Arabic-to-Malay Machine Translation Using Transfer Approach

Mohammed M. Abu Shquier

1 School of Computer Science and Information Technology, Jerash University, Jerash, Jordan

Correspondence: Mohammed M. Abu Shquier, School of Computer Science and Information Technology, Jerash University, Jerash, Jordan.

Received: June 23, 2020 Accepted: August 18, 2020 Online Published: September 30, 2020
doi:10.5539/cis.v13n4p1 URL: https://doi.org/10.5539/cis.v13n4p1

Abstract
Translation from/to Arabic has been widely studied recently. This study focuses on the translation of Arabic as a source language (SL) to Malay as a target language (TL). The proposed prototype will be conducted to map the SL “meaning” with the most equivalent translation in the TL. In this paper, we will investigate Arabic-Malay Machine Translation features (i.e., syntactic, semantic, and morphology), our proposed method aims at building a robust lexical Machine Translation prototype namely (AMMT). The paper proposes an ongoing research for building a successful Arabic-Malay MT engine. Human judgment and bleu evaluation have been used for evaluation purposes, The result of the first experiment prove that our system (AMMT) has outperformed several well-regarded MT systems by an average of 98, while the second experiment shows an average score of 1-gram, 2-gram and 3-gram as 0.90, 0.87 and 0.88 respectively. This result could be considered as a contribution to the domain of natural language processing (NLP).

Keywords: machine translation, transfer-based approach, ANLP, AMMT

1. Introduction
Arabic is one of the natural languages that is spoken by hundreds of millions of people as a native language, besides, it is the language of prayers for around 1.4 billion Muslims around the world (Shaalan. et.al., 2019). Arabic is considered as a derivation-language, subject pronoun-drop, and Subject-Verb-Object (SVO) structural language by default. On the other hand, Malay is the mother language of many people in southeast Asia (Hamza et. al, 2019). Morphological Speaking, Malay words are formed as:

1. (affixation): affex(es) + root.
2. (composition): composition of a compound word.
3. (reduplication): words or part of words repetition.

At the level of syntax, the default Malay sentence structure is Subject-Verbal-Object (SVO). Besides, Malay is a Verbal grammatical language (i.e., possessive, adj.).

According to (Al Saket et. al., 2014), Malay Language has a robust lexical features, besides, it is a language with no inflections at all for its verbs or nouns, recall, its morphology is formed by using affixes, composition and reduplication. According to Almeshrky et al. (2012), researchers should take into their consideration three types of knowledge to obtain a proper translation for this pair of languages.

1. Comprehend the source language (lexicon, morphology, syntax, and semantics) to understand the meaning of the source text.
2. Comprehend the following features (lexicon, morphology, syntax, and semantics) in the target language to produce a better translation.
3. Understand “the subject matter”.

2. Related Work
Several researchers publish their articles in this domain, particularly for this pair of languages. Abdalla (2012) introduced a rule-based MT, he went through the morphological and syntactical analysis of the SL to obtain a syntactic structure, to be used for the final representation of the TL using a the transfer approach. Almeshrky et al. (2012) developed a machine translation from Arabic language to Malay for dialogue purposes. Ahmed Alsaket et. al., (2014) demonstrated
a rule-based Arabic to Malay MT system. They have used the BLEU for evaluating their hypothesis, beside the natural judgment to evaluate the correctness of their system. Abodina et al. (2015) developed an Arabic to Malay MT system. Unlike Abdalla (2012), Almeshrky et al. (2012) and Ahmed Alsaket et. al., (2014), Abodina et al. study focuses on medical domain that contains fifty dialogue sentences (50) (dialogue between doctor and patient).

According to Tatabahasa Dewan (2008), Malay Language covers four structures:

Table 1. Malay Basic Structures

| Category | Phrases                          |
|----------|----------------------------------|
| 1        | Noun Phrase + Noun Phrase        |
| 2        | Noun Phrase + Verb Phrase        |
| 3        | Noun Phrase + Adjective Phrase   |
| 4        | Noun Phrase + Prepositional Phrase |

**Ambiguity in Arabic-Malay Translation System.** These are several challenges that need to be taken into consideration in automation of Malay language.

- Several meaning for the same Arabic Word, let us take these two examples:

1. \(\text{كَتَابُ البِكْرَةُ} \rightarrow \text{kindhearted}\) may be translated as ("Baik budi" or "baik hati").

2. \(\text{كَتَابُ البِكْرَةُ} \rightarrow \text{beloved}\) may be translated as ("bintang hati", "buah hati", "mahkota hati", "rangkai hati", "tajuk mahkota", "tangkai hati", "tangkai kalbu" or "bintang terang").

- Several meaning for the same Malay Word, for example:

1. "mata air" \(\rightarrow\) (lover, spring).
2. "orang putih" \(\rightarrow\) (pious man, European people).
3. "air muka" \(\rightarrow\) (face, pride)
4. "bawa diri" \(\rightarrow\) (to be independent).
5. "Bagai cicak makan kapur" \(\rightarrow\) (pleased).
6. "Ada air adalah ikan" \(\rightarrow\) (people in a country, fortune is everywhere)

3. **System Design and Architecture**

According to (Shaalan, 2010), the transfer-based translation passes through three phases:

1. Analysis process.
2. Transferring process.
3. Generation process

Initially, the input is analysed to have a certain SL structure that maps the "meaning" to generate a proper equivalent translation in the TL.

3.1 **Analysis Module**

we have analyzed the prototype lexically, morphologically and syntactically:

3.1.1 Lexical Databases
The information or features assigned to every individual words are usually defined as lexical resources, however, in our approach, we have developed a lexicon for Arabic-Malay words/phrase and we then assigned each words/phrase meaning with its features (i.e., number, gender, person, case, humanity, and alive/non-alive).

3.1.2 Tokenisation

The work “token” means splitting text into smaller units. The tokenization in our system extracts clitics, the prefixes and the suffixes of each word in the input sentence (Attia, 2007). The process is shown in figure 1, a list of Arabic_words_list will be returned as shown in figure 1 below.

3.1.3 Morphological Analysis

In this process, each word will be analyzed morphologically according to derivational rules (Badaro et al., 2019)(Habash, 2008). The derivation algorithm invokes certain features (i.e., verb-adj, sub-noun, etc) of the input considering (number, gender, humanity, alive, etc...) (Shquier MMA, 2019, 2013).

3.1.4 Syntactic Analysis

Many researchers consider this process as a major component of any MT system, this particular process analyses the SL to determine a reasonable grammatical structure, then this information will be used to split the sentence into smaller unit. However, once the normalizer/tokenizer finished their task, the parser takes the input and return a list of their part of speech as shown in figure 2. Stanford parser has been used for this purpose [?].

3.2 Transformation Module

The transformation is carried out using two processes:

1. Lexical Transfer.
2. Structural Transfer.

The transformation is carried out as follows:

1. Calling bilingual dictionary Arabic-Malay.
2. Calling parser to get POS.
The prototype framework is shown as a flow chart in Figure 2.

![Figure 2: Transformation Flowchart](http://cis.ccsenet.org)

3.3 Generation Module

In this process, the output of the TL will be rendered according to certain form concerning language grammar and meaning.

1. Accepts the Malay word to generate a well-format sentence.

2. Considering agreement and reordering as shown in figure 3, certain rules are considered during this process:

   - Malay ignores the definite article in general.
   - Malay dual nouns are translated by adding the word “dua” before the noun.
   - Malay nouns are indirectly inflected for gender.
   - Malay affixes attached to adjectives are mostly similar to those attached to verbs.
   - Malay pronouns depend on the speakers’ status.
   - Malay possessive pronouns are not attached to noun.
   - Malay classifiers (Penjodoh Bilangan) precedes nouns to show their amounts as follows:

     - orang (person, people) → orang 在 negara
     - ekor (tail) → ekor 在 negara
     - buah (fruit) → buah 埔殖在 negara
     - biji (seed) → biji 埔殖 in negara
     - batang (stick) → batang 埔殖 in negara

   - orang (person, people) is used for humans.
   - ekor (tail) is used for animals.
   - buah (fruit) is used for most inanimate objects. eg. books, tables, cars, houses, schools.
   - biji (seed) is used for small, round objects such as eggs, sweets and fruits.
   - batang (stick) is used for long, slim items such as pencils, pens, or sticks.
keping (pieces → گوشت گوشت) is used for a piece/pieces of paper, bread, cake, cheques, photographs.

pucuk (shoots → گوشت) is used for letters and arms.

Most verbs are preceded by a verbal prefix(es), (i.e., meng- for active voice, di- for passive voice and ber- for intransitiveness).

In Malay noun phrases, modifiers generally follow the head but quantifiers usually precede it.

No inflections in Malay, instead, prepositions are used to indicate syntactical relations.

Malay tense is normally denoted by time adverbs (i.e., ”sudah” → ”already”)

Malay imperfect verb/adverb denoted by tense indicators (i.e., ”sedang, telah” → ”still, already”).

Malay has no concatenated pronouns, instead they are separately written based on number, gender and tense features.

let us take an example on how the system handles the SL-TL word ordering based on the rules mentioned earlier, for the (SL) گوشت گوشت گوشت گوشت گوشت گوشت, the associated Arabic sentence matches the rule VD/1;NS/2;;N/3;J/4; then, the corresponding Malay mapping structure would be NNS/1;VBD/2;NNS/3;JJ/4, hence, the equivalent TL translation would be pelajar/NNX menyelesaikan/VBX ۱ masalah/NNX yang sukajJ. The flow of the agreement and ordering process is shown in figure 3. it is worthy stressing that we have built 183 structures to map the SL sequence structure with its corresponding Malay structure, a sample of this table is shown in Table

4. Implementation and Design

We have exhibited the entire process of our prototype in Figure 4, the developed design utilizes a framework developed by Hamdy N. Agiza (2012).

1. Analysis Module (Arabic text)

(a) (SL) input
(b) Tokenization
(c) Parsing 1.)

---

1VBX: (VB, VBZ, VBD, VBN) like (verb, base (believe), verb, -s (believes), verb, verb, past tense (believed), past participle (believed))
Table 2. A Sample of Arabic-Malay Mapping Patterns

| Features          | Rules    | Description                                      |
|-------------------|----------|--------------------------------------------------|
| Arabic Pattern    | VBX/1;NNX/2 | Pattern structure with word order               |
| Subject           | 2        | This means that the subject is the 2nd word      |
| Main verb         | 1        | This means that the main verb is the 1st word    |
| Object            | NULL     | This means: there is no object in this pattern   |
| Verb Agreement    | 1/2      | Agreement to be handled between both words       |
| Adj. Agreement    | NULL     | There is no adjective in this pattern            |
| Complement Feature| No       | This particular pattern has no complement        |
| Malay Pattern     | NNX/1;VBX/2 | Represents the equivalent pattern in Malay      |

Figure 4: System Architecture flowchart
2. **Transfer 2)**

(a) Arabic-Malay Lexicon
   i. Adding POS, to Arabic words in Lexicon.
   ii. Arabic-Malay dictionary Arabic_POS_list[i].
   iii. Accepts Arabic_words_list[i] and Arabic_POS_list[i].
   iv. get Malay_words_list[i].

3. **TL Generation**

(a) Synthesis TL (Malay) - rule-basis
   i. Accepts Malay_words_list[i].
   ii. Link Malay_words_list[i] based on the Malay_structure_list[i].
   iii. TL Generation.

(b) Malay Morphology
   i. Invoke Rules reordering_Malay_words_list [i]to get the most proper translation in the TL 3.

Full representation of the prototype is shown in Figure 4 and figure 5 respectively.

5. **Experiment and Results**

To judge the translation accuracy received by AMMT; we have tested our approach against human translation.

1. Test the prototype against the selected test examples.
2. Compare the output with the human translation.
3. Assign the reason behind the ill-translation to its corresponding category.
4. Assign a score (0-10) for each problem.
5. Compute Accuracy.

5.1 **Experiment**

Three well-regarded MT systems (i.e., Microsoft, Google, Yandex) are analyzed against our proposed system to evaluate the performance of the AMMT. In the first experiment, human judgment methodology is used for this purpose, while in the second experiment, we evaluate our system with iBLEU metric (Papineni et. al., 2002).

5.1.1 Human Judgment

Basically we have compared the output of our proposed system against the human translation, we have built a test example (test suit) out of 130 examples that were carefully selected from scientific books, popular media channels, the result is as shown in table and figure 6 below.

|                  | Microsoft | Google | Yandex | AMMT |
|------------------|-----------|--------|--------|------|
| Matches Sentences| 99        | 109    | 92     | 117  |
| Mismatches Sentences | 31       | 21     | 38     | 13   |
| Matches Sentences Total | 990      | 1090   | 920    | 1170 |
| Mismatches Sentences Total | 232.5    | 168    | 285    | 104  |
| Total Score      | 1222.5    | 1258   | 1205   | 1274 |
| Percentage       | 94.03%    | 96.76% | 92.69% | 98.0% |

To judge the evaluation properly, we have constructed a matrix to relate the issue of translation to certain score according to the following criteria:

1. **Def-Noun**: This problem arises when the system fails to distinguish between the articles "a(n)" or "the". 9.
Figure 5: System Architecture with example
2. Noun-Adj and Sub-Verb: 8.

3. Pronouns and Nouns: 8.

4. Subjects and Adjectives: 7.

5. Addition and Deletion: 7.

![Figure 6: Test Suit results](image)

Table shows error type in Microsoft, Google, Yandex, and AMMT, along with their frequencies. To illustrate the first row of Table (i.e., the Def-Noun agreement), we could notice that this particular issue has been shown 4 times in Microsoft, 4 times in Google, 4 times in Yandex, and twice in AMMT. Therefore, hence, Def-Noun agreement has been arisen 14 times in all systems under evaluation. Figure 6 and Figure 7 represent the frequencies of these issues after conducting the human judgment experiment.

Table 4. Type of Error Frequencies against AMMT

| Error | Error Type                  | Frequency | Error % | Microsoft | Google | Yandex | AMMT |
|-------|-----------------------------|-----------|---------|-----------|--------|--------|------|
| 1     | Def-Noun                    | 14        | 13.39%  | 4         | 4      | 4      | 2    |
| 2     | Noun-Adj                    | 16        | 15.53%  | 4         | 3      | 5      | 4    |
| 3     | Sub-Verb                    | 16        | 15.53%  | 6         | 2      | 6      | 2    |
| 4     | Nouns                       | 5         | 4.85%   | 0         | 3      | 1      | 1    |
| 5     | Pronouns                    | 16        | 15.53%  | 6         | 6      | 3      | 1    |
| 6     | Subjects                    | 13        | 12.62%  | 5         | 1      | 6      | 1    |
| 7     | Adjectives                  | 6         | 5.82%   | 2         | 1      | 2      | 1    |
| 8     | Successive words form an expression | 3 | 2.91% | 1 | 0 | 2 | 0 |
| 9     | Addition or Deletion        | 14        | 13.39%  | 3         | 1      | 9      | 1    |
| Frequencies of Errors | **103** | **31** | **21** | **38** | **13** |

The experiment shows that our system outperformed other systems with an average of 98.0, statistically speaking, only 2% out of the test examples have shown errors during the human judgment experiment.

5.1.2 The Bleu Evaluation

The BLEU metric ranges between 0 and 1, some translations may score 1, otherwise, they are quite similar. Due to this reason, even a human translator may not score 1. It is worth stressing that the higher score requires more reference translations per sentence. However, in this experiment, we compute the iBLEU scores (1gram, 2grams, and 3grams) for all test suit sentences. Afterward, we compute the overall average of each n-gram iBLEU scores. Table presents iBLEU score of Yandex against AMMT for 1gram, 2gram, and 3gram. Table and Figure 8 show the iBleu scores of the 2 systems against two references on the test suit mentioned above. As shown in Table the average score of 1 gram, 2gram and 3gram for Yandex is 0.60, 0.48 and 0.44 respectively, while the average score of 1gram, 2gram and 3gram for AMMT system is 0.90, 0.87 and 0.88 respectively. Thus, we can claim that the AMMT system outperform other well-regarded MT systems in the translation of specific-domain sentences.
Table 5. Experiment 2 Results: The iBLEU Score for Yandex and AMMT

| No. | Yandex | AMMT |
|-----|--------|------|
|     | n1-gram | n2 | n3 | n1-gram | n2 | n3 |
| S1  | 0.67    | 0.5 | 0.5 | 1        | 1  | 1  |
| S2  | 0.25    | 0.25| 0.25| 1        | 0.25 | 1  |
| S3  | 0.67    | 1   | 0.08| 0.5      | 1  | 0.6 |
| S4  | 0.5     | 0.5 | 1   | 1        | 1  | 1  |
| S5  | 0.17    | 0.5 | 0.13| 1        | 1  | 1  |
| S6  | 0.6     | 0.67| 0.08| 0.67     | 1  | 0.25|
| S7  | 0.67    | 0.33| 0.25| 1        | 1  | 1  |
| S8  | 0.5     | 0.13| 0.25| 1        | 0.83| 1  |
| S9  | 1       | 0.33| 1   | 0.5      | 1  | 0.6 |
| S10 | 1       | 0.33| 1   | 1        | 1  | 1  |
| S11 | 0.25    | 0.25| 1   | 1        | 1  | 1  |
| S12 | 0.33    | 0.13| 0.08| 1        | 1  | 0.6 |
| S13 | 0.33    | 0.5 | 1   | 0.67     | 0.25| 1  |
| S14 | 1       | 1   | 0.13| 1        | 1  | 1  |
| S15 | 1       | 0.25| 0.08| 1        | 0.5 | 0.5 |
| S16 | 0.67    | 1   | 0.25| 1        | 1  | 1  |
| S17 | 0.75    | 0.5 | 0.25| 0.67     | 0.83| 1  |
| S18 | 0.5     | 0.5 | 0.25| 1        | 1  | 1  |
| S19 | 0.75    | 0.67| 0.25| 1        | 0.83| 1  |
| S20 | 0.5     | 0.33| 1   | 1        | 1  | 1  |
| AVG | **0.605** | **0.483** | **0.441** | **0.900** | **0.874** | **0.877** |
6. Conclusion and Future Work

In this study, we developed a lexical MT system using a scalable transfer-based architecture for the translation of MSA into Latin-based Malay. The deliverable of this study: first: Arabic-Malay transformation structures development, second: the development of MT prototype based on transfer approach. third: Shed light on Arabic to Malay MT system challenges and proposes methods for handling them, and fourth: Test example development. These examples have been used in evaluating AMMT against Microsoft, Google, and Yandex. (i.e., "human judgment"), and iBLEU metric. The two experiments prove that AMMT has outperformed other systems.

References

Abdalla, H. (2012). Malay To Arabic Rules-Based Machine Translation Based on Word Ordering And Agreement. MSc Thesis, Universiti kebangsaan, Malaysia, Bangi.

Abodina, A. A., & Antisar, Y. A. (2012). Arabic to malay medical dialogue translation system based on the grammatical rules. Diss. MSc Thesis, Universiti kebangsaan Malaysia, Bangi, 2012.

Agiza, H. N., Ahmed, E. H., & Noura, S. (2012). An English-to-Arabic Prototype Machine Translator for Statistical Sentences. https://doi.org/10.4236/im.2012.41003

Almeshrky, H. A., & Mohd, J. A. A. (2012). Arabic Malay machine translation for a dialogue system. JApSc, 12(13), 1371-1377. https://doi.org/10.3923/jas.2012.1371.1377

Alqudsi, A., Nazlia, O., & Khalid, S. (2019). A Hybrid Rules and Statistical Method for Arabic to Malay Translation. 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS). IEEE, 2019. https://doi.org/10.1109/CAIS.2019.8769545

Alsaket, A. J., & Mohd, J. A. A. (2014). Arabic-malay machine translation using rule-based approach. Journal of Computer Science, 10(6), 1062. https://doi.org/10.3844/jcssp.2014.1062.1068

Attia, M. (2007). Arabic tokenization system. Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources. 2007. https://doi.org/10.3115/1654576.1654588

Badaro, G., Ramy, B., Hazem, H., Wassim, E. H., Khaled, B. S., Nizar, H., Ahmad, A. S., & Ali, H. (2019). A survey of opinion mining in Arabic: a comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(3), 1-52. https://doi.org/10.1145/3295662

Habash, N. (2008, June). Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In Proceedings of ACL-08: HLT, Short Papers (pp. 57-60).

Hamza, M. A., Mohd, J. A. A., & Nazlia, O. (2019). Identification of Sentence Context based on Thematic Role Rules for Malay Short Essay Assessment. International Journal of Software Engineering and Computer Systems, 5(2), 66-77. https://doi.org/10.15282/ijsecs.5.2.2019.5.0061

Papineni, K. et al. (2002). BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002. https://doi.org/10.3115/1073083.1073135

Safiah, K. N. et al. (2010). Tatabahasa Dewan Edisi Ketiga. Kuala Lumpur: Dewan Bahasa dan Pustaka (2010).

Shaalan, K. (2010). Rule-based approach in Arabic natural language processing. The International Journal on Information and Communication Technologies (IJICT), 3(3), 11-19.

Shaalan, K., Siddiqui, S., Alkhatib, M., & Monem, A. A. (2019). Challenges in Arabic natural language processing. Computational Linguistics. https://doi.org/10.1142/9789813229396_0003

Shquier, M. M. A. (2013). Computational Approach to the Derivation and Inflation of Arabic Irregular Verbs in English-Arabic Machine Translation. International Journal of Advancements in Computing Technology, 5(15).

Shquier, M. M. A. (2019). Novel Prototype for Handling Arabic Natural Language Processing: Smart Morphological Analyser. Second International Conference on Artificial Intelligence for Industries(AI4I) 1-8 (2019).

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).