УДК 004.912

Мусбах Заід Енвейджи, аспірант

Обробка арабської мови, заснована на правилах

Київський національний університет імені Тараса Шевченка, 03680, м. Київ, пр-т. Академіка Глушкова 4д,
e-mail: s.cc85@yahoo.com

Musbah Zaid Enweiji postgraduate

Arabic natural language processing with Rule based approach features

Taras Shevchenko National University of Kyiv, 03680, Kyiv, Glushkova st., 4d,
e-mail: s.cc85@yahoo.com

Підхід на основі правил був успішно використаний при розробці багатьох систем обробки природньої мови. Ця стаття присвячена вивченню проблем та особливостей використання даного підходу до сімейства арабських мов. Особливу увагу було приділено характеристикам арабської мови та показано можливості машинного перекладу, заснованого на правилах.

Ключові слова: обробка арабської мови, машинний переклад, морфологічний аналіз

The rule-based approach has successfully been used in developing many natural language processing systems. Systems that use rule-based transformations are based on a core of solid linguistic knowledge. The linguistic knowledge acquired for one natural language processing system may be reused to build knowledge required for a similar task in another system. In this paper we explained how to involved rule-based approach for different Arabic natural language processing tasks. Also We focused on the characteristics of Arabic language as rich morphologically language and the rule of ruled based approach in some applications that can be used such as machine translation .

Key Words: Arabic language processing, machine translation , morphological analysis

I. INTRODUCTION

ARABIC is a Semitic language spoken by more than 330 million people as a native language, in an area extending ill-formed learner input. Furthermore, the linguistic knowledge acquired for one natural language processing from the Arabian/Persian Gulf in the East to the Atlantic Ocean in the West. Moreover, it is the language in which 1.4 billion Muslims around the world perform their daily prayers. Arabic is a highly structured and derivational language where morphology plays a very important role [2].

Over the last few years, Arabic natural language processing (ANLP) has gained increasing importance, and several state of the art systems have been developed for a wide range of applications. These applications had to deal with several complex problems pertinent to the nature and structure of the Arabic language [7]. The lack of available resources and their limitations have motivated many scholars to follow the rule- based approach and rely on hand-constructed linguistic rules in developing their tools, systems, and resources. ANLP tools based on this approach generally include morphological analyzers/generators and syntactic analyzers/generators. The approach is also used for some specific tasks. Moreover, rule-based ANLP systems include machine translators, named entity recognizers, and intelligent computer assisted language learning systems.

Systems and tools that use rule-based transformations are based on a core of solid linguistic knowledge [8]. The characteristics of a rule-based approach are :

1 . It has a strict sense of well-formedness in mind .

2 . It imposes linguistic constraints to satisfy well- formedness .

3 . It allows the use of heuristics (such as a verb cannot be

preceded by a preposition)

.

4 . It relies on hand-constructed rules that are to be acquired from language specialists rather than automatically trained from data.

The advantages of this approach are that it is easy to incorporate domain knowledge into the linguistic knowledge which provides highly accurate results. Domain rules have been used in generating Arabic sentences [9] and analysis of ill-formed learner input [1]. Furthermore, the linguistic knowledge acquired for one natural language processing system may be reused to build knowledge required for a similar task in another system.

Because time was of the essence, and in the absence of complete computationally viable grammars of Arabic, statistical approaches that rely primarily on training data and parallel texts gained momentum. Machine learning approach usually gives good results when the training set and the testing data are similar. There is also a point at which more training data does not make significant improvement. Moreover, there may be some structures or entities that are sparse. In this case the machine learning component does not have enough data to make the right generalization. Regardless of sparseness of the data, the statistical-based or machine learning approaches have some difficulties with specific natural language processing tasks such as distinguishing between well-formed and ill-formed input, whereas the rule-based approaches have the advantage of providing detailed analyses of the Arabic learner's answer using linguistic (morphological and syntactic) knowledge which, in applications such intelligent tutoring systems, enables feedback elaboration that helps learners to understand better their knowledge gab.

II. ASPECTS OF THE ARABIC LANGUAGE

Arabic is rooted in the Classical or Quranic Arabic, but over the centuries, the language has developed to what is now accepted as MSA. MSA is a simplified form of Classical Arabic, and follows its grammar [8]. The main differences between Classical Arabic and MSA are that MSA has a larger (more modern) vocabulary, and does not use some of the more complicated forms of grammar found in Classical Arabic. For example, short vowels are omitted in MSA such that letters of the Arabic text are written without diacritic signs. The Arabic language is written from right to left. It has 28 letters, some of which have one form (like " د"), while others have two forms (" س " ;"ســ "), three forms (" ه" ;" ـهـ " ;"هـ ") or four forms ("ع" ; "ـع; ""ـعـ; " عـ") [14]. Arabic words are generally classified into three main categories [19]: noun, verb, and particle.

Arabic is a language of rich and complex morphology, both derivational and inflectional [7]. Word derivation in Arabic involves three concepts the root, pattern, and form. Word forms (e.g. verbs, verbal nouns, agent nouns, etc.) are obtained from roots by applying derivational rules to obtain corresponding patterns. Generally, each pattern carries a meaning which, when combined with the meaning inherent in the root, gives the target meaning of the lexical form. For example, the meaning of the word form " كاتب " (writer) is the combination of the meaning inherent in the root " كتب "(write) and the meaning carried by the pattern (or 'template') " ف -ا-ع-ل " (fa'il) which is the pattern of the doer of the root. Arabic inflectional morphology involves adding morph syntactic features such as tense, number, person, case, etc. Arabic also has some more morphological peculiarities. For example, an indefinite word can be made definite by attaching the prefix definite article " الـ "(the) to it, but there is no indefinite article. As another example, a verb can take affix pronouns such as " سأعطيكما " (will-I-give-you); this also shows that the verb is conjugated with the dual suffix pronoun " كما "(you). An Arabic inflected verb can form a complete sentence, e.g. the verb " سمعتك " (heard-I-you) contains a complete syntactic structure in just a one-word sentence. Moreover, the rich morphology of Arabic allows the dropping of the subject pronoun ('pro-drop'), i.e. to have a null subject when the inflected verb includes subject affixes. There are two types of Arabic sentences [10]: nominal and verbal sentence. An Arabic compound sentence is formed from a simple sentence followed by a complementary sentence [11], such as a conjunction form ,(عطف) e.g " نحن نرغب في تأجير سيارة وسنحتاج لساحة انتظار قريبة من الفندق " .(We want to rent a car and we- will-need to park near the-hotel), or a quasi-sentence شبه جمله ,( ) e.g " بالفندق " .(in-the-hotel). Agreement is a major syntactic principle that affects the generation of an Arabic sentence. Agreement in Arabic is full or partial and is determined by word order [10]. An adjective in Arabic usually follows the noun it modifies) "الموصوف" (and fully agrees with it with respect to number, gender, case, and definiteness. The verb in Verb-Subject-Object order agrees with the subject in gender, e.g " جاء الولد/الأولاد " .(came the- boy/the-boys) versus " جاءت البنت

" البنات/(came the-girl/the- girls). In Subject-Verb-Object (SVO) order, the verb agrees with the subject with respect to number and gender, e.g الولد جاء " . " الأولاد جاءوا/(came the-boy/the-boys) versus البنت " " البنات جئن / جاءت(came the-girl/the-girls).

III. MACHINE TRANSLATION

Machine translation (MT) is the area of information technology and applied linguistics dealing with the translation of human languages such as English and Arabic. There are three different approaches of rule-based translation systems : direct, transfer, and interlingual. The simplest approach is the direct translation approach where a word-by-word translation (lexical transfer) from the source language to the target language is performed. From a linguistic point of view, what is missing in this approach is any analysis of the internal structure of the source text, particularly the grammatical relationships between the constituents of the sentences. Such systems gave the kind of translation that was characterized by frequent mistranslations at the lexical level and largely inappropriate syntactic structures which mirrored too closely those of the source language.

In the transfer approach, 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. We developed a transfer-based machine translation system of English noun phrase to Arabic. The architecture is shown in Figure 1 [13]. In our noun phrase translator, the actual translation occurs in the transfer step in which one side of the tree-to-tree transfer rules is matched against the input structure, resulting in the structure on the right-hand-side. Figure 2 [14] illustrates the translation of the noun phrase (NP) “networks performance evaluation” into “تقييم أداء شبكة ”, which shows the switching of words that is indicated by the following transfer rule:

$$[w_i : \$1, w_{i+1}: \$2, \ldots, w_k: \$\ k] \qquad (1\le i\le k)$$

$$[w_k : \$\ k, w_{k-1}:\$\ k\text{-}1, \ldots, w_i: \$\ i] \qquad (1\le i\le k)$$

This rule says that the translation of the word at level i is switched with the word at level k-i+1 . Where k is the number of NPs equivalent to maximum (sub)tree level. Transfer-based approach in some cases cannot give a correct transfer of distance relationships. For example, consider the translation of “intelligent tutoring systems” into “نظم التعليم الذكي ” (intelligent.adj.fem tutoring.noun.masc systems.noun.fem) and incorrectly takes the noun directly preceding the adjective as the noun it modifies.

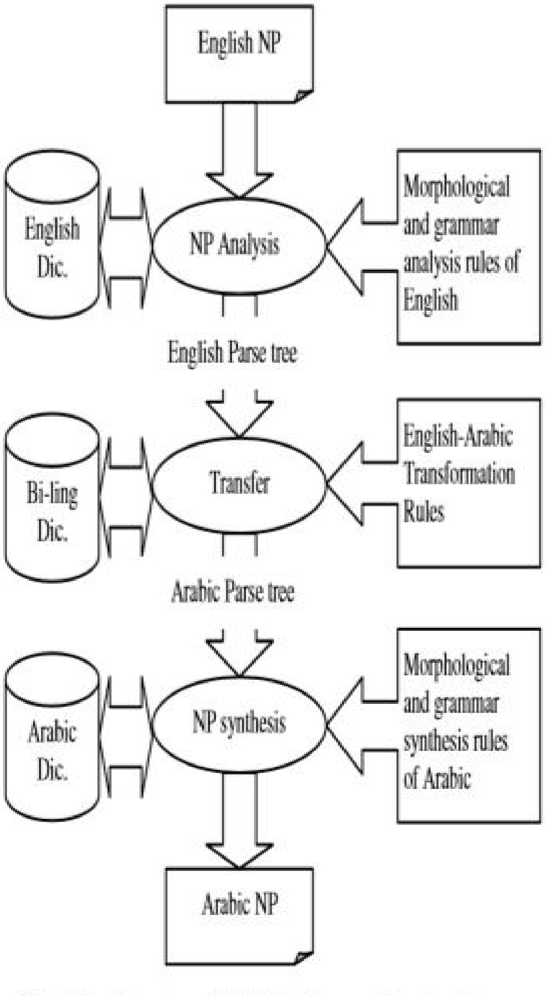
English NP
English Dic.
NP Analysis
Morphological and grammar analysis rules of English
English Parse tree
Bi-ling Dic.
Transfer
English-Arabic Transformation Rules
Arabic Parse tree
Arabic Dic.
NP synthesis
Morphological and grammar synthesis rules of Arabic
Arabic NP

Fig. 1 Overall structure of English-Arabic noun phrase translator

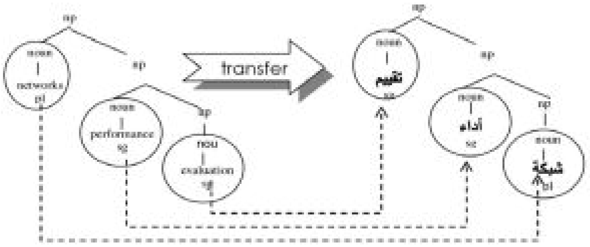

Fig. 2 . Simple transfer

The Interlingual approach MT is used successfully in multilingual translation. It aims to achieve the translation task in two independent steps. First, meanings of the source- language sentences are represented in an intermediate language-independent (Interlingua) representation. Then, sentences of the target language are generated from those meaning representations. The Interlingua representation plays an important role in the accuracy of the translation as it should be a language-neutral representation and captures the intended meaning of the source sentence.

IV. CONCLUSION

We have presented the rule-based approach in Arabic natural language processing. One possible criticism of the rule- based approach is that it is a traditional and widely studied topic especially when it regards to European languages. The current studies and researches still gives a steps towards helping Arabic language technology to catch up with other mature languages technology such as English.

The rapid development of rule-based systems is feasible, especially in the absence of linguistic resources and the difficulties faced in adapting tools from other languages due to peculiarities the nature of Arabic language. Finally, the necessity of adopting general solutions as much as possible, as this increases the chances that the linguistic knowledge and tools can be used in other domains and systems as well.

Список використаних джерел

1. *Magdy M.* Lexical Error Diagnosis for Second Languag Learners of Arabic. / M. Magdy, K. Shaalan, A. Fahmy // In the Proceedings of The Seventh Conference on Language Engineering, Egyptian Society of Language Engineering (ELSE). - Cairo, Egypt, 2007. – pp. 5-6.

2. *Attia M.* A large scale computational processor of Arabic morphology and applications / M. Attia. - Cairo University, Egypt. - 1999. – 223p.

3. *Farghaly A.* Arabic Natural language processing challenges and solution, ACM Transaction on asian language information processing (TALIP) / Farghaly A., Shaalan K.. Proceedings to the Association for Computing Machinery (ACM). TALIP Vol. 8, Issue 4, December 2009. – pp.2-10.

References

1. MAGDY M., SHAALAN K., FAHMY A.. (2007) Lexical Error Diagnosis for Second Language Learners of Arabic. *In the Proceedings of The Seventh Conference on Language Engineering, Egyptian Society of Language Engineering (ELSE), Cairo, Egypt.* – pp. 5-6.

2. ATTIA M.(1999) A large scale computational processor of Arabic morphology and applications. Cairo University, Egypt.

3. FARGHALY A., SHAALAN K. (2009) Arabic Natural language processing challenges and solution, ACM Transaction on asian language information processing (TALIP), *Proceedings to the Association for Computing Machinery (ACM). TALIP* Vol 8, Issue 4. - pp.2-10.

4. *Abdel Monem A.* Generating Arabic Text in Multilingual Speech-to-Speech Machine Translation Framework, Machine Translation /A. Abdel Monem, K. Shaalan, A. Rafea, H. Baraka. - Cairo University, Egypt. - 2008. – 350 p.

5. *Ryding K.* Reference Grammar of Modern Standard Arabic / K. Ryding. - Cambridge University Press, Cambridge, UK. - 2005. – 150 p.

6. *Mace J.* Arabic Grammar: A Reference Guide. /J.Mace. -Edinburgh University Press, Edinburg, UK. - 1998. – 200p.

7. *Trujillo A.* Translation Engines: Techniques for Machine Translation / A.Trujillo. - Springer Verlag, USA. - 1999. – 220p.

8. *Shaalan K.* Machine Translation of English Noun Phrases into Arabic / K. Shaalan, A Rafea, A .Abdel Monem, H.Baraka // The International Journal of Computer Processing of Oriental Languages (IJCPOL),World Scientific Publishing Company, 17(2), 2004. – p.121-134.

9. *Shaalan K.* Rule-based Approach in Arabic Natural Language Processing, International / K. Shaalan // Journal on Information and Communication Technologies, Vol. 3, No. 3, June 2010. - pp.3-7.

4. ABDEL MONEM A., SHAALAN K., RAFEA A., BARAKA N. (2008) Generating Arabic Text in Multilingual Speech-to-Speech Machine Translation Framework, Machine Translation. Cairo University, Egypt.

5. RYDING K. (2005) Reference Grammar of Modern Standard Arabic, Cambridge University Press, Cambridge, UK.

6. MACE J. (1998) Arabic Grammar: A Reference Guide. Edinburgh University Press, Edinburgh, UK.

7. TRUJILLO A. (1999) Translation Engines: Techniques for Machine Translation, Springer Verlag, USA.

8. SHAALAN K., RAFEA A., ABDEL MONEM A., BARAKA N. (2004) Machine Translation of English Noun Phrases into Arabic, The International Journal of Computer Processing of Oriental Languages (IJCPOL),World Scientific Publishing Company, 17(2):121-134.

9. SHAALAN K. (2010) Rule-based Approach in Arabic Natural Language Processing. International Journal on Information and Communication Technologies, Vol. 3, No. 3, pp.3-7.

Надійшла до редколегії 01.12.2015