.

(1) P =: Pvi | Pvt | Pvd | Pvl | Pn | Pad
(2) PP =: P
(3) PP =: P, P, . . . Pcj P
(4) PP =: BP PP

For example:

- In the pond is a large fish.
- Alice cried and laughed for a movie.

H. ADVERBIAL PHRASE (BP)

An adverbial phrase is a word group that modify an action, an event or a state.

(1) B =: Bav | Bpp | Bvi | Bvt | Bvd | Bvl | Bn
(2) BP =: B
(3) BP =: B, B . . . Bcj B
(4) BP =: BP BP
(5) BP =: SBci

For example:

- He happily accepts the invitation. (Bav)
- He sadly and slowly told the story with a deep sorrow. (Bav)

I. COMPLEMENTARY PHRASE (CP)

A complementary phrase is a word group to complement an object of a verb that cannot give it in accordance with the linguistic rule. The specific case is that a verb attempts to express additional meaning it itself or its object cannot express. The rules are:

(1) C =: Cad | Cvi | Cvt | Cvd
(2) CP =: C
(3) CP =: C, . . . Cci C
(4) CP =: BP CP

For example:

- John painted the wall red. (Cad)

J. CLAUSAL PHRASE (SP)

A clausal phrase is a phrase or clause that can independently exist.

SP =: Sn | Sad | Svi | Svt | Svd | Snm | Spa | Sit | S*ci in which S*ci =*ci + phrase, * = nominative | accusative | adverbial. For example:

- Wa, all friends are here. (Sit)
- 345265t, he spoke loudly. (Spa)

1) MOOD

Mood is a grammatical category to tell the mood of a sentence. It is the third category of a verb besides the voice and tense. In the previous section, we have already described the voice (active and passive) and tense (16 kinds defined).

In MUG, there are five moods in sentence. Thus, five kinds of sentences are defined by an ending mark of a sentence. In particular, each mood is represented by a punctuation mark as follows:

(1) ".": declarative
(2) "!": exclamatory
(3) "?": interrogative
(4) "@": imperative
(5) "*": Conditional

V. SENTENCE AND TRANSLATION

The human languages consist of letters, words, and sentences. Each language has its special script and terminology, for all languages consist of nominal and verbal sentences, and these sentences include a noun which functions as the subject, object or case. Language is generated through the input of words, groups of words and sentences, and in order to produce sentences that represent language the input must be processed, and through the application of the rules, the output will be compatible with the input as in Figure 5.

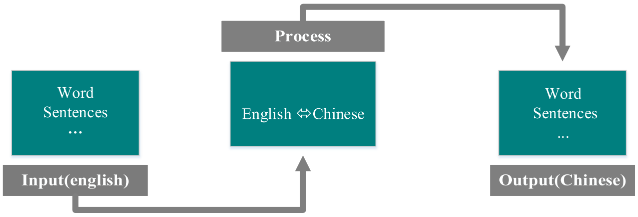


FIGURE 5. Process of sentence translation.

A. LOCAL GRAMMAR

The local mode consists of local to mediation and mediation to local. In this paper, within the MUG structure, Part of Speech (POS) is the foundation, which means that all other structures are built on the basis of POS. In the dictionary, the base and affixes, which suggest tense, voice or plurality, of a term are predefined, among the big group of MUG, original POS in local grammar will be reclassified or redefined. For instance, article in English grammar is traditionally treated as a single POS and appears in Indo-European language family but not in Chinese and MUG, therefore, it needs to join the group of nouns regarded as an adjective. Local grammar in local layer has error analysis that includes rules which are capable of parsing ill-formed input and which apply if the grammatical rules fail. The grammar rule of the unrestricted object construct is augmented by an error analysis rule that checks the inputting sentence against every possible ill-formed construction and produces the appropriate feedback to users.

To enable every rule in local grammar to be fitted into cases of MUG, normalization in the local layer is created by grammar rules (patterns) matching process. We compare each sentence user inputted with an ordered list of patterns, which are regular expressions that can also include additional constraints on case types based on the MUG. These patterns represent local sentence structures that are commonly used to express the various relation types in the local layer, such as "NP is a NP", represent sentence structures that contributors commonly used when entering knowledge as free text. Before a sentence goes through the pattern matching process, local grammars which are grammar errors, special particles, tenses and so on are tagged using the CONEX dictionary and predefined rules. The normalization process tries to solve localized grammar. For example, in the view of grammar, the special particle should be solved, as it has a traditional meaning, as a part of speech that cannot be inflected, and a modern meaning, as a function word associated with another word or phrase to impart meaning. Mandarin Chinese contains a number of grammatical particles. These can have a number of different functions depending on their placement in a sentence. In English, a similar particle exists too. For example:

Sentence 1 (Chinese): *他交流**的**方法是有效的*

*[He communication **DE(的)** method is effective.]*

Sentence 2: *He spoke so well **that** everybody was pleased.*

Therefore, in local mode, the tagging particle distinguishes the kinds of particles function according to their property in the syntax. To normalized particles, the system detects the type of particle then takes further action. There are two types of transformation in the processing of sentence.

Addition (ADD): LABEL words are particles and have no real meaning, such as 的(DE) and 把 (BA). However, in particular languages, some particles are necessary and cannot be neglected, meanwhile, the difference of property of noun and tense should add particles to target language.

Deletion (DEL): This operation is similar with Addiction, in some situations, we need to delete particle or others at first.

Another issue is the developments of a complete tagged local corpus. This corpus would consist of representative number of rules that can be used to evaluate further inputted sentence.

B. MAPPING SENTENCE

The goal and result of mapping are a target-language case phrase, a list of *iid* pairs, whose contents reflect the content of the mediation, expressed in terms of syntactic and lexical properties of the target language. The mapper uses MUG, mapping rules and a CoDic to convert the human language 1 into human language 2. The local document is syntactically analyzed into a MUG document. In our framework, a one-to-one mapping approach achieves semantic unambiguity and reduces complexity crossing languages. The mapping process involves two stages:

Stage 1: *Internal Mapping*. Performing lexical look-up for the lexical entries in order to associate lexemes with case phrase concepts and values. Group the words to construct the order. Such as the order of noun and adjective in a case phrase structure.

Stage 2: *External Mapping*. Determining the case phrase structure. It performs two tasks:

- Use *aspect marker* to determine the sentence mode.
- Use the sentence mode to reorder the sentence to form the case phrase structure.

Internal Mapping: The mapping defines the relation between terms or *iid*, which represents the meaning conveyed in the source text, and the target lexical items. Internal mapping is used to ensure the PoS relations between various elements within the case, internal structure of a case needs to be altered according to the pattern of target language using formed grammar and it gives a new sentence having different word orders. The mapping rules use the mapping lexicon to transform lexical values into the corresponding target words. Figure 6 contains examples of general internal mapping.

External mapping: It is a case phrase structure mapping, and local mapping rules are used to determine the case structure of the sentence. These case phrase representations are used to construct the syntactic structure that will be used to generate a target sentence. The structural mapping rules follow the transformation grammar formalism to order the recognized constituents from constructed case phrase representation that reflects the syntactic structure of the target sentence. They are processed in order and use the pattern shown in Table 6 to map the sentence. The sequence of the iid-value pairs corresponds to the syntactic structure that will be used to generate the target sentence.

Sentence: I need your help.
Map_value (*iid*= 383, i, pronoun, “我”).
Map_value (*iid* =454, need, verb, “需要”).
Map_value (*iid* =432, your, pronoun, “你的”).
Map_value (*iid* =783, help, noun, “帮助”).

FIGURE 6. Examples of general internal mapping.

SI: *She completed her literature review, but she still needs to work on her methods section.*
LS: **[She]** *NP* **[completed]** *PP* **[her literature review]** *AP* [,]*mk* [T-past, M- declarative,V-active] **but** *Ncj* **[she]** *NP* **[still needs to work on]** ***PP*** **[her methods section]** *AP* [.] *mk* [T-present, M- declarative,V-active]
Sentence Mode: [T-past, M- *declarative*, V-*active*]
Map to Chinese: **[她]** *NP* **[完成]** *PP* **[她的文献综述]** *AP* [,] *mk*
Sentence Mode: [T-present, M- declarative, V-active]
Map to Chinese: **但是** *Ncj* **[她]** *NP* **[仍然需要去工作在]** *PP* **[她的方法部分]** *AP* [.] *mk*

FIGURE 7. Simple mapping steps.

TABLE 6. Case phrase structure of the target sentence.

Type of Sentence	Case Phrase Structure	Example of Target English sentence
Statement (.)	NP PP DP AP	*Amy gives me a gift.*
	NP PP AP	*I’m planning a vacation.*
Question (?)	Question marker NP PP AP	*Do you know this application?*

In the following, we describe a mapping example to explain how the mapper maps an English sentence into a target Chinese sentence.

C. TRANSLATION

1) CONTEXT-FREE GRAMMAR

A semantic document is a complex syntactic and semantic phenomena subject to the individual context. Thus, its representation is also complex and contextual, needing to decompose it into clear simple and context types in both syntax and semantics. To realize the goal of context-free document representation that is compatible with most existing information systems, a context-free document syntax is designed and implemented on the base of a sign description theory. The one core of our framework makes the document universal and computer-readable, Context-free Grammar accepts all languages that comply with the word order that begins with the subject such as English while accepts languages that are compatible with all word orders. A CFG has been set up which acts as a language device that has a universal grammar to examine language compatibility based on predefined grammar and knowledge. In this section, we describe the context-free grammar of the word order. The basic idea is that for the sentence f that is to be parsed, we want to create a (monolingual) context-free grammar F that generates strings (f 0) of words in the language that are permutations of the sentence. In the following, the definitions of CFG is introduced.

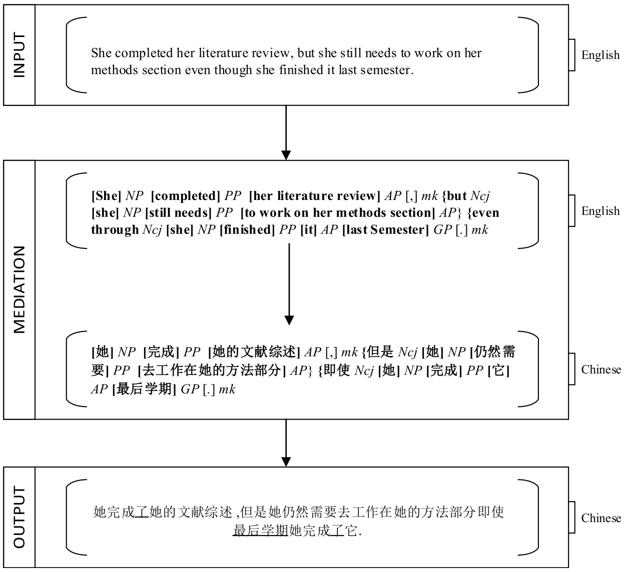
Definitions: A context-free grammar (CFG) is a type of formal grammar, which has the form as follows:

$$G =< \Delta, \sum, S, R >$$

where a set of rules govern how a sentence is produced in a language. With this formal form, G can always produce a rule like $T \rightarrow w$, $T \in \Delta$, and $w \in (\Delta \cup \sum)^*$. It means that T can always be substituted by the word w without needing to consider the occurrence of T in the context. This is why G is called context free grammar (CFG). With this concept in mind, we define a universal context free grammar with the element of $G =< \Delta, \sum, S, R>$ as follows:

- Δ is set of non-terminals (non-terminal symbols) or a finite set of variables. They represent different types of phrases or clauses in a sentence.
- $\sum$ is set of terminals (i.e., a lexicon as a dictionary, e.g., CONEX Dictionary consists of a set of words, where each word has a particular part of speech).
- S is start symbol used to represent the whole sentence (one of the non-terminals).
- R is a relation from Δ to $(\sum \cup \Delta)^*$, which at least exists a set of rules/productions of the form $Y \rightarrow \gamma$, where $Y \in \Delta$ is a nonterminal, and $\gamma \in (\sum \cup \Delta)^*$ is a sequence of terminals (or words in a dictionary) and non-terminals (may be empty).
- A grammar G generates a language L.

Figure 8 presents simple processing that how a source sentence is translated to the target sentence through CFG. Firstly, Inputting document mapping a local document from MUG based on meaning group by predefined rules, all cases in the rearranged sentence undergo IID sequences to make the case and PoS order closer to be universal, the mapping between two languages is possibly performed by using the case as the universal structure, then, the new document maps to target language from internal and external structures until meets target language grammar. After that term by term, translation is performed using the CONEX dictionary.



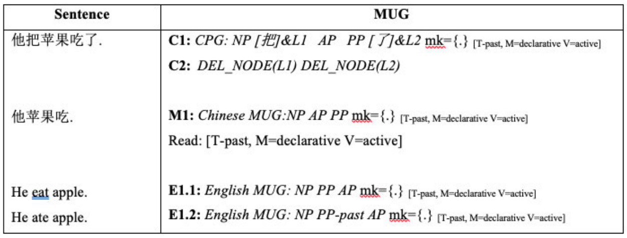
She completed her literature review, but she still needs to work on her methods section even though she finished it last semester.

[She] *NP* **[completed]** *PP* **[her literature review]** *AP* [,] *mk* **{but** *Ncj* **[she]** *NP* **[still needs]** *PP* **[to work on her methods section]** *AP*} **{even through** *Ncj* **[she]** *NP* **[finished]** *PP* **[it]** *AP* **[last Semester]** *GP* [.] *mk*

[她] *NP* **[完成]** *PP* **[她的文献综述]** *AP* [,] *mk* **{但是** *Ncj* **[她]** *NP* **[仍然需要]** *PP* **[去工作在她的方法部分]** *AP*} **{即使** *Ncj* **[她]** *NP* **[完成]** *PP* **[它]** *AP* **[最后学期]** *GP* [.] *mk*

她完成了她的文献综述 ,但是她仍然需要去工作在她的方法部分即使最后学期她完成了它.

FIGURE 8. Translation steps of a sentence.



Sentence	MUG
他把苹果吃了.	**C1:** *CPG: NP [把]&L1 AP PP [了]&L2* mk={.} [T-past, M=declarative V=active]
	C2: *DEL_NODE(L1) DEL_NODE(L2)*
他苹果吃.	**M1:** *Chinese MUG:NP AP PP* mk={.} [T-past, M=declarative V=active]
	Read: [T-past, M=declarative V=active]
He eat apple.	**E1.1:** *English MUG: NP PP AP* mk={.} [T-past, M=declarative V=active]
He ate apple.	**E1.2:** *English MUG: NP PP-past AP* mk={.} [T-past, M=declarative V=active]

FIGURE 9. Sample of translation from Chinese to English.

2) RULES EXAMPLE

Figure 9 describes how sentences work between Chinese to English. First, we need to state that the order in which the rules are written in CPG is important and to achieve input constraint. To translate the sentence, we first generate a case phrase structure by CPG. Within this step, we can also add or delete features. The definite particles "把" and "了" in Chinese which does not have a real feature must be realized as the in target English language. This is achieved by the C2 which L1 and L2 are deleted. Although the English and Chinese sentences follow the "NP PP AP" in general, "BA" sentence in Chinese construct in "NP AP PP", so the sentence needs to transfer into English declarative mode: "NP PP AP" (E1.1). Next, reading the aspect marker knows the tense, voice and sentence type to change features based on the aspects, the verb is transformed to past tense and the best surface structure will be rendered by the target language (E1.2). Below present two examples to illustrate the CF-MUG transfer work.

Sentence 1: *An adhesive activated by ultraviolet secures the sensor housing to the middle bracket.*

After the CPG, we get a syntax tree as shown in Figure 10.

In the English sentence, after PP appears BP- "to" between AP and DP, therefore, when Chinese recognized the structure, it changes to "BA" sentence, the sentence will be transformed by matching rules. After the transformation, we get a new syntax tree in Chinese as shown in Figure 11.

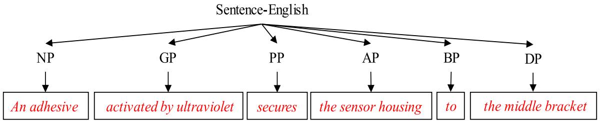


FIGURE 10. Syntax tree of Sentence 1 before reordering.

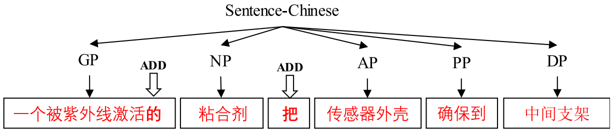


FIGURE 11. Syntax tree of Sentence 1 after reordering.

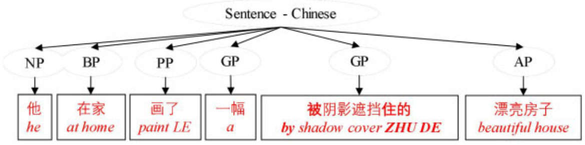


FIGURE 12. Syntax tree of Sentence 2 before reordering.

Transformation: *NP GP PP AP BP DP =>*
*GP **ADD_NODE(L1)(CHN=的)** NP*
***ADD_NODE(L1)(CHN=[把)** AP PP DP*

Sentence 2: *他在家画了一幅**被**阳光遮挡**住的**美丽房子*
(*He at home plain LE a by sun cover LE DE beautiful house.*)

In sentence 2, particles ‘’*LE*’’ and “*DE*’’, verb tense and sentence order should be rearranged and modified. We transfer the syntax tree in Figure 12 to a new syntax tree in Figure 13.

Transformation: ***NP BP PP-***$_{past}$*{PP+LE&L1}* ***GP GP****{BEI NP PP 住&L2 DE& L3}* ***AP*** *{GP AP} =>*
DEL_NODE(L1) (L2) (L3)
NP PP-past AP GP*{ADD_NOTE(L4) (ENG=[that]) PP-*$_{past}$ *ADD_NOTE(L5)(ENG=[by]) ADD_NOTE(L6)(ENG=[the]) NP}* ***BP***.

In all, according to CPG, while a parser takes a string as input and produces case phrase structures as output, a mapper takes a case phrase-structure as input and produces all of the surface strings as output. Of course, the transfer is not as simple as might be conceived from the demo example. There are a great number of structural divergences that must be taken care of. Transfer proves to be considerably convenient as it relieves the grammar writer from worrying about the word order and surface structure in the target language. This confirms the common belief that structural parallelism achieved at the structure level facilitates the translation process.

The basic principle of mapping analysis is to breakdown terms from IID sequences into a structure of case features (lexical category and morphosyntactic properties). Each case rule is responsible for applying a case on given IID sequences to yield an inflected case form. The IID sequences are represented as a basic case structure. Table 7 can be clarified by an example showing the mapping. An example sentence is: “*She completed her literature review, but she still needs to work on her methods section.*”. Based on GC, the sentence structure and syntactic phrase constructions of English can be rewritten as:

*[**She**] NP [**completed**] PP [**her literature review**] AP* [,] *mk*
but *Ncj [**she**] NP [**still needs**] PP [**to work on her methods section**] AP* [.] *mk*

Another verb-based example, we supposed that a sentence just considers subject, objective, verb, then according to the feature of the verb to determine further case shown in Figure 14.

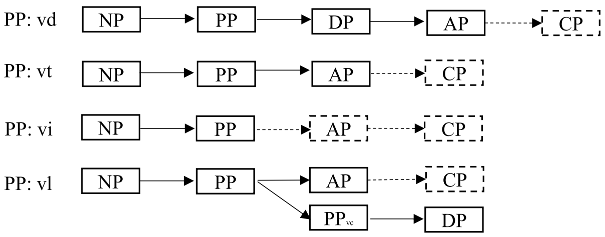


FIGURE 14. Parsing of verb-based sentence.

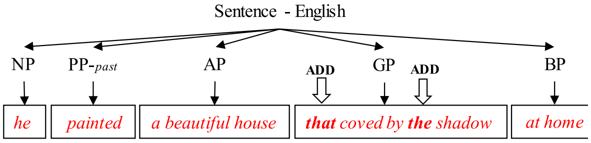


FIGURE 13. Syntax tree of Sentence 2 after reordering.

TABLE 7. Case phrase parsing.

Sentence: *She completed her literature review, but she still needs to work on her methods section.*	
→	**She** *Nn*[a]
→	**She** *Nn* **completed** *Pvt*[a]
→	**[She]** *NP* **completed** *Pvt* **her** *An*[a]
→	**[She]** *NP* **[completed]** *PP* **her** *An* **literature** *An* **review** *An*, *mk*[a]
→	**[She]** *NP* **[completed]** *PP* **[her literature review]** *AP* [,] *mk*
→	**but** *Ncj*[a]
→	**but** *Ncj* [read meaning group→] **she** *Nn* **still** *Bav*[a]
→	**but** *Ncj* [→] **[she]** *NP* **still** *Bav* **needs** *Pvt*
→	**but** *Ncj* [→] **[she]** *NP* **[still]** *BP* **needs** *Pvt* **to** *App*[a]
→	**but** *Ncj* [→] **[she]** *NP* **[still]** *BP* **[needs]** *PP* **to** *App*
→	**but** *Ncj* [→] **[she]** *NP* **[[still]** *BP* **[needs]** *PP***]]** *PP* **to** *App*
→	**but** *Ncj* [→] **[she]** *NP* **[still needs]** *PP* **to** *App* **work** *Avt*[a] **on** *App*
→	**but** *Ncj* [→] **[she]** *NP* **[still needs]** *PP* **[to work]** *AP* **on** *App* **her** *An*
→	**but** *Ncj* [→] **[she]** *NP* **[still needs]** *PP* **[to work]** *AP* **on** *App* **her** *Nn* **methods** *An* **section** *An*. *mk*
→	**but** *Ncj* [→] **[she]** *NP* **[still needs]** *PP* **[to work]** *AP* **[on her methods section]** *AP* [.] *mk*
→	OK!

VI. METHODOLOGY

To describe and prove the validity of our approach in crossing language semantic, this paper introduces the definition of representation for one sentence or plaintext or any sentence-based documents.

A. DEFINITION

A local document (LD) consists of words, words can generate into terms consists of words and phrases, a sentence consists of terms, and hence a document structure will be:

Definition 1: LD = Ti = {T1, T2, T3, ..., Tn}, where T is a set of terms.

T is described as lower representation where IID, POS and Term are all elementary structures. Particularly, while T structure connects a relationship between Chinese and English in "apple", identifier (IID) is a unique identifier, PoS plays a recognizer that when a computer reads each term. For example, the T structure of "apple" can be notated as T [001, noun, "*apple*"]. IID, POS and Term are dependent, where '|' notates a dependence property. T structure is defined as:

Definition 2: T = {IID | POS | Term}

After the operation of the local layer, a universal document (UD) is generated through cases, definition 3 states that a sentence consists of a couple of cases (C) and T structure, a case can contain any T_i and a T can be conveyed in any case structure. A case does not need to imply a term and a term does not need to relate to a particular case. Thus, a case on its own is meaningless, referring to nothing but merely an existence of a construct. A UD where a sentence is sequences of structure and its terms are conveyed in the case.

Definition 3: UD = S_i = (Case, Term) = (C, T), Term ∈ Case, where S is a sentence, C is case.

Any UD's sentences can be generically modeled as a set of cases as Definition 4 because a UD is a set of sentence-based representations. Every sentence is constructed under the case format. The MUG is implemented as collaborative mediation. It takes a centrally managed architecture that fully replicates IID and case structure of different natural languages. The challenge of MUG implementation is the structure consistency maintenance between multiple language structures, where the key issue is to ensure homogeneous case structure, it prevents case concepts generated in S1 of one natural language from being translated as an inequivalent concept in S2 of another natural language.

Definition 4: $C_i = \{C_1, C_2, C_3, \ldots\ldots, Cn\}$

Due to the diversity of languages, the sentence order of sentences in different languages is different. The paper adopts the case structure and allows a relationship among cases. A definition is as follows:

Definition 5: Case associative relation (CAR): C1=: C1 C2 for C1, C2 ∈ C

A case phrase can be combined with another case Phrase and becomes a new case phrase. For example, the following shows some transfers between case phrases, for example, a GP meets a NP, GP and NP become NP. We allowed the CAR between Chinese and English as follows.

NP =: GP NP AP =: GP AP DP =: GP DP

GP =: BP GP PP =: BP PP

Meanwhile, there exists a non-associative relationship among cases. If the system recognizes the relationship, the sentence does not hold.

Definition 6: Case non-associative relationship (CNAR): C1 ≠ (C1 C2) for C1, C2 ∈ C

When the sentence is mapped, the position of the cases changes. Therefore, controlling the position of cases ensures the validity of the sentence.

Definition 7: Case commutive relationship (CCR): (C1 *C2) = (C2 *C1) for C1, C2 ∈ C

For example:

(GP*NP) = (NP*GP); (PP *BP) = (PP *BP);

(GP *AP) = (AP *GP); (GP *DP) = (DP *GP).

Definition 8: Case non-commutive relationship (CNCR): (C1 *C2) ≠ (C2 *C1) for C1, C2 ∈C

For example:

(NP *PP) ≠ (PP *NP); (AP *PP) ≠ (PP *AP).

B. CONDITION

We have discussed the basic definitions. In the real world, documents are complex and heterogeneous. A heterogeneous document is a document that has a different way of constructing an equivalence in both structure and semantic, compared with others. The heterogeneous document exists because of different semantic communities, which provide different contexts for document representations. This section will represent heterogeneous document. When contexts are involved, a document representation will have the following form:

Condition 1 (Structure Equivalence): Given two crossing language sentences of S1 [$C1_i$ [$T1_i$ [IID, POS, Term]]] and S2 [$C2_i$ [$T2_i$ [IID, POS, Term]]] through definitions, then S1 and S2 are structure-consistent if and only if:

(1) $T1_i \in C1_i \in$S1, and $T2_i \in C2_i \in$ S2;
(2) TID is a unique identifier of T such that T (IID, POS, Term)→IID;
(3) There exists a Mapping Relationship (MP): IID ∈ S1 ⇔ IID ∈ S2;
(4) C is a universal structure, there exists a Mapping Relationship (MP): $C1_i \Leftrightarrow C2_i$ (structurally).
(5) Meet definition 5, 6, 7 and 8.

Generically, Condition 1 achieves structure consistency by converging all heterogeneous structures onto an isomorphic structure Map ($C_1, \ldots, C_n$) and Map ($IID_1, \ldots, IID_n$), For example, given the heterogeneous structures, if IID1=[001, 002,003,004] from context 1 and IID2=[004,003,001,002] from context 2, meanwhile there are universal Case structure as mediation for both languages, C1=[NP,BP,CP,DP,...] and C2 =[BP,CP,NP,DP,...], then heterogeneous S1 and S2 are structurally consistent on Map(IID1 ∈ S1, IID2 ∈ S2),Map(C1 ∈ S1, C2 ∈ S2), Map(IID1⇔IID2) and Map(C1⇔C2).

Moreover, how does "apple", "epal" and "苹果" in different languages maintain the same semantics? In this paper, the heterogeneous semantic integration condition described below is a condition of achieving semantic substitutability.

Condition 2 (Semantic Equivalence): Given two crossing language sentences of S1 [$C1_i$ [$T1_i$ [IID, POS, Term]]] and S2 [C2i [T2i [IID, POS, Term]]] through the definitions, then S1 and S2 are semantic-consistent if and only if:

(1) Term is unique concept of T such at T (IID, POS, Term)→Term;
(2) IID ⇐ Term;
(3) Terms ∈ S1 ⇔ Terms ∈ S1 (semantically).

Condition 2 guarantees that two heterogeneous semantics are semantically consistent. For example, the semantic consistency among concept 1 [IID=002, Pos=Noun, Term= "apple"], concept 2 [IID=002, Pos=Noun, Term= "epal"], and concept 3 [IID=002, Pos=Noun, Term= "苹果"] could be guaranteed if and only if we could guarantee that "apple", "epal" and "苹果" are semantically the same.

In this paper, we employ a collaboration mechanism, available to all different contexts. Through a universal grammar in the mediation layer, heterogeneous structures and contexts could be collaboratively mediated. The common context achieved by collaboration mechanism is referred to as heterogeneous context integration condition as follows:

Condition 3 (Context Equivalence): Given two crossing language context Doc1 of $S1_i$ [$C1_i$ [$T1_i$ [IID, POS, Term]]] and Doc2 of $S2_i$ [$C2_i$ [$T2_i$ [IID, POS, Term]]], then LD1 and LD2 are context-consistent if and only if:

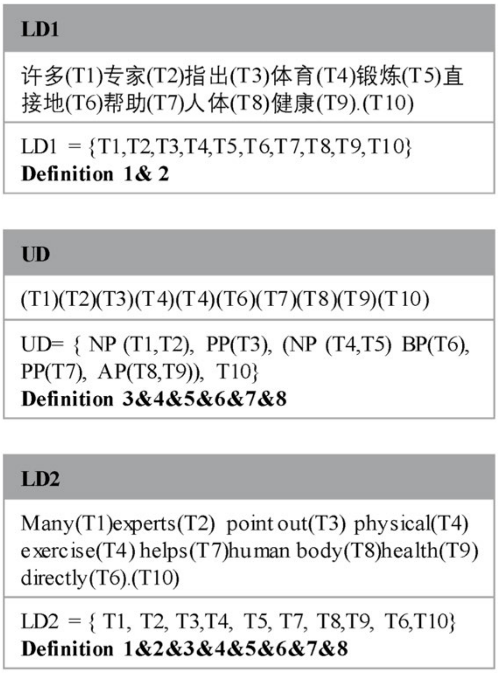
(1) Both condition 1 and condition 2 are satisfied.
(2) $T1_i \in S1_i \in$ LD1 and $T1_i \in S2_i \in$ LD2.

For example, if Doc 1 designer can work with Doc 2 designers on a collaborative system to negotiate that 'apple' and are semantically equivalent under the constraints of 3 conditions above, the concept of 'apple' and "苹果" can be semantically consistent. Meanwhile, any machine-based reconciliation between two heterogeneous documents for structure and concept mapping is important and sufficient. This is because machines can only infer representation equivalence through pre-encoded rules.

Figure 15 shows an example of justification of LD1 and LD2, from the set of LD1 and LD2, we found both of LD1 and LD2 have similar elements of Ti, and UD presents a universal case to compatible with LD1 and LD2. Therefore, LD1 is structure equivalence to LD2. Meanwhile, comparing the translation result of LD1 and LD2, LD1 and LD2 are semantic equivalence.

VII. EVALUATION

In order to research the acceptance of translation for users, the evaluation of acceptance adapts a questionnaire analysis in different translation tools among machine translation tools (Google, Bing, and Baidu) and our proposed CF-MUG. To assess the acceptability of translation, currently, there are two feasible approaches - automatic evaluation and human evaluation, BLEU is the most common machine evaluation technique. Automatic evaluation-BLEU [28] is the most common machine evaluation tech- nique and only applicable within machine translation through the strong correlation between human judgments, and shows how many words are shared between MT output and human-made reference, benefiting sequential words. We randomly pick up 50 Chinese sentences (N=50) and 50 English sentences (N=50) and evaluate our MNLM using the standard BLEU score metric and to make our results comparable to Google, Bing and Baidu translation tools.



LD1

许多(T1)专家(T2)指出(T3)体育(T4)锻炼(T5)直接地(T6)帮助(T7)人体(T8)健康(T9).(T10)

LD1 = {T1,T2,T3,T4,T5,T6,T7,T8,T9,T10}
Definition 1& 2

UD

(T1)(T2)(T3)(T4)(T4)(T6)(T7)(T8)(T9)(T10)

UD= { NP (T1,T2), PP(T3), (NP (T4,T5) BP(T6), PP(T7), AP(T8,T9)), T10}
Definition 3&4&5&6&7&8

LD2

Many(T1)experts(T2) point out(T3) physical(T4) exercise(T4) helps(T7)human body(T8)health(T9) directly(T6).(T10)

LD2 = { T1, T2, T3,T4, T5, T7, T8,T9, T6,T10}
Definition 1&2&3&4&5&6&7&8

FIGURE 15. Justification from LD1 to LD2.

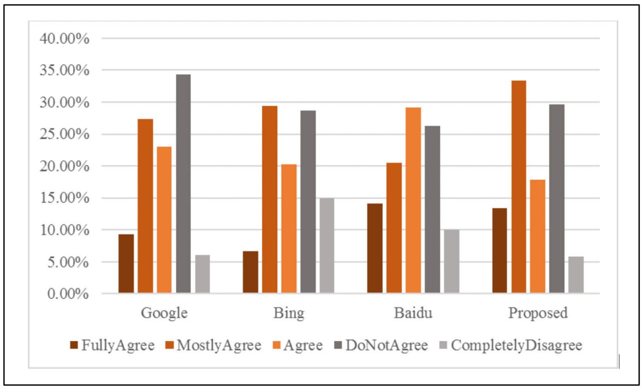


FIGURE 16. Acceptance of translation from English to Chinese.

TABLE 8. BLEU sores among four tools.

Translation tools	En-Ch	Ch-En
Bing	0.33	0.43
Baidu	0.31	0.37
Google	0.34	0.36
CF-MUG	0.43	0.42

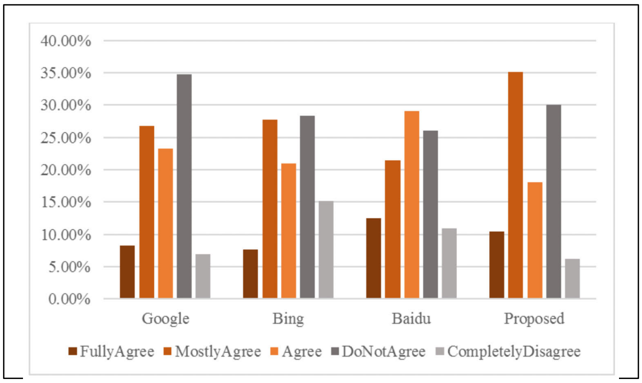


FIGURE 17. Acceptance of translation from Chinese to English.

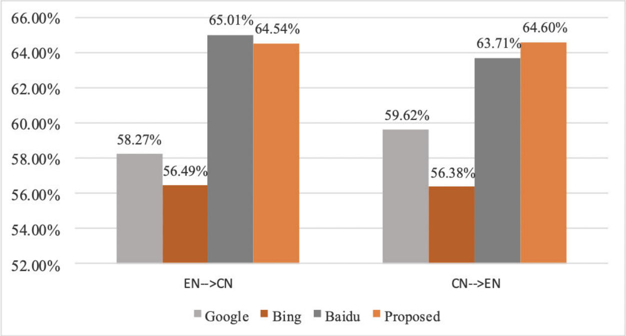


FIGURE 18. Percentage of the acceptable translation result.

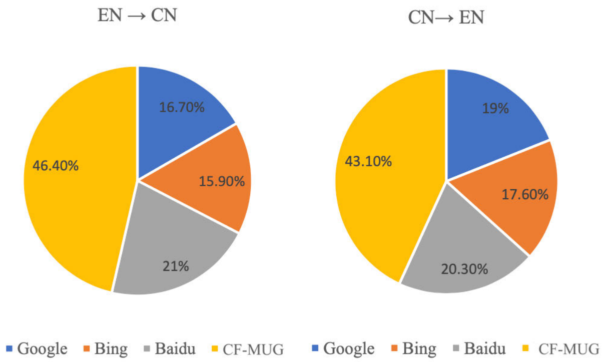


FIGURE 19. Percentage of the semantic accuracy result.

From the result of Table 8, it shows a comparison of the average scores of BLEU, but there is no obvious and better BLEU score for MNLM compared with other methods. The reasons under three issues of BLEU method are: (1) It does not consider meaning. (2) It does not directly consider sentence structure, and (3) It does not map well to human judgments. Therefore, from the view of the semantic or meaning sentences, human semantic evaluation method plays a more important role in the acceptance and semantic accuracy of translation.

In this paper, the questionnaire is used for human evaluation to analysis the human judgments in acceptance based on the assigned questionnaire. The questionnaire collected sentences between Chinese and English randomly from websites. To ensure the validity of the data, approximately 450 valid data support our analysis. The evaluators were asked to consider each machine translated outputs from four selections which are Google, Bing, Baidu, and our proposed method in *Fully Agree, Mostly Agree, Agree, Don't Agree* and *Completely Disagree*. Formula (1) indicates how to calculate the percentage of acceptance:

$$Acceptance(A_t) = \frac{\sum_{i=1}^{8}(Nr_i/N_t)}{S} \times 100\% \tag{1}$$

where A indicates percentages of acceptance in a translation tool, t indicates the Likert scale, t = [Fully Agree | Mostly Agree | Agree | Do not Agree | Completely Disagree], i indicates the number of sentences, Nri indicates the number of selecting scale in the t, Nt is the number of total evaluators, and S indicates the sum number of sentences. Figs. 16 and 17 show the results of acceptance between Chinese and English sentences through Microsoft Excel data analysis tool using formula (1).

From Figure 16 and 17, comparison of the percentage of acceptance between the CF-MUG and machine translation tools indicate that four MT tools perform a better result as *CompletelyDisagree* and *FullyAgree* have the smallest percentage in overall. Translation results of the two languages are not much different but the google translation is higher in *DoNotAgree*, accounting for about 34%. If we only consider the *Agree* and the above are successful sentences as shown in Figure 18, then, Google and Bing translations are relatively lower than others. In general, Baidu translation is more acceptable in Chinese-English translation among machine translation tools.

BLEU scores and acceptance of translation are not good at evaluation in semantic accuracy, therefore, to better understand the accuracy of semantics, we made another questionnaire - semantic accuracy of translation. We randomly selected 30 English sentences and 30 Chinese sentences from websites to conduct a survey on 321 valid (N = 321) Chinese native participants as the survey focuses on Chinese and English, and we told the participants a basic knowledge of semantic and its accuracy to make better judgments. The participants were asked to consider each machine translated outputs from four tools which are Google, Bing, Baidu, and our proposed CF-MUG. If the questions were identical to be extremely accurate, they would select it.

From Figure 19, comparisons of the percentage of *semantic accuracy* between the CF-MUG and machine translation tools indicate that the CF-MUG performances more significantly accuracy in both translations- English to Chinese and Chinese to English which is 46.4% and 43.1% respectively. Especially, within machine translation tools, Baidu tool performances better result and the lowest semantic accuracy is the Bing tool.

Through the survey, we validated translation results based on questionnaires but there are two limitations:

1) The number of languages examined was significantly limited. Here, acceptance is made by MT tools between English and Chinese. It is possible that if this same study was carried out translating other languages, the results may be different. 2) In the current study, the translated sentence in our proposed CF-MUG is manually generated based on well-defined rules and grammars, and it probably makes a slight difference in real.

VIII. DISCUSSION AND CONCLUSION

In the document exchange, the accuracy of translation and disambiguation in the exchange of semantic documents across different languages is critical and ensures that senders and receivers understand the meaning of the exchanged documents. A large amount of text available online is becoming more and more linguistic, providing more useful information. However, because language barriers prevent cross-language searches, most users do not have easy access to most of this information. Due to the ambiguity of the language, the use of synonyms to express a single idea creates problems. To realize semantic unambiguity and accuracy in crossing document exchange and natural language processing, this paper designs a new framework of CF-MUG, CF-MUG draws on the advantages of the collaboration concept, due to each word tags a unique ID, document exchange transfers to ID exchange, this method not only ensures computer-readable, understandable and further automatically executable in Artificial Intelligence but also minimizes the information lack from the source language to the target language.

Based on the grammatical case, this paper proposed MUG grammar, also be called universal grammar MUG that accepts all coming languages. MUG aims to reduce the complexity of mapping compared with existing methods through the third-party language translation. So far, we have initially proposed the theoretical framework. In the future, we need to improve and implement the framework.

REFERENCES

[1] G. Xiao, J. Guo, Z. Gong, and R. Li, "Semantic document exchange for E-business: Trends and issues," in *Proc. IEEE 12th Int. Conf. e-Bus. Eng.*, Oct. 2015, pp. 147–153.

[2] S. Bergamaschi, S. Castano, M. Vincini, and D. Beneventano, "Semantic integration of heterogeneous information sources," *Data Knowl. Eng.*, vol. 36, no. 3, pp. 215–249, Mar. 2001.

[3] L. A. Zadeh, "A key issue of semantics of information," in *Proc. IEEE 15th Int. Conf. Cognit. Informat. Cognit. Comput. (ICCI CC)*, Aug. 2016, p. 1.

[4] J. Guo, "Collaborative conceptualisation: Towards a conceptual foundation of interoperable electronic product catalogue system design," *Enterprise Inf. Syst.*, vol. 3, no. 1, pp. 59–94, Feb. 2009.

[5] V. Macketanz, E. Avramidis, A. Burchardt, J. Helcl, and A. Srivastava, "Machine translation: Phrase-based, rule-based and neural approaches with linguistic evaluation," *Cybern. Inf. Technol.*, vol. 17, no. 2, pp. 28–43, Jun. 2017.

[6] M. Costa-Jussa, M. Farrús, J. B. Marino, and J. Fonollosa, "Study and comparison of rule-based and statistical Catalan-Spanish machine translation systems," *Comput. Informat.*, vol. 31, no. 2, pp. 245–270, 2012.

[7] A. Barreiro, B. Scott, W. Kasper, and B. Kiefer, "OpenLogos machine translation: Philosophy, model, resources and customization," *Mach. Transl.* vol. 25, no. 2, p. 107, 2011.

[8] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*. [Online]. Available: http://arxiv.org/abs/1609.08144

[9] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Torat, F. Viégas, M. Wattenberg, G. Corrado, and M. Hughes, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Trans. Assoc. Comput. Linguistics* vol. 5, pp. 339–351, Oct. 2017.

[10] S. Bergamaschi, S. Castano, M. Vincini, and D. Beneventano, "Semantic integration of heterogeneous information sources," *Data Knowl. Eng.*, vol. 36, no. 3, pp. 215–249, Mar. 2001.

[11] B. J. Dorr, P. W. Jordan, and J. W. Benoit, "A survey of current paradigms in machine translation," *Adv. Comput.*, vol. 49, pp. 1–68, 1999.

[12] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," *J. Emerg. Technol. Web Intell.*, vol. 2, no. 3, pp. 258–268, Aug. 2010.

[13] J. Hollan, E. Hutchins, and D. Kirsh, "Distributed cognition: Toward a new foundation for human-computer interaction research," *ACM Trans. Comput.-Hum. Interact.*, vol. 7, no. 2, pp. 174–196, Jun. 2000.

[14] R. P. Cohen, "EDI basics," GXS Washington Blvd, Gaithersburg, MD, USA, Tech. Rep., 2013.

[15] *W3C XML Schema–World Wide Web Consortium*. [Online]. Available: http://www.w3.org/XML/Schema

[16] A. Constantin, S. Peroni, S. Pettifer, D. Shotton, and F. Vitali, "The document components ontology (DoCO)," *Semantic Web*, vol. 7, no. 2, pp. 167–181, Feb. 2016.

[17] B. McBride, "The resource description framework (RDF) and its vocabulary description language RDFS," in *Handbook on Ontologies*. Berlin, Germany: Springer, 2004, pp. 51–65.

[18] *International Handbooks on Information Systems*. Berlin, Germany: Springer.

[19] G. Antoniou and F. Van Harmelen, "Web ontology language: Owl," in *Handbook on Ontologies*. Berlin, Germany: Springer, 2004, pp. 67–92.

[20] J. Guo, I. H. Lam, C. Chan, and G. Xiao, "Collaboratively maintaining semantic consistency of heterogeneous concepts towards a common concept set," in *Proc. 2nd ACM SIGCHI Symp. Eng. Interact. Comput. Syst. (EICS)*, 2010, pp. 213–218.

[21] J. Guo, "SDF: A sign description framework for cross-context information resource representation and interchange," in *Proc. Enterprise Syst. Conf.*, Aug. 2014, pp. 255–260.

[22] J. Guo and C. Sun, "Context representation, transformation and comparison for ad hoc product data exchange," in *Proc. ACM Symp. Document Eng. (DocEng)*, 2003, pp. 121–130/

[23] G. Xiao, J. Guo, and Z. Gong, "Syntactic file and semantic file alignment for E-Business document editing and exchange," in *Proc. IEEE 10th Int. Conf. e-Bus. Eng.*, Sep. 2013, pp. 112–117.

[24] G. Xiao, "Semantic document exchange for electronic business through user-autonomous document sense-making," Ph.D. dissertation, Univ. Macau, Zhuhai, China, 2015.

[25] V. J. Cook, "Chomsky's universal grammar and second language learning," *Appl. Linguistics*, vol. 6, no. 1, pp. 2–18, 1985.

[26] J. Lidz, "The grammar of accusative case in Kannada," *Language*, vol. 82, no. 1, pp. 10–32, 2006.

[27] L. Yager, N. Hellmold, H. A. Joo, M. T. Putnam, E. Rossi, C. Stafford, and J. Salmons, "New structural patterns in moribund grammar: Case marking in heritage German," *Frontiers Psychol.* vol. 6, p. 1716, Nov. 2015.

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Philadelphia, PA, USA, 2001, pp. 311–318.

QUANYI HU received the master's degree in electronic and computer engineering from the University of Macau, Macau, China, in 2017, where he is currently pursuing the Ph.D. degree in computer science with the Department of Computer and Information Science. His research interests include indoor localization, blockchain, and deep learning.

JIE YANG is currently pursuing the Ph.D. degree with the University of Macau. He is also a Lecturer with Chongqing Industry and Trade Polytechnic, China. He has published several papers in reputable journals and conferences. His research interests include machine learning, biomedicine, and evolutionary computing.

PENG QIN (Graduate Student Member, IEEE) received the master's degree in e-commerce technology from the University of Macau, Macau, China, in 2017, where he is currently pursuing the Ph.D. degree in computer science with the Department of Computer and Information Science. His research interests include semantic document exchange, semantic integration, and e-commerce.

SIMON FONG (Member, IEEE) received the B.Eng. degree (Hons.) in computer systems and the Ph.D. degree in computer science from La Trobe University, Bundoora, VIC, Australia, in 1993 and 1998, respectively. He held various managerial and technical posts, such as a Systems Engineer, an IT Consultant, and an E-Commerce Director in Australia and Asia. He is currently an Associate Professor with the Computer and Information Science Department, University of Macau, Macau, China, and an Adjunct Professor with the Faculty of Informatics, Durban University of Technology, Durban, South Africa. He is a Co-Founder with the Data Analytics and Collaborative Computing Research Group, Faculty of Science and Technology. He has published over 450 international conference papers and peer-reviewed journal articles in data mining, data stream mining, big data analytics, and metaheuristics optimization algorithms and their applications. He is an Active Researcher with leading positions, such as the Vice-Chair of the IEEE Computational Intelligence Society Task Force on Business Intelligence and Knowledge Management, the TC Chair of the IEEE ComSoc E-Health SIG, and the Vice-Director of the International Consortium for Optimization and Modeling in Science and Industry. He serves on the editorial boards of *Journal of Network and Computer Applications* (Elsevier), *IT Professional* magazine (the IEEE), and various special issues of SCIE-indexed journals.