An English-Arabic Bi-Directional Machine Translation Tool in the Agriculture Domain

A Rule-based transfer approach for translating expert systems

K. Shaalan*, A. Hendam **, A. Rafea***

**The British University in Dubai, Informatics*
Dubai International Academic City,
Dubai, P.O. Box 345015,UAE
khaled.shaalan@buid.ac.ae
Honorary Fellow, School of Informatics, University of Edinburgh

***Central Lab. For Agricultural Expert Systems (CLAES), TEUES*
6 El Nour St., Giza, 1123 Egypt,
a_hendam@mail.claes.sci.eg

**** American University in Cairo, SSE,*
AUC Avenue, P.O. Box 74, New Cairo
1183 Egypt
rafea@aucegypt.edu

ABSTRACT: *The present work reports our attempt in developing an English-Arabic bi-directional Machine Translation (MT) tool in the agriculture domain. It aims to achieve automated translation of expert systems. In particular, we describe the translation of knowledge base, including, prompts, responses, explanation text, and advices. In the central laboratory for agricultural expert systems, this tool is found to be essential in developing bi-directional (English-Arabic) expert systems because both English and Arabic versions are needed for development, deployment, and usage purpose. The tool follows the rule-based transfer MT approach. A major design goal of this tool is that it can be used as a stand-alone tool and can be very well integrated with a general (English-Arabic) MT system for Arabic scientific text. The paper also discusses our experience with the developed MT system and reports on results of its application on real agricultural expert systems.*

KEYWORDS: machine translation, transfer-based translation, rule-based analysis, rule-based generation, Arabic natural language processing, bilingual agricultural expert systems.

1. Introduction

Arabic is the fourth most-widely spoken language in the world. It is a highly inflectional language, with a rich morphology, relatively free word order, and two types of sentences (Ryding, 2005): nominal and verbal. Arabic natural language processing has been the focus of research for a long time in order to achieve an automated understanding of Arabic (Al-Sughaiyer et al., 2004). With globalisation and expanding trade, demand for translation is set to grow. Computer technology has been applied in technical translation in order to improve speed and cost of translation (Trujillo, 1999). *Speed*: Translation by or with the aid of machines can be faster than manual translation. *Cost*: Computer aids to translation can reduce the cost per word of a translation. In addition, the use of machine translation (MT) can result in improvements in *quality*, particularly in the use of consistent terminology within a scientific text or for a specific domain.

With the recent technological advances in MT, Arabic has received attention in order to automate Arabic translations (Farghaly et al., 2009). In this paper, we follow a transfer-based MT approach. In the transfer approach (Trujillo, 1999), the translation process is decomposed into three steps: analysis, transfer, and generation. In the analysis step, the input sentence is analyzed syntactically (and in some cases semantically) to produce an abstract representation of the source sentence, usually an annotated parse tree. In the transfer step, this representation is transferred into a corresponding representation in the target language; a collection of tree-to-tree transformations is applied recursively to the analysis tree of the source language in order to construct a target-language analysis tree. In the generation step, the target-language output is produced. The (morphological and syntactic) generator is responsible for polishing and producing the surface structure of the target sentence. For each natural language processing component, i.e., analysis, transfer, and generation, we followed the rule-based approach. The advantage of the rule-based approach over the corpus-based approach is clear for (Abdel Monem et al., 2008; Shaalan, 2010): 1) less-resourced languages, for which large corpora, possibly parallel or bilingual, with representative structures and entities are neither available nor easily affordable, and 2) for morphologically rich languages, which even with the availability of corpora suffer from data sparseness.

English is a universal language that is widely used in the media, commerce, science and technology, and education. The size of the modern English content (e.g. literature and web content) is far larger than the amount of Arabic content available. Consequently, English-to-Arabic MT is particularly important. English-Arabic MT systems are mainly based on the transfer approach. For example, Ibrahim (1991) discussed the problem of the English-to-Arabic translation of embedded idioms and proverb expressions with the English sentences. Rafea et al. (1992) developed an English-to-Arabic MT system which translates sentences from the domain of political news from the Middle East. Pease et al. (1996) developed a system which

translates medical texts from English-to-Arabic. El-Desouki et al. (1996) discussed the necessity of modular programming for English-to-Arabic MT. Translation of an English subset of a knowledge base to the corresponding Arabic phrases is described in (El-Saka et al., 1999). Mokhtar et al. (2000) developed an English-to-Arabic MT system, which is applied on abstracts from the field of Artificial Intelligence. Shaalan et al. (2004) developed an MT system for translating English noun phrases into Arabic that was applied to titles of theses and journals from the computer science domain. On the contrary, little work has been done in developing Arabic-to-English MT systems. Al-barhamtoshy (1995) proposes a translation method for compound verbs. Shaalan (2000) described a tool for translating the Arabic interrogative sentence into English. Chalabi (2001) presented an Arabic-to-English MT engine that allows any Arabic user to search and navigate through the Internet using the Arabic language. Othman et al. (2003) developed an efficient chart parser that will be used for translating Arabic sentence.

The proposed rule-based transfer MT tool described here is part of an ongoing research to automate the translation of expert systems between Arabic and English. This process translates the knowledge base, in particular, prompts, responses, explanation text, and advices. In CLAES[1], this tool is found to be essential in developing bilingual (English-Arabic) expert systems because both English and Arabic versions are needed for development, deployment, and usage purpose.

The next section outlines the overall architecture of the proposed English-Arabic bi-directional MT tool with illustrative examples of simple and complex transfers. In following section, we present the results of evaluation experiments. In a concluding section, we present some final remarks. Appendix I presents a classification of problems in the evaluation experiments.

2. The System Architecture

The structure of the bi-directional MT tool is shown in Figure 1. In this figure the arrows indicate the flow of information. The oval blocks indicate the basic modules of the system. Rectangular blocks represent the linguistic knowledge. This architecture describes the translation of a knowledge base in the agricultural domain, in particular, see Table 1: 1) prompts: noun phrases in the form of interrogative expressions, 2) responses: legal values in the form of noun phrases, 3) advices: in the form of imperative expressions and noun phrases, and 3) explanation text: in the form of verbal and nominal sentences.

1. Stands for Central Laboratory of Agricultural Expert Systems (CLAES), Agricultural Research Centre (ARC), Egypt, http://www.claes.sci.eg .

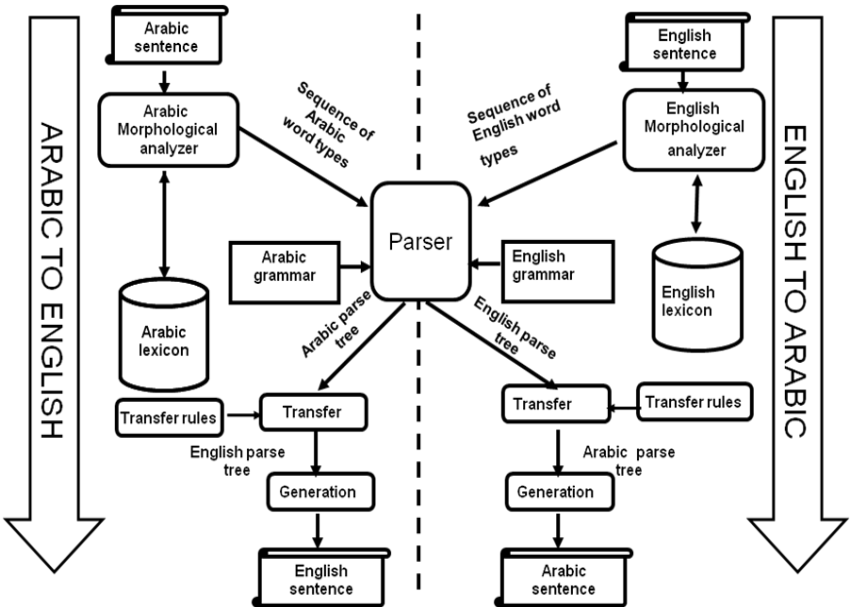

Figure 1. *Overall Structure of English-Arabic bi-directional sentence Translator*

	English	Arabic
Prompts	what is the abnormal leaves colour in the tunnel?	ما لون الأوراق الغير الطبيعى فى الصوبة؟
	what is the level of the nitrogen in the soil surface?	ما مستوى النتروجين فى سطح التربة؟
Responses (legal values)	bean mottle virus	فيروس تبقع الفول
	white growth with large black sclerotia	نمو أبيض مع أجسام حجرية سوداء
Advices (decisions)	Get rid of the remnants of the previous crop	تخلص من بقايا المحصول السابق
	spray when the number of nymphs is 3 on leaf	رش عندما يكون عدد الحوريات 3 على الورقة
Explanation	The unit for micro element for manganese during vegetative stage two	وحدة العناصر الصغرى من المنجنيز خلال مرحلة النمو الخضرى الثانية
	the added fertilization elements are determined during the flowering stage by using the watery fertilization elements index	تحدد عناصر التسميد المضافة خلال مرحلة التزهير بإستخدام ترتيب عناصر التسميد المائية

Table 1. *Examples of English-Arabic textual knowledge*

The proposed system is based on the transfer approach with three main components for each direction of translation: analysis, transfer, and generation. The analysis component consists of two steps morphological analysis and parsing. For accomplishing morphological analysis the lexicon is necessary, which is a repository of word stems. As Arabic is morphologically rich language, the morphological analysis of Arabic-to-English MT is an important step that is needed before we proceed with parsing the input sentence (Rafea et al., 1993). The transfer component has a collection of tree-to-tree transformations to the analysis tree of source sentence in order to construct a target analysis tree. The generation component generates the target language words according to the semantic features of the source language words. In our bi-directional English-Arabic translator, the actual translation occurs in the transfer phase. To explain how the sentence transfer process is performed by our translation system, we provide illustrative examples in Figure 2 through Figure 3 to show simple transfer of a noun phrase and compound transfer of a complete sentence, respectively. The former is an example showing that the syntactic transfer between English and Arabic noun phrase parse trees yields a representation in which word order is reversed. The later is a wider example showing the syntactic transfer between English sentence parse tree and Arabic verbal sentence parse tree yields a representation in which the Arabic VSO (verb-subject-object) order is transformed into the English SVO order.

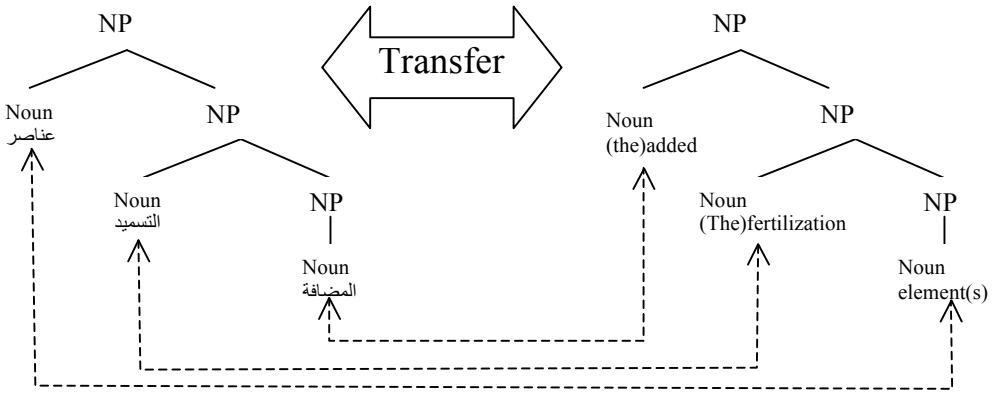

Figure 2. *Simple Transfer of Noun Phrase*

3. Automatic Evaluation

To meet the demands of a rapid MT evaluation method, various automatic MT evaluation methods have been proposed in recent years. These include the BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002; Akiba et al., 2004). BLEU has attracted many MT researchers, who have used it to demonstrate the quality of their novel approaches to developing MT systems. BLEU is an automatic scoring method based on the precisions of N-grams The precision of N-grams is calculated against reference translations produced by human translators. The results of BLEU

is a score in the range of [0,1], with 1 indicating a perfect match. In order to evaluate the quality of our MT system by the Bleu tool we conducted two experiments in each direction of translation, i.e., from English to Arabic, and vice versa.

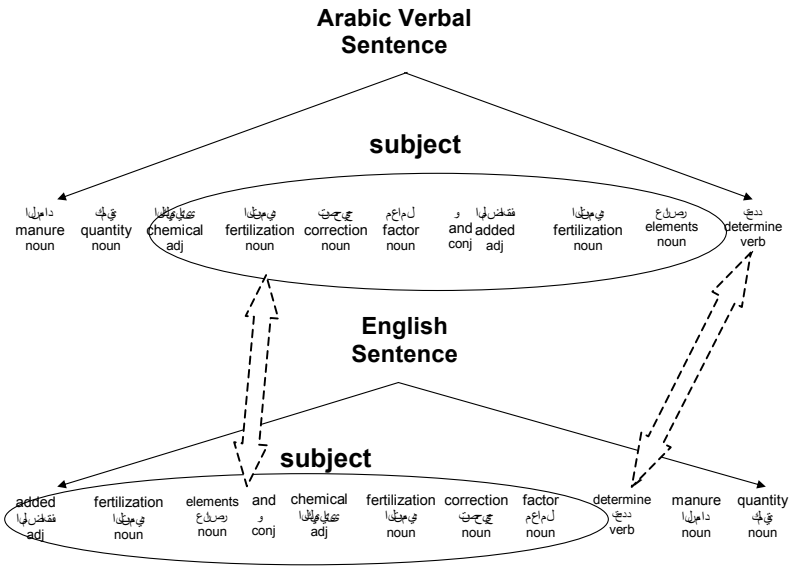

Figure 3. *Compound Transfer of verb and subject of a sentence*

A set of real parallel 100 phrases and sentences from both English and Arabic versions of agricultural expert systems at CLAES, was used as a gold standard reference test data. This set consists of 23 advises, 46 prompts, and 31 explanation and responses. The evaluation methodology is performed as follows: 1) Run the system on the test data, 2) Automatically evaluate the system output against the reference translation and get results of the BLEU score, 3) Classify the problems that arise from mismatches between the two translations, 4) For problems that needs an alternative reference translation such as synonyms, prepare a second reference translation for the identified problems, and 5) Rerun the system on the same test data using both reference translations and present the results of improvements.

3.1. *English to Arabic evaluation experiment*

The automatic evaluation results of experiment I are shown in Table 2. There are 9 classifications of problems that arise from the divergences and mismatches between system output and reference translation which is shown in Table 6. As for problems 1, 4, 5, and 6, we made the changes on a second reference translation but for the remaining problems they are not solved at the

moment as more research is needed to decide on their translations. Table 3 presents the automatic evaluation results of experiment IV which shows an improvement from 0.4504 to 0.6427.

	BLEU Score
Advices	**0.5147**
Prompts	**0.4433**
Explanation and responses	**0.4703**
Overall	**0.4504**

Table 2. *Results of automatic evaluation in Experiment I*

	BLEU Score
Advices	**0.7673**
Prompts	**0.6549**
Explanation and responses	**0.6156**
Overall	**0.6427**

Table 3. *Results of automatic evaluation in Experiment II*

3.2. *Arabic to English evaluation experiment*

The automatic evaluation results of experiment III are shown in Table 4. There are 4 classifications of problems that arise from the divergences and mismatches between system output and reference translation which is shown in Table 7. As for problems 1 and 4, we made the changes on a second reference translation but for problems 2 and 3 they are not solved at the moment as more research is needed to decide on their translations. Table 5 presents the automatic evaluation results of experiment IV which shows an improvement from 0.4581 to 0.8122.

	BLEU Score
Advices	**0.4019**
Prompts	**0.4988**
Explanation and responses	**0.5616**
Overall	**0.4581**

Table 4. *Results of automatic evaluation in Experiment III*

	BLEU Score
Advices	**0.8682**
Prompts	**0.7851**
Explanation and responses	**0.8169**
Overall	**0.8122**

Table 5. *Results of automatic evaluation in Experiment IV*

6. Conclusions

In this paper, we described the development of a novel English-Arabic bi-directional rule-based transfer MT tool in the agriculture domain. The translation between monolingual English and Arabic expert systems leads to rapid development and deployment of agricultural expert systems when one version is available. However, in the current version we may need to resort to minor post editing. Moreover, this tool would facilitate knowledge acquisition process to be either in English when international agricultural domain experts are available or in Arabic from local domain experts, which lead to bridging the gap of the language barrier.

A set of gold standard parallel English-Arabic phrases and sentences selected from agricultural expert systems developed at CLAES, is used to evaluate our approach, as well as the quality of the output of the MT tool. The problems found are classified, explained, and possible improvements, to some extent, are dealt with. The overall evaluation results, according to the presented evaluation methodology, were satisfactory. The automatic evaluation under one reference set achieved a BLEU score of 0.4504 for English-to-Arabic direction and 0.4581 for Arabic-to-English direction, whereas for two reference sets achieved 0.6427 for English-to-Arabic direction and 0.8122 for Arabic-to-English direction. However, more investigations are needed in order to make further improvements. On possible future direction is to use semantic processing. Another direction is to invest in building parallel corpora in the agriculture domain and employ the statistical machine translation approach.

7. References

Akiba Y., Federico M., Kando N., Nakaiwa H., Paul M., Tsujii J. "Overview of the IWSLT04 evaluation campaign", *In: Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, pp. 1—12, 2004.

Abdel Monem, A., Shaalan, K., Rafea, A., Baraka, H., *"Generating Arabic Text in Multilingual Speech-to-Speech Machine Translation Framework*", Machine Translation, Springer, Netherlands, 20(4): 205-258, December 2008.

Al-barhamtoshy A., "Arabic to English Translator of Compound Verbs", *In Proceeding of the Annual Conference on Statistics, Computer Science, and Operations Research*, Cairo University, December 1995.

Al-Sughaiyer I., Al-Kharashi I., "*Arabic morphological analysis techniques: A comprehensive survey*". Journal of the American Society for Information Science and Technology, 55(3):189—213, 2004.

Chalabi A. Sakhr "Web-based Arabic-English MT engine", *In proceeding of the ACL/EACL Arabic NLP Workshop*, 2001.

El-Desouki A., Abd Elgawwad A., and Saleh M., "A Proposed Algorithm For English-Arabic Machine Translation System", *In Proceeding of the 1st KFUPM Workshop on Information and Computer Sciences (WICS): Machine Translation*, Dhahran, Saudi Arabic, 1996.

El-Saka T., Rafea A., Rafea M., Madkour M., "English to Arabic Knowledge Base Translation Tool", *In Proceedings of the 7th International Conference on Artificial Intelligence Applications*, Cairo, Feb., 1999.

Farghaly A., Shaalan K. *"Arabic Natural Language Processing: Challenges and Solutions*", ACM Transactions on Asian Language Information Processing (TALIP), the Association for Computing Machinery (ACM), 8 (4):1-22, December 2009.

Ibrahim M., A Fast and Expert Machine Translation System involving Arabic Language, Ph. D. Thesis, Cranfield Institute of Technology, UK, 1991.

Mokhtar H., Darwish N., Rafea A., "An automated system for English-Arabic translation of scientific texts (SEATS)". *MT2000: machine translation and multilingual applications in the new millennium: international conference*, University of Exeter, British Computer Society, 20-22 November 2000.

Othman E., Shaalan K., and Rafea A. "A Chart Parser for Analyzing Modern Standard Arabic Sentence", *In proceedings of the MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches*, New Orleans, Louisiana, USA, 2003.

Papineni K., Roukos S., Ward T., Zhu W-J. BLEU: a Method for Automatic Evaluation of Machine Translation. *In: 40th annual meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, pp.311—318, 2002.

Pease C. and Boushaba A., "Towards an Automatic Translation of Medical Terminology and Texts into Arabic", *In Proceedings of the Translation in the Arab World, King Fahd Advanced School of Translation*, Nov. 27-30, 1996.

Rafea, A., Sabry, M., El-Ansary, R., Samir, S., "Al-Mutargem: A Machine Translator for Middle East News", *In Proceedings of the 3rd International Conference and Exhibition on Multi-lingual Computing*, December 1992.

Rafea A., Shaalan K., "Lexical Analysis of Inflected Arabic words using Exhaustive Search of an Augmented Transition Network", *Software Practice and Experience*, John Wiley & sons Ltd., UK,23(6):567-588, June 1993.

Ryding, K. Reference Grammar of Modern Standard Arabic, Cambridge University Press, Cambridge, UK, 2005.

Shaalan K., "Machine Translation of Arabic Interrogative Sentence into English", *In Proceeding of the 8th International conference on Artificial Intelligence Applications*, pp. 473-483, Egyptian Computer Society (EGS), Egypt, 2000.

Shaalan, K. ,Rafea, A., Abdel Monem, A., Baraka, H., *"Machine translation of English noun phrases into Arabic*", The International Journal of Computer Processing of Oriental Languages (IJCPOL), 17(2):121—134, 2004.

Shaalan K., *"Rule-based Approach in Arabic Natural Language Processing*", Special Issue on Advances in Arabic Language Processing, the International Journal on Information and Communication Technologies (IJICT), Serial Publications, New Delhi, India, June 2010. Submitted for publication.

Trujillo A. Translation Engines: Techniques for Machine Translation, Springer Verlag, London, UK. 1999.

8. Appendix I: classification of problems in experiments I & III

1. Difference due to using a synonym of the target Arabic noun	the added fertilization elements are determined during the flowering stage by using the watery fertilization elements index	Source
	يحدد عناصر التسميد المضاف خلال مرحلة التزهير باستخدام **فهرس** عناصر التسميد المائي	Reference
	يحدد عناصر التسميد المضاف خلال مرحلة التزهير باستخدام **ترتيب** عناصر التسميد المائي	Output
2. Different translation of a preposition	The melted fertilization elements index **in** water for nitrogen during the second vegetative growth stage **in** kgm Fert/m^3	Source
	ترتيب عناصر التسميد المذابة فى الماء من النيتروجين خلال مرحلة النمو الخضرى الثانية **فى** كجم تسميدة/متر^3	Reference
	ترتيب عناصر التسميد المذابة فى الماء من النيتروجين خلال مرحلة النمو الخضرى الثانية **بـ**كجم تسميدة/متر3	Output
3. Misinterpret Arabic conjunction of words as English conjunction of phrases	The used fertilizers units total quantity determines the season length based on **current and previous quantity of manure**	Source
	يحدد أجمالى كمية وحدات الاسمده المستخدمه طول العروه بناء على **الكميه الحاليه و الكمية السابقه للسماد**	Reference
	يحدد اجمالى وحدات الاسمدة المستخدمة طول العروة بناء على **الحالية و كمية سابقة للسماد**	Output
4. An optional pronoun might come after the Arabic interrogative particle	What is the abnormal growth colour on the fruits?	Source
	ما **هو** لون النمو الغير طبيعى للثمار؟	Reference
	ما لون النمو الغير طبيعى للثمار؟	Output
5. Some words may have either sound	What is the shape of the irregular **fruits**?	Source
	ما شكل **الثمار** غير المنتظمة؟	Reference

plural feminine noun or broken (irregular) plural	ما شكل **الثمرات** غير المنتظمة؟	Output
6. Non-standardization of the Arabic Written letters	"soil", "second", etc.	Source
	"تربه" "الثاني"	Reference
	"تربة" "الثاني"	Output
7. Disagreement in present tense prefix of an Arabic verb	the added fertilization elements quantity and the chemical fertilizer correction factor determines the manure quantity and the unit during the flowering stage	Source
	تـحدد كمية عناصر التسميد المضافة و معامل تصحيح السماد الكيميائى كمية و وحدة السماد خلال مرحلة التزهير	Reference
	يـحدد كمية عناصر التسميد المضافة و معامل تصحيح السماد الكيميائى كمية و وحدة السماد خلال مرحلة التزهير	Output
8. Disagreement in gender between the adjective and the noun it modifies	The fertilization quantity from magnesium during the second vegetative growth stage in kg Fert/m^3	Source
	كمية التسميد من الماغنسيوم خلال مرحلة النمو الخضرى **الثانية** بكجم تسميدة/متر^3	Reference
	كمية التسميد من الماغنسيوم خلال مرحلة النمو الخضرى **الثاني** بكجم تسميدة/متر^3	Output
9. missing definite article in the Arabic noun	the drippers number and the dripper flow rate determine the irrigation motor time in **minutes**	Source
	يحدد عدد النقاطات و معدل تصرف النقاط وقت موتور الرى **بـالدقائق**	Reference
	يحدد عدد النقاطات و معدل تصرف النقاط وقت موتور الرى **بدقائق**	Output

Table 6. *Classification of problems in Experiment I*

1. Difference due to synonyms of a target English noun	**كمية السماد العضوى**	Source
	the organic **fertilizer** quantity	Reference
	the organic **manure** quantity	Output
2. Selecting ambiguous category of a source Arabic word	**رش** المساحة المصابة فقط	Source
	spray the infected area only	Reference
	the infected area was **sprayed** only	Output
3. Misinterpret English conjunction of words as Arabic conjunction of phrases	حساب كمية المياه الكلية فى كل مرحلة بإستخدام **تاريخ البداية و النهاية**	Source
	The total water quantity calculation for every stage by using the **start and the end date**	Reference
	the total water quantity calculation for every stage by using the **start date and the end**	Output
4. Variant translation without the preposition "of"	كمية السماد	Source
	The quantity **of** fertilizer	Reference
	The fertilizer quantity	Output

Table 7. *Classification of problems in Experiment III*