Advanced Engineering Informatics xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

A model based transformation paradigm for cross-language collaborations

Kunmei Wen [a,*], Suo Tan [b], Jie Wang [a], Ruixuan Li [a], Yuan Gao [a]

[a] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China
[b] Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada H3G 1W8

ARTICLE INFO

Article history:
Received 8 March 2012
Received in revised form 11 October 2012
Accepted 22 October 2012
Available online xxxx

Keywords:
Cross-language transformation
Online collaboration
Recursive object model
Model-based paradigm
CLT-ROM

ABSTRACT

Online collaboration is a big challenge in the field of international product development in a cross-language environment. It serves two purposes: cross-language translation and design requirement clarification. Though many approaches and tools are developed for each of the purposes, not a solution serves both of them well. Especially, the traditional statistical methods for cross-language translation cannot preserve the whole semantic information, which intend to incur misunderstanding and ineffective collaboration. This results in potential problems in clarifying the design requirements. In this paper, we proposed a method to online collaboration, named Cross-Language Transformation based on Recursive Object Model (CLT-ROM). The proposed method consists of two steps. Firstly, a natural language sentence is transformed into a source ROM diagram. Secondly, a corresponding target ROM diagram is generated by a transformation algorithm. The proposed method is a model-based communication tool which facilitates collaborations. Since the ROM has been proven effective in requirements clarification, some examples are given to illustrate that the CLT-ROM has a good capability of semantic preserving in requirement engineering for product development.

1. Introduction

With the gradual integration of the world economy, communication and cooperation are playing an increasing role for enterprisers in achieving their business goals effectively. For those enterprisers who expand their business into global markets through international collaborations in product development, cross-language and cross-cultural challenges emerge. These challenges may prevent the business partners from understanding each other, and result in ineffective collaborations that harm their business. Though many researchers have put much effort into finding solutions [1–3], not a solution satisfied the common needs for international engineering collaborations. Large enterprisers are able to face the challenges since they have enough resources to deal with. But this may not be the case for small and medium enterprisers (SMEs) in developing and underdeveloped countries.

Let us consider such a scenario: an English-speaking company and a Chinese-speaking company are collaborating on rapid product development in multi-disciplinary design situations. Employees from both companies are having difficulties in expressing themselves clearly in a language other than their mother tones. For such a case, the very basic activities in product development that encounter cross-language barrier can be extracted as real-time tasks (for example, video and telephone conference, instant message, face-to-face online discussion) and static tasks (such as document translation and email correspondence). Obviously, once a solution to resolving the real-time tasks is found, static tasks can be resolved without a doubt. Furthermore, if we could use conference calls over face-to-face discussion, we could save a large amount of money. And thus, the solution must be efficient for the collaboration. If other real-time tasks apart from face-to-face online discussion are renamed as online collaboration, the problems in cross-language collaborations can be simplified as a single critical one – online collaboration.

To break the barrier, enterprisers have to hiring experts as interpreters and/or purchase software kits (language tool). However, for interpreters who well know both languages may do not have domain specific knowledge about the design concepts. Therefore they cannot perform documents translations and real-time interpretations. Even though perfect interpreters are available, they are always too expensive for SMEs to hire. For software kits, they are less expensive compared to experts in the long run. Unfortunately, current software kits still offer poor translations in engineering applications. A substantial demand of a more reliable solution in engineering applications is increasing for SMEs.

Before giving a solution, we must understand the purpose of online collaboration for cross-language collaborations in product development. Based on the previous analysis, online collaboration serves two objectives: cross-language translation and design requirements clarification. The first purpose is to make a translated sentence readable in semantics. The other one is to ensure the

* Corresponding author. Tel.: +86 27 87544285.
E-mail address: kmwen@hust.edu.cn (K. Wen).

1474-0346/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aei.2012.10.007

translated sentences reasonable and meaningful for communication, especially for engineering applications. They are in two different levels. Although human translators are still widely used in cross-language collaborations, in this paper, human translation is not considered.

Computational linguistics is an interdisciplinary field dealing with natural language modeling from a computational perspective. Natural Language Processing (NLP) is also an interdisciplinary field of computer science, artificial intelligence, and linguistics, which focuses on the interactions between computers and human natural languages. Sometimes, computational linguistics and NLP can be treated as a same concept. Machine Translation (MT) is a sub-field of computational linguistics or NLP that investigates the automatic translation of text or speech from one natural language to another. Their common goal is to enhance human–machine communication in natural languages. Language translation, also referred as MT [4], is the most commonly used method to implement cross-language collaborations by transforming a source sentence to a target sentence, for example, transforming an English sentence to a Chinese sentence. However, language translation maintains a challenge despite of the efforts in MT. Semantic integrity is a basic standard for cross-language engineering applications. If the semantic information included in natural languages is lost, the translation results will be incorrect. As a result, the design product cannot satisfy the users' requirements. Whether designers can clearly understand customers' requirements or not largely determines the practicability and efficiency of the design results [5]. Therefore, the semantics should be preserved in the process of cross-language translation. There are two important features of semantics: constant and imparity.

Constant: The linguistic relation between words of a sentence tends to be constant.
Imparity: The semantics that different words carry are imparity.

Traditional language translation methods cannot resolve the semantic preserving problem, in other words, the constant and imparity semantics are lost. The ambiguity is also unavoidable. Since the traditional methods cannot well preserve the semantics in translation, it motivates us to propose a different paradigm for cross-language collaborations. We attempted to preserve the semantics included in natural language by cross-language model transformation.

In this paper, we propose a method called Cross-Language Transformation based on Recursive Object Model (CLT-ROM) for the cross-language online collaboration issue, in order to improve the effectiveness and efficiency of global collaborations. The contributions are as follows: (1) a model-based transformation paradigm is proposed for cross-language collaborations. (2) A prototype of the model-based transformation method is implemented. (3) Examples are provided to illustrate the feasibility of the method.

The rest of the paper is organized as follows. Section 2 interprets the related work. In Section 3, we present a brief introduction to the ROM. The CLT-ROM method is proposed in Section 4. Case study and experiments are provided in Section 5 to illustrate the effectiveness of the proposed method. Finally, we give the conclusions and the future work in Section 6.

2. Problem analysis and related work

A natural language is a typical tool used for communication on daily basis. However, it is not the only option. As shown in Fig. 1, diverse ways are used to describe the communication universe. For instance, symbol and mathematical expressions are neat in science. A model language, Unified Modeling Language (UML) for example, is used for structural representation of objects in software engineering. Apparently, communication can be natural-language-independent. This inspired us that if an appropriate model with primitive symbols is generated to carry semantics of natural language and to represent the relationships of language elements intuitively, perhaps a novel paradigm could be proposed for cross-language collaboration. The Recursive Object Model (ROM) [6] is such a graph-based approach used to model the syntactic and semantic relations between words in a sentence. It can preserve the semantics of natural languages. The constant feature of the linguistic relation between the words can be kept by the predefined elements. Meanwhile, the different relations defined in the ROM can be well used to distinguish the imparity semantics according to the restricted semantics and non-restricted semantics. Therefore, the correct model of a sentence is unique. And ambiguity issue is reduced or eliminated. A detailed introduction to the ROM is given in Section 3. In spite of language transformation issue, the ROM has been proven very effective in requirements gathering in various engineering applications [7–11], such as eliciting and refining detailed requirements by asking questions [7], and specifying the requirements into the current widespread UML diagram [8]. Thus, a tool based on the ROM is able to fulfill our goal in online collaboration. Back to the aforementioned scenario, if such a model-based method is used for online collaboration, we do not need direct translation between different natural languages. We only need to transform requirement sentences, in either English or Chinese, into models in another language, since the models are carrying all information and are intuitive for understanding.

As shown in Fig. 2, language translation methods translate a source sentence into a target sentence directly. In comparison, model-based transformation method, shown in Fig. 3, transforms a sentence into a source model first. And then the source model is transformed into a target model. The main differences between the traditional language translation method and our model-based transformation method can be summarized as follows: firstly, traditional translation is based on statistical methods by translating one sentence into another language. Our transformation method transforms one sentence into its ROM diagrams in different languages. It provides a new paradigm. Secondly, since the CLT-ROM method is based on the ROM, it can preserve the semantics, while the traditional method may lose some, especially the relations among words.

The aim of cross-language translation is simple: given a source sentence, its equivalent target sentence is to be generated. However, implementation of language translation is complex as inherent ambiguity. In other words, a single word has different meanings with different contexts and idioms. Therefore, without considering the differences between the source language and the target language, translating a source sentence merely into a word sequence of a target language cannot implement a good translation. The word sequence must be reordered before generating the target language sentence to ensure a high quality translation. As a matter of fact, reordering the translated words to fit the logic of a target language is one of the most challenging problems in MT.

When solving the translation problems, language translation is divided into the two categories: Rule-Based Machine Translation (RBMT) and Corpus-Based Machine Translation (CBMT) [12]. The main ideas of RBMT are based on linking the structure of a given input sentence with the structure of a demanded output sentence, necessarily preserving their unique meaning. The RBMT is a large-scale and time-consuming method. It is hard to improve the translation performance because adding a new rule into a large-scale rule set is difficult to implement. Some linguistic information still needs to be set manually, resulting in a huge and changeful rule base. It is too hard to guarantee the consistency of the rules in

Please cite this article in press as: K. Wen et al., A model based transformation paradigm for cross-language collaborations, Adv. Eng. Informat. (2012), http://dx.doi.org/10.1016/j.aei.2012.10.007

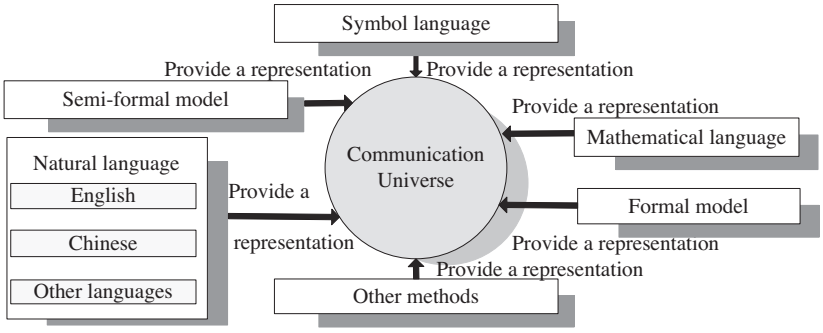


Fig. 1. Methods of describing the communication universe.

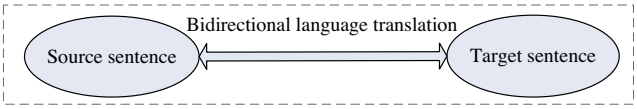


Fig. 2. Language translation.

big systems. CBMT involved to Example-Based Machine Translation (EBMT) and Statistical Machine Translation (SMT). Most of the efforts were made on the SMT methods as the EBMT methods [13] build a base to store translation examples first.

The SMT were evolved in traditional word-and-phrase-based model and syntax model. Brown et al. [14] proposed a word-based SMT method by using a noisy channel model. The method was an important work for the SMT methods. The model was designed to model the lexical dependencies between the words of sentences. However, this method did not resolve the reordering problem. The phrase-based models [15] were introduced to remedy the reordering model. Target language strings were broken into phrases which are independently translated as blocks. These translated phrases were reordered to produce the source sentence according to a distortion probability. Talbot et al. [16] designed a lightweight evaluation framework for language translation reordering. Reordering models involved some different models, such as fundamental distance-based distortion model [16,17], flat reordering model [18,19], lexicalized reordering model [20], hierarchical phrased model [21] and maximum entropy-based phrase reordering model [22]. These methods cannot carry out the complex reordering operations including long distance dependencies and variable contexts. To obtain a global reordering result, the syntax-based SMT methods were developed to capture the syntactic knowledge by simply swapping the children nodes of a parse tree [23]. These methods totally depend on the syntax structure to complete translation [24–26]. These syntax-based SMT methods partly resolve the local reordering problem. However, the reordering results cannot be consistent with syntactic structures. Furthermore, a syntax-based SMT model is more complex than a phrased-based SMT model.

Overall, the methods mentioned above brought some improvements. However, in order to produce a good translation for a source sentence, the existing methods must balance the translated effect and the reordering consequence. It brings a problem: these methods cannot fully preserve the original semantics included in natural language. Based on different classification methods, there are several kinds of semantic information, such as humor semantics [27] and ontology semantics [28]. Correct and complete semantic information plays a significant role in cross-language understanding. Focusing on the constant feature of the linguistic relation between words and the imparity feature of semantics, we proposed the CLT-ROM method. The differences between the traditional methods and the proposed transformation method can be summarized in Table 1, for an English to Chinese translation case.

3. Introduction to the ROM

The ROM [6] is a graphic model that carries the semantic information included in natural language. It is a tool used to analyze natural languages, especially to extract major ontological words. It uses two primitives, object and relation, to capture the linguistic structure of a natural language (shown in Table 2). The objects are further classified into object and compound object, and relations are characterized into constraint, connection, and predicate. These five elements are derived based on strong mathematical foundations including set theory, mathematical relations and logic. The primitives are proven sufficient for technical English through enumeration [6]. Every word and phrase in sentences can be treated as an object and a compound object, respectively. Some applications have demonstrated the usefulness of the ROM in engineering design [8,10,11,29,30].

We use the notation R(A, B) to denote the representations of ROM diagrams, where R denotes the relation between two objects A and B. Fig. 4 illustrates the different representations of R(A, B). From the left to the right in Fig. 4, they are constraint, connection and predicate respectively.

4. The proposed method – CLT-ROM

The proposed method, CLT-ROM, is for cross-language collaborations in this paper. Our focus is on the transformation from English sentences to Chinese ROMs. The inverse transformation can be implemented by using this method under different rules. Fig. 5

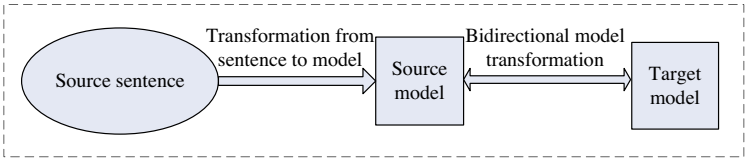


Fig. 3. Model based transformation.

Please cite this article in press as: K. Wen et al., A model based transformation paradigm for cross-language collaborations, Adv. Eng. Informat. (2012), http://dx.doi.org/10.1016/j.aei.2012.10.007

Table 1
Differences between traditional method and our transformation method.

	Traditional language translation method	CLT-ROM transformation method
Method type	Statistical method	Semantic method
Input	An English sentence	An English sentence
Output	A Chinese sentence	A Chinese model
Semantic preservation	Part of the semantics	Most of the semantics
Ambiguity	Unavoidable	Partly avoidable

illustrates the main process of the CLT-ROM method. Firstly, a source sentence is transformed to a Source ROM (S-ROM) by using mapping rules between the representations of grammar parser and the ROM relations. The mapping rules are discussed later. Then, a Target ROM (T-ROM) is generated by the CLT-ROM algorithm addressed in Section 4.2.

To facilitate the discussion, some notations and concepts are defined in Section 4.1, and an example sentence is given to show the process of transformation: "*Design a tool for fixing brake linings onto brake shoes for internal drum brakes.*"

4.1. Concepts and notations

Syntax is the study of the principles and processes by which sentences are constructed. Semantics is the study of meaning. It focuses on the relations between words or phrases. The lexical of a language is its vocabulary. Therefore, the definition of semantics can explain better through the concepts of restricted semantics and non-restricted semantics. Restricted semantics indicate that there are some relations between objects. On the contrary, non-restricted semantics indicate that there is no any relation between objects.

We assume that there are n words in a sentence $S = W_1 W_2 \cdots W_n$. If S is presented by a ROM diagram with n objects, $S_O = \{O_1, O_2, \ldots, O_i, \ldots, O_n\}$ is able to denote the sentence. The object O_i is corresponding to the word W_i. In addition, there are m ROM relations denoted as $S_R = \{R_1, R_2, \ldots, R_i, \ldots, R_m\}$. R_i can be expressed as $R_i(O_k, O_j)$, where O_k, $O_j \in S_O$ and $1 \leqslant k \leqslant n$, $1 \leqslant j \leqslant n$ and $k \neq j$. The three relations defined in the ROM are denoted as $R_i = \{Predicate|Constraint|Connection\}$. They imply different semantics for any $R_i(O_k, O_j)$:

(1) If $R_i(O_k, O_j) = Predicate$, it indicates that the object O_k restricts the semantic of the object O_j. Meanwhile, O_j restricts the semantic of O_k as well. The *Predicate* relation is a bidirectional restricted semantic relation.
(2) If $R_i(O_k, O_j) = Constraint$, it indicates that the object O_k restricts the semantic of the object O_j whereas O_j does not restrict the semantic of O_k. The *Constraint* relation is a non-bidirectional restricted semantic relation.

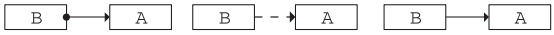


Fig. 4. Different representations of R(A, B).

(3) If $R_i(O_k, O_j) = Connection$, it indicates that the status of the objects O_k and O_j is independent. There is no any restricted semantic between the two objects with connection relation.

Therefore, the objects can be categorized into two types: *restricted semantic objects* and *non-restricted semantic objects*. For an object O_i, at least, has one object restricting its semantics, is a restricted semantic object. It can be formally represented as:

For any $O_i \in S_O$,
if $\exists R(O_k, O_i) = \{Predicate|Constraint\}$ or $\exists R(O_i, O_j) = \{Predicate\}$, $0 < j < n$, $0 < k < n$
then $O_i \in R\_object$

On the contrary, if an object O_i has no other object restricting its semantics, it is a non-restricted semantic object. It is formally represented as:

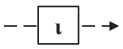
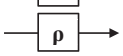
For any $O_i \in S_O$,
if $O_i \in S_O - R\_object$
then $O_i \in NR\_object$

Other than the these two objects, there is another special type of objects – *compound objects*. It is helpful to reduce the complexity of cross-language transformation. A definition is given to the compound objects as: A compound object is an object which includes two or more objects in it. Compound objects always consist of phrases, such as noun phrases, verb phrases, verbal phrases, and gerund phrases. The structure of a ROM diagram is more clear if compound objects are used. In the CLT-ROM method, determining compound objects is one of the important steps.

4.2. CLT-ROM algorithm

The algorithm for the CLT-ROM is shown in Algorithm 1. The input of the CLT-ROM algorithm is an English source sentence, whereas the output is a Chinese Target ROM (T-ROM) diagram. The Chinese T-ROM can be obtained in four steps. Step 1, the S-ROM is generated for the English source sentence S. Some preparations are necessary in step 1. We need to record some additional information for the objects of the S-ROM to classify these objects. In step 2, the compound objects are determined. Step 3, the objects of the S-ROM are transformed into the target objects. For the restricted semantic objects and non-restricted semantic objects, different transformational operations are performed. The structure of the S-ROM is updated and the Chinese T-ROM is obtained in step 4.

Table 2
Elements defined for the ROM [6].

Type		Graphic representation	Description
Object	Object		Everything in the universe is an object
	Compound object		It is an object that includes at least two objects in it
Relation	Constraint relation		It is a descriptive, limiting, or particularizing relation of one object to another
	Connection		It is to connect two objects that do not constrain each other
	Predicate relation		It describes an act of an object on another or that describes the states of an object

Please cite this article in press as: K. Wen et al., A model based transformation paradigm for cross-language collaborations, Adv. Eng. Informat. (2012), http://dx.doi.org/10.1016/j.aei.2012.10.007

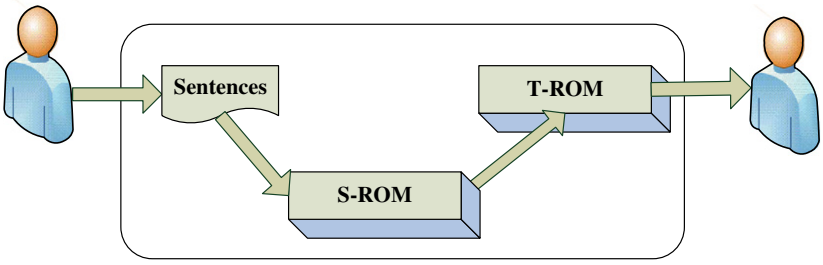


Fig. 5. Main process of the CLT-ROM method.

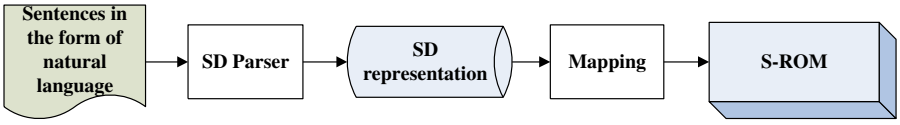


Fig. 6. How to generating S-ROM.

Table 3
A part of the mapping rules.

SD	amod	cc	conj	cop	csubj	det
ROM	Constraint	Connection	Connection	Predicate	Predicate	Constraint

Algorithm 1. CLT-ROM

1: Generating S-ROM and Preparation
2: Determining the compound objects
3: Object Transformation
4: Updating structure and obtaining T-ROM

4.2.1. Generation of S-ROM and preparation

Fig. 6 illustrates how to generate an S-ROM from a source sentence. Where *SD* (Stanford typed Dependencies) [31] represents sentences by using a grammar parser. It acts as a bridge between a natural language sentence and its ROM diagram. The S-ROM is generated by mapping an *SD representation* into a ROM representation. The mapping rules preserve the syntax logic relation between two objects. In order to improve the quality of mapping results, specified rules are developed to revise some special *SD representations*. Table 3 gives some mapping rules. The *amod* and the *det* are mapped into ROM constraint relations. The *cc* and the *conj* are mapped into ROM connection relations. The *cop* and the *csubj* are mapped into ROM predicate relations. The complete mapping rules are determined once a source language is selected.

The algorithm for S-ROM generation and preparation is shown in Algorithm 2. Once an S-ROM is generated (line 1), the ROM representation set S_{ROM} and the object set S_O (line 2) are obtained. By analyzing the objects in the S-ROM, some preparations are then completed, e.g., classifying parts of speech (POS), differentiating tense (present progressive, present perfect, or pluperfect), identifying voice (active or passive), and determining morphology (singular or plural). During the process of preparation, additional information (AI) is also obtained (lines 5 and 6). The information will be utilized in the later object transformation. Object lemmatization is also needed for the object set S_O (line 7), which is critical in getting the candidate transformation. Most lemmas in the dictionary are stored in terms of original morphology.

Algorithm 2. Generating S-ROM and preparation

1: Generating the S-ROM
2: Initial S_{ROM}, S_O
3: **for** *all* $O \in S_O$ **do**
4: $AOC[I_O] \leftarrow 0$
5: Record additional information (AI) for O
6: $HashMap\langle I_O, AI\rangle$
7: $O \leftarrow Lemmatize(O)$
8: **end for**
9: **for** all $R(O_t, O_h) \in S_{ROM}$ **do**
10: **if** R = Predicate **then**
11: $AOC[I_{O_t}] \leftarrow AOC[I_{O_t}] + 1$
12: $AOC[I_{O_h}] \leftarrow AOC[I_{O_h}] + 1$
13: **end if**
14: **if** R = Constraint **then**
15: $AOC[I_{O_n}] \leftarrow AOC[I_{O_n}] + 1$
16: **end if**
17: **end for**
18: **for** all $O \in S_O$ **do**
19: **if** $AOC[I_O] = 0$ *or* $AI(O).POS = preposition$ **then**
20: $S_n \leftarrow S_n \bigsqcup O$
21: **else**
22: $S_r \leftarrow S_r \bigsqcup O$
23: **end if**
24: **end for**

All of the elements in the set S_{ROM} are traversed. The restricted semantic relation and non-semantic relation are obtained from S_{ROM}. Concurrently, the number of restricted semantic relations is recorded for each object (lines 9–17). According to the definition in Section 3.1, classifying parts of speech (POS), as one kind of AI information, is useful to categorize the objects into *R_object* or *NR_object* (line 18–23). We used an *AOC* (the count of an object) array to store the pairs $\langle O, count\rangle$. The O is corresponding to the object whereas *count* is corresponding to the number of the object's restricted semantics. The elements of *AOC* are resorted in a descending order according to the value of *count*. To illustrated, the generated S-ROM for the example sentence is shown in Fig. 7.

4.2.2. Determination of compound objects

As aforementioned, the complexity of cross-language transformation can be reduced by using compound objects. Following principles are followed to determine compound objects.

Please cite this article in press as: K. Wen et al., A model based transformation paradigm for cross-language collaborations, Adv. Eng. Informat. (2012), http://dx.doi.org/10.1016/j.aei.2012.10.007

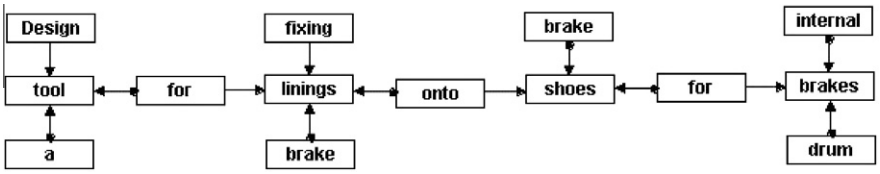


Fig. 7. The generated S-ROM for the example sentence.

(1) A compound object must contain as many objects as possible.
(2) A compound object has a definite semantic. Therefore, a compound object must be matched in the dictionary.
(3) Determining the compound objects should be conducted according to a certain order.

Furthermore, Algorithm 3 is proposed to serve the purpose. Firstly, all of the objects in S_O are stored in a queue Q_O (line 1). Secondly, an effort is made to find as many objects as possible to compose a compound object. The first object O_e of the queue Q_O is seen as the last object of a compound object (line 3). The variable x indicates the position of O_e (line 4). The object $StartP(O_e)$ which has the least ROM relation connections with O_e is regarded as the first object of the compound object, and y records the position of $StartP(O_e)$ (line 5). A recursion is used to identify the first object of a compound object. Then the current objects generated by composing all the objects between O_x and O_y is supposed to be a compound object. For every compound object $O_c = \{O_1, O_2, \ldots, O_x\}$, $O_i (i = 1,2,\ldots,x)$ is a simple object and its stem is stem (O_i). We traverse all possibilities that O_c can reach by O_i itself or its stem (O_i). Once one of the case matches the dictionary phrase we record it. If the current objects cannot be matched in the dictionary, they will be discarded. At the same time, the transformation of a determined compound object is obtained from the dictionary. The current objects are removed from the queue thereafter (line 7). This process is repeated until all the compound objects are determined (line 8). Finally, after all the compound objects are determined, the set of ROM representations S_{ROM} and the set of objects S_O (line 10) are updated. Fig. 8 shows the S-ROM with determined compound objects for the example sentence.

Algorithm 3. Determining compound objects

1: $Q_O \leftarrow S_O$
2: **while** $Q_O \neq empty$ **do**
3: $O_e \leftarrow getHead(Q_O)$
4: $x \leftarrow I_{O_e}$
5: $y \leftarrow StartP(O_e)$
6: $z \leftarrow Determine(O_x, O_y)$
7: $Q_O \leftarrow Q_O - \{O_z \cdots O_x\}$
8: $S_{CO} \leftarrow S_{CO} \cup \{O_z \cdots O_x\}$
9: **end while**
10: Update S_{ROM}, S_O, AOC

4.2.3. Object transformation

The third step of the CLT-ROM is object transformation, shown in Algorithm 4. Different transformational operations are carried out for the restricted semantic objects and non-restricted semantic objects. Usually, an object with more restricted semantics is the focus in understanding a sentence. Therefore, the transformation is conducted in a descending order according to the number of the object's restricted semantics (line 3). In the algorithm, $AOC[I_O]$ stores the number of an object's restricted semantics. The object with the largest restricted semantic number is transformed prior to the others.

Algorithm 4. Object transformation

1: $S = O_1 O_2 O_3 \cdots O_n$
2: **while** $S \neq empty$ **do**
3: **if** $O_i \in S \cap AOC[O_i]$ is the maximum **then**
4: $T \leftarrow O_i$
5: **end if**
6: **if** $O_i \in StopList$ **then**
7: record the translation of O_i
8: continue
9: **end if**
10: **for** each $O_x \in S$ **do**
11: **if** O_x restrict O_i and $O_x \notin StopList$ **then**
12: $T \leftarrow T \cup O_x$
13: **end if**
14: **end for**
15: $O_i' \leftarrow Trans(T)$
16: $S \leftarrow S - O_i$
17: $T \leftarrow NULL$
18: **end while**

For a restricted semantic object, the transformation is directly obtained by querying the dictionary (line 11). Multiple candidate transformations may exist. In this case, other related objects make contributions to reduce the ambiguity. Several methods were proposed to reduce or eliminate the ambiguity. The first method is calculating the co-occurrence probability for the related objects to confirm the transformation. The second method is computing the semantic similarity of the related objects to eliminate the ambiguity. The semantics of a word in HowNet [32] can be used to compute the semantic similarity. Meanwhile, the AI information, such as POS, is also useful to optimize the transformation by reducing the variation of candidate transformations.

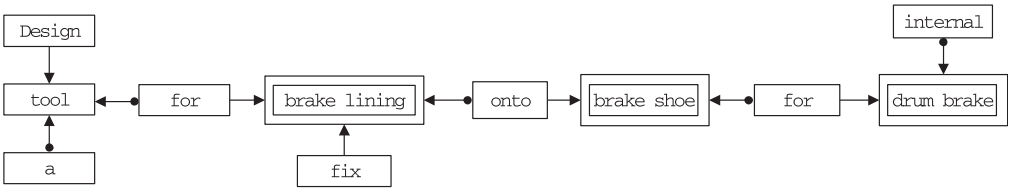


Fig. 8. S-ROM with determined compound objects for the example sentence.

Please cite this article in press as: K. Wen et al., A model based transformation paradigm for cross-language collaborations, Adv. Eng. Informat. (2012), http://dx.doi.org/10.1016/j.aei.2012.10.007

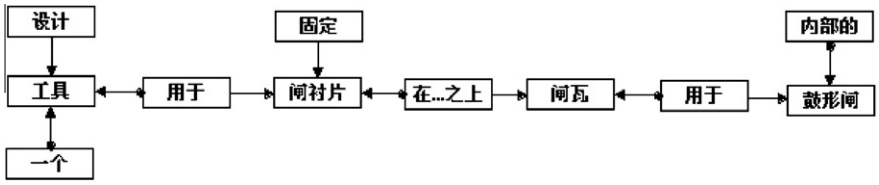


Fig. 9. T-ROM for the example sentence.

For the non-restricted semantic objects, a *Stop_List* is created to collect them. They are used to describe the statement between the restricted semantic objects. The corresponding candidate transformations of the *Stop_List* are also stored in the dictionary. Meanwhile, rules for the non-restricted semantic object transformation are developed. For example, some non-restricted semantic preposition objects, such as *for*, *on* and *to*, are transformed by combining with the related verbs. Some non-restricted semantic object transformations are combined with the voice and tense. Some non-restricted semantic objects have no meanings. After the object transformation is completed, the structure of ROM is updated, and the T-ROM is output. Fig. 9 shows the T-ROM for the example sentence.

5. Case study and evaluation

5.1. Case study

A prototype system, ROM-based Communication Tool (RCT), was developed to support online cross-language communication based on the CLT-ROM method. It is programmed and run in Windows XP laptop (1.83 GHZ Inter Core 2 Due CPU and 2 GB RAM) with a Java environment (MyEclipse 6.5 and JDK version 1.6.0_32). The RCT includes two clients: an English end and a Chinese one, for demonstration purpose. Each of them can take a text sentence as an input and output a ROM diagram, i.e. an English sentence for the English client and a Chinese sentence for the Chinese one. When a user submits an English sentence at the English interface, the system outputs the corresponding English ROM diagram on the same client, and the transformed Chinese ROM diagram on the Chinese client. It shares the same principle for a Chinese-to-English process. In this way, users from both clients can communicate with each other, sharing their ideas, and discussing the design problems. Just like popular instant messaging tools, people can use RCT to share their design requirements in real time without language barriers.

Assuming that the RCT has been adopted by aforementioned companies for online collaboration in product development, a conversion is initialized by a Chinese employee. A simplified process is demonstrated in Figs. 10 and 11: (1) the Chinese employee submits a sentence to ask an English-speaking partner to give a brief description about the use of the system, as shown in Fig. 10a. The RCT outputs the Chinese ROM diagram on the Chinese client and the English ROM diagram on the English client, as shown in Fig. 10b. (2) The English user reads the English ROM and replies the Chinese user as per request in English, as shown in Fig. 10c. The RCT transforms the English sentence to an English ROM diagram on the English interface, and outputs the corresponding Chinese ROM diagram on the Chinese client (see Fig. 10d). (3) The conversion goes onto clarify requirements, as illustrated in Fig. 11.

5.2. Evaluation

Since the transformation results of the RCT are graphs not sentences, the evaluation criterion of language translation cannot be directly applied. We define a criterion for transformation based on ROM evaluation, named *TRE* (Transformation based on ROM evaluation), which is a variation of *BLEU* (Bilingual evaluation understudy) [33].

5.2.1. BLEU (bilingual evaluation understudy)

Before giving an introduction to the *TRE*, some concepts about *BLEU* is addressed here. The *BLEU* is an evaluation criterion of language translation. The evaluation requires two ingredients, a numerical metric *BLEU* and a corpus of high quality reference translations. The *BLEU* can be computed by:

$$BLEU = BP \cdot \exp\left(\sum_{1}^{N} w_n \log P_n\right) \tag{1}$$

where P_n shown in Eq. (2) is a modified *n-gram* precision to capture two aspects of translation: adequacy and fluency. A brevity penalty factor *BP* is introduced to penalize the short candidate sentence. A translation using the same words as in the reference translations tends to satisfy the adequacy.

$$P_n = \frac{\sum_{C\in\{candidates\}}\sum_{n-gram\in C} Count_{clip}(n-gram)}{\sum_{C\in\{candidates\}}\sum_{n-gram'\in C} Count(n-gram')} \tag{2}$$

Considering the recall of language translation, the length of a candidate sentence should match the effective length of a reference sentence. The *BP* can be calculated by:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leqslant r \end{cases} \tag{3}$$

where c is the length of a candidate sentence. r is the length of a reference sentence.

5.2.2. TRE (transformation based on ROM evaluation)

Accordingly, the evaluation of cross-language transformation also requires two ingredients, a numerical metric *TRE* and a corpus of high quality reference transformations. *TRE* can be computed by the Eq. (4) which is similar to *BLEU*.

$$TRE = BP \cdot P \tag{4}$$

For the criterion *TRE*, since we transform a source sentence to a target ROM graph, P_1 in *BLEU* can be directly used in the *TRE*. A ROM based precision P_r is defined, and can be computed by:

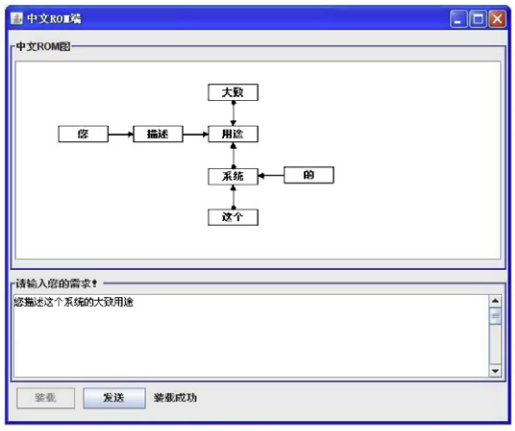
$$Pr = \left(\frac{Count(RR \cap TR)}{Count(TR)}\right) \tag{5}$$

where *RR* is a reference ROM representation. The candidate ROM representation is denoted by *TR*. First, we add up the clipped ROM representation counts. Then, the result is divided by the total number of *TR*. Hence, the ratio *P* of the criterion *TRE* can be computed by:

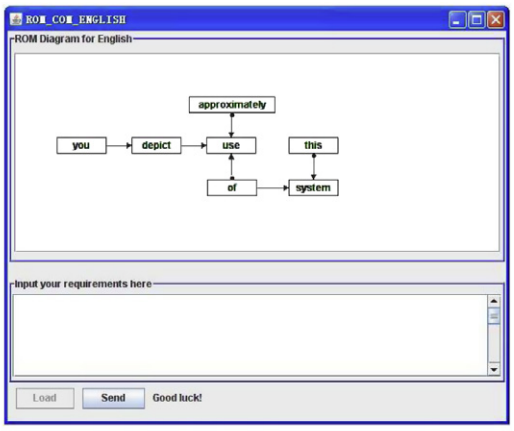
$$P = \exp(w_1 \log P_1 + w_r \log P_r) \tag{6}$$

The score of *P* indicates the capability of semantic preserving. w_1 and w_r are corresponding to the weight of P_1 and P_r. Here, $w_1 = w_r$. Similarly, considering the recall, we also introduce the brevity penalty *BP* to penalize the candidate ROM representation when the number of a candidate ROM representation is smaller

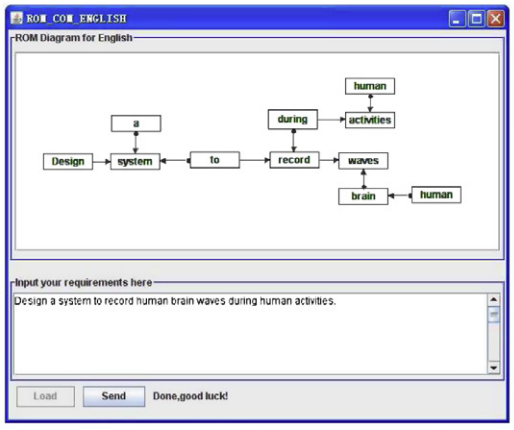
Please cite this article in press as: K. Wen et al., A model based transformation paradigm for cross-language collaborations, Adv. Eng. Informat. (2012), http://dx.doi.org/10.1016/j.aei.2012.10.007



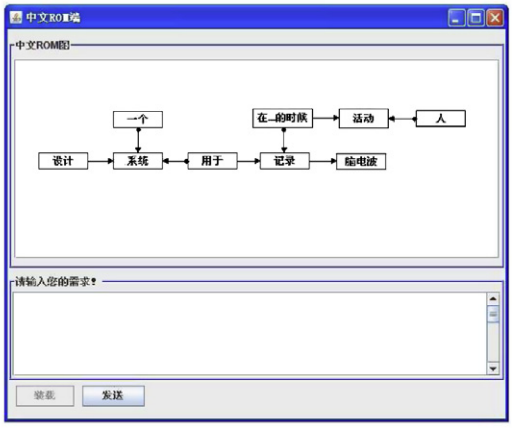
(a) Chinese user submits a sentence to request a system description



(b) English ROM results are shown simultaneously for the English user



(c) English user replies by submitting an English sentence



(d) Chinese ROM results are shown simultaneously for the Chinese User

Fig. 10. Chinese user asking for a system description.

than the number of a reference ROM representation. *BP* can be computed by the same method that is defined in the *BLEU*. Therefore, we get the evaluation criterion *TRE* shown in Eq. (7) by using the log form to facilitate the following comparison.

$$\log TRE = \min(1 - r/c, 0) + \log P \tag{7}$$

where r is the effective number of the reference ROM representation, c is the number of the candidate ROM representation.

5.2.3. Results of evaluations

5.2.3.1. Comparing with Google Translate in terms of sematic preserving. In order to indicate the capability of semantic preserving, we compared the results of CLT-ROM method with the results of Google Translate. It does not apply grammatical rules, since its algorithms are based on statistical analysis rather than rule-based analysis. The transformed ROM representation of a Google translated sentence is denoted by *GR*. Eq. (8) was used to calculate the precision of the *GR*, denoted by P'_r.

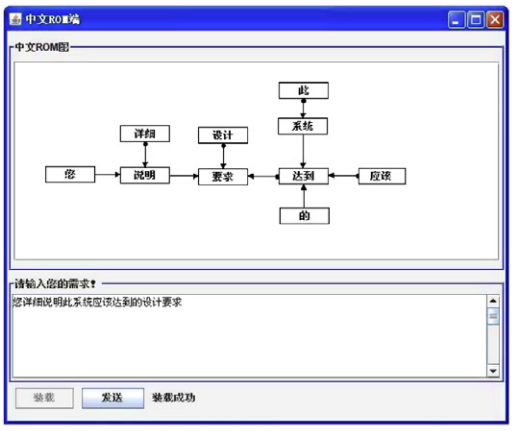
$$P'_r = \frac{Count(RR \cap GR)}{Count(GR)} \tag{8}$$

By using the criterion *TRE*, the respective score of semantic preserving were derived for the CLT-ROM method and Google Translate. The following eight sample sentences adopted from the book by Hubka et al. [34] are used as the test corpus. For each sentence, the scores P_1 and P_r, corresponding to both the CLT-ROM method and the Google Translate, are showed in Figs. 12 and 13, respectively.

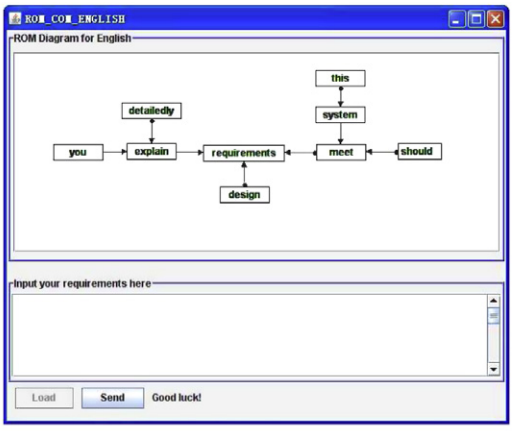
(1) Design a tool for fixing brake linings onto brake shoes for internal drum brakes.
(2) The user of this tool is a car mechanic.
(3) The working height of the user should follow ergonomic standards.
(4) The use of this tool should conform to the related industry safety standards.
(5) The service life of this tool should be around 5 years.
(6) The tool should be easy for transportation and maintenance.
(7) It will be manufactured in a specific workshop, which has specified equipments.
(8) The cost of this tool cannot be over $190.00.

For the score of P_1, the Google Translate can make the right translations for part of the words in the sentences. The results from the CLT-ROM method are slightly better than from the Google Translate. However, for the score of P_r, which implies more semantic information, a significant difference exists between the two methods. The CLT-ROM method has a higher P_r score than the Google Translate does. The results indicate that the Google Translate lost some semantics during the process of translation, while the CLT-ROM had a better capability of semantic preserving. The

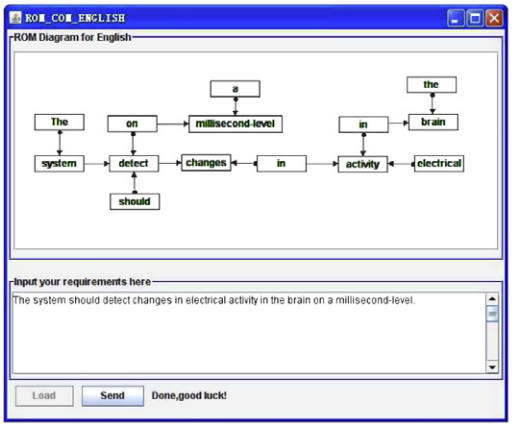
Please cite this article in press as: K. Wen et al., A model based transformation paradigm for cross-language collaborations, Adv. Eng. Informat. (2012), http://dx.doi.org/10.1016/j.aei.2012.10.007



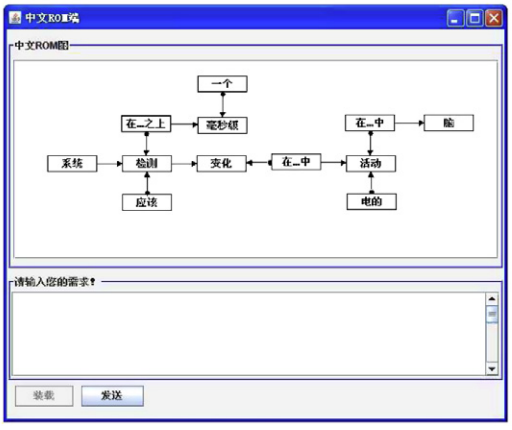
(a) RCT other requirements-Chinese user



(b) RCT other requirements-English user



(c) RCT reply for other requirement-English user



(d) RCT reply for other requirement-Chinese user

Fig. 11. Communication about other requirements.

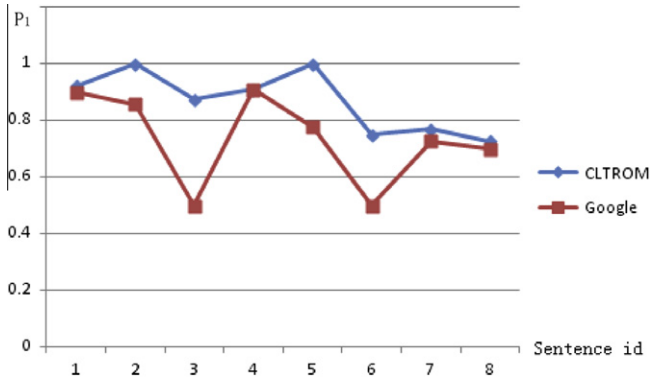


Fig. 12. P_1 generated by the CLT-ROM and the Google Translate for each sentence.

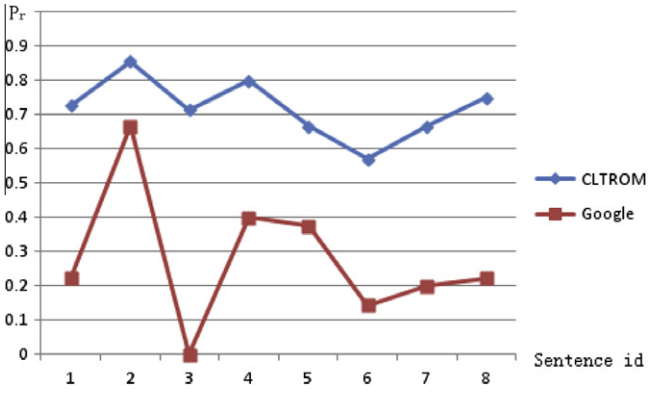


Fig. 13. P_r generated by the CLT-ROM and the Google Translate for each sentence.

results of comparative evaluations for total test corpus are shown in Table 4. According to the table, the CLT-ROM method has a higher score of P_r and *TRE* than that of the Google Translate. It also indicates that the proposed method has a better performance than the Google Translate in the capability of semantic preserving for cross-language transformation.

5.2.3.2. More on correctness tests. To evaluate the correctness of the transformation from English sentences to Chinese ROM diagrams, 50 sentences from all five basic sentence patterns in English are involved. The sentences chosen here, as listed as follows, are focusing on industrial requirements sharing and design problems discussing.

Pattern 1. Subject + intransitive verb
The use of this tool should conform to the related industry safety standards.
Pattern 2. Subject + link verb + subject Complement Ultrasonic motors are of great interest due to the flexibility of

Table 4
Overall results for the test corpus.

Transformation	P_1	P_r	*TRE*
Google	0.737	0.265	0.195
CLT-ROM	0.864	0.718	0.260

Table 5
Detailed results for the dataset.

Sentence patterns	Object correctness	Relation correctness
Subject + intransitive verb	0.878	0.874
Subject + link verb + subject complement	0.887	0.862
Subject + transitive verb + object	0.880	0.887
Subject + transitive verb + indirect object + direct object	0.840	0.795
Subject + transitive verb + object + object complement	0.839	0.800
Total sentences	0.877	0.860

miniaturization in comparison with conventional electromagnetic motors whose efficiency decreases significantly.

Pattern 3. Subject + transitive verb + object
The EEG can detect changes in electrical activity in the brain on a millisecond level.
Pattern 4. Subject + transitive verb + indirect object + direct object
The approach in designing these piezoelectric motors will bring the system more reliability.
Pattern 5. Subject + transitive verb + object + object complement
I think it impossible for us to finish the design task in such a short time.

We used two measures to evaluate the transformed Chinese ROM diagrams generated by the CLT-ROM method: the correctness of the objects and the correctness of the relations. For each sentence in the dataset, we draw a correct Chinese ROM diagram for the comparing purpose. After compared all the experimental results with the correct ones for the 50 sentences, we have achieved 81.7% for the object correctness and 82.7% for the relation correctness. The details results are shown in Table 5, indicating very favorable outcomes from our initial prototype system.

6. Conclusions and future work

In this paper, the CLT-ROM method is proposed to transform English sentences into Chinese ROM diagrams, for international collaboration on product development. The CLT-ROM method is able to capture the semantics and preserve information included in natural languages. A prototype, RCT, is developed for online collaboration. The case study and examples show that the CLT-ROM method along with the RCT obtains an advantage in the capability of semantic preserving, and it is feasible, intuitive, and effective for cross-language collaborations. It provides a well-initial basis for the further applications.

However, there are some limitations of our method compared with the Google Translate. People who use it much have a clear understanding of the ROM elements and their rules, though there are few of them. The CLT-ROM method is better to be used in some specific cross-language collaborations, for example, cross-language product collaborative design. The implemented prototype is still at its preliminary stage, it can transform one sentence to a target ROM diagram at a time. For a group of sentences that required to a merged ROM diagram, it cannot help so far. It cannot be used in any application that requires sentence to sentence translation. Nevertheless, we believe we are on the right track for an innovative cross-language translation.

The future work includes three aspects. Firstly, the current experiments did not use a large-scale data set. The data set is to be improved to cover all types of grammatical structures. A more in-deep test is going to be conducted thereafter. Secondly, a ROM-net-based dictionary is to be built. With the dictionary, the transformation results will be improved. Thirdly, since user feedback is valuable for the quality of language transformation, we plan to use the feedback recursively to make the prototype adaptable. Translating of target ROM diagrams to a target sentences with ROM merge functions will be implemented in the next version of the prototype system as well.

Acknowledgments

This paper is supported by National Natural Science Foundation of China under Grant Nos. 61173170, 60873225 and 70771043, CCF Opening Project of Chinese Information Processing, central university basic research special funding (Innovation Fund of Huazhong University of Science and Technology under Grant Nos. 2011TS135 and 2010MS068). The authors would like to extend their heartfelt gratitude to Prof. Yong Zeng for his knowledge and valuable suggestions. The authors are also grateful to anonymous reviewers for their constructive comments on this paper.

References

[1] U. Sekaran, Methodological and theoretical issues and advancements in cross-cultural research, Journal of International Business Studies 14 (1983) 61–73.
[2] X. Xiao, C. Zhao, S. Zhang, Do we talk differently: cross culture study on conference call, in: Proceedings of the 2nd International Conference on Usability and Internationalization, UI-HCII'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 637–645.
[3] N. Schadewitz, N. Zakaria, Cross-cultural collaboration wiki: evolving knowledge about international teamwork, in: Proceedings of the 2009 International Workshop on Intercultural Collaboration, IWIC '09, ACM, New York, NY, USA, 2009, pp. 301–304.
[4] M. Khalilov, J.A. Fonollosa, Syntax-based reordering for statistical machine translation, Computer Speech and Language 25 (2011) 761–788.
[5] A. McKay, A. de Pennington, J. Baxter, Requirements management: a representation scheme for product specifications, Computer-Aided Design 33 (2001) 511–520.
[6] Y. Zeng, Recursive object model (ROM) – modelling of linguistic information in engineering design, Computers in Industry 59 (2008) 612–625.
[7] M. Wang, Y. Zeng, Asking the right questions to elicit product requirements, International Journal of Computer Integrated Manufacturing 22 (2009) 283–298.
[8] D.Y. Zhang, Y. Zeng, L. Wang, H. Li, Y. Geng, Modeling and evaluating information leakage caused by inferences in supply chains, Computers in Industry 62 (2011) 351–363.
[9] X. Sun, Y. Zeng, F. Zhou, Environment-based design (EBD) approach to developing quality management systems: a case study, Journal of Integrated Design and Process Science 15 (2011) 53–70.
[10] X. Deng, G. Huet, S. Tan, C. Fortin, Product decomposition using design structure matrix for intellectual property protection in supply chain outsourcing, Computers in Industry 63 (2012) 632–641.
[11] Y. Zeng, L. Wang, X. Deng, X. Cao, N. Khundker, Secure collaboration in global design and supply chain environment: problem analysis and literature review, Computers in Industry 63 (2012) 545–556.
[12] H. Somers, Review article: example-based machine translation, Machine Translation 14 (1999) 113–157.
[13] E. Sumita, H. Iida, Experiments and prospects of example-based machine translation, in: Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, ACL '91, pp. 185–192.
[14] P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, R.L. Mercer, The mathematics of statistical machine translation: parameter estimation, Computational Linguistics 19 (1993) 263–311.
[15] D. Xiong, Q. Liu, S. Lin, Maximum entropy based phrase reordering model for statistical machine translation, in: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-44, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006, pp. 521–528.
[16] D. Talbot, H. Kazawa, H. Ichikawa, J. Katz-Brown, M. Seno, F.J. Och, A lightweight evaluation framework for machine translation reordering, in:

Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT '11, Association for Computational Linguistics, Stroudsburg, PA, USA, 1988, pp. 12–21.
[17] F.J. Och, H. Ney, The alignment template approach to statistical machine translation, Computational Linguistics 30 (2004) 417–449.
[18] R. Zens, H. Ney, T. Watanabe, E. Sumita, Reordering constraints for phrase-based statistical machine translation, in: Proceedings of the 20th International Conference on Computational Linguistics, COLING '04, Association for Computational Linguistics, Stroudsburg, PA, USA, 2004, pp. 205–211.
[19] S. Kumar, W. Byrne, Local phrase reordering models for statistical machine translation, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, Association for Computational Linguistics, Stroudsburg, PA, USA, 2005, pp. 161–168.
[20] C. Tillmann, A unigram orientation model for statistical machine translation, in: Proceedings of HLT-NAACL 2004: Short Papers, HLT-NAACL-Short '04, Association for Computational Linguistics, Stroudsburg, PA, USA, 2004, pp. 101–104.
[21] D. Chiang, A hierarchical phrase-based model for statistical machine translation, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, Association for Computational Linguistics, Stroudsburg, PA, USA, 2005, pp. 263–270.
[22] D. Xiong, Maximum entropy based phrase reordering model for statistical machine translation, in: Proc. of COLING-ACL, pp. 521–528.
[23] D. Zhang, M. Li, C.H. Li, M. Zhou, Phrase reordering model integrating syntactic knowledge for SMT, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 533–540.
[24] K. Yamada, K. Knight, A syntax-based statistical translation model, in: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, ACL '01, Association for Computational Linguistics, Stroudsburg, PA, USA, 2001, pp. 523–530.
[25] C. Quirk, A. Menezes, Dependency treelet translation: the convergence of statistical and example-based machine-translation, Machine Translation 20 (2006) 43–65.
[26] Y. Liu, Q. Liu, S. Lin, Tree-to-string alignment template for statistical machine translation, in: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-44, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006, pp. 609–616.
[27] S. Attardo, V. Raskin, Script theory revis(it)ed: joke similarity and joke representation model, International Journal of Humor Research 4 (2009).
[28] V. Raskin, C.F. Hempelmann, J.M. Taylor, The ontological semantic alternative: a meaning-based path to natural language understanding, in: Proceedings of the 2011 Annual Meeting of the Society for Design and Process Science, 2011.
[29] Y. Zeng, G. Cheng, On the logic of design, Design Studies 12 (1991) 137–141.
[30] Y. Zeng, S. Yao, Understanding design activities through computer simulation, Advanced Engineering Informatics 23 (2009) 294–308.
[31] P.-C. Chang, H. Tseng, D. Jurafsky, C.D. Manning, Discriminative reordering with chinese grammatical relations features, in: Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation, SSST '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 51–59.
[32] Z. Dong, Q. Dong, HowNet and the Computation of Meaning, World Scientific Publishing Co., New Jersey, USA, 2006.
[33] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 311–318.
[34] V. Hubka, M.M. Andreasen, W.E. Eder, Practical Studies in Systematic Design, first ed., Butterworth-Heinemann, London, UK, 1988.

Please cite this article in press as: K. Wen et al., A model based transformation paradigm for cross-language collaborations, Adv. Eng. Informat. (2012), http://dx.doi.org/10.1016/j.aei.2012.10.007