A Bilingual Approach for Arabic Paraphrases Acquisition: Preliminary Experiments

Rania Al-Sabbagh **Khaled Elghamry**

re832003@yahoo.com elghamryk@ufl.edu

Faculty of Al-Alsun (Languages), Ain Shams University, Cairo, Egypt

Abstract

This paper presents preliminary experiments on a bilingual approach for Arabic paraphrase acquisition; a research which is motivated by the importance of paraphrasing for overcoming sparseness of data and its importance for many NLP applications such as Question Answering (QA) and Information Retrieval (IR). The proposed approach develops an unsupervised bilingual algorithm to acquire Arabic paraphrases at the phrase level which is rather more challenging than the elementary word-level paraphrasing and is less efficiently handled by current Arabic paraphrasing systems. Preliminary results show that our approach manages to get term variations – orthographic, lexical and syntactic – for ~ 70% of 4000 randomly selected phrases.

I. Introduction

To paraphrase is to restate the same information using different lexical and/or syntactic structures. According to Callison-Burch (2007), paraphrasing proves to be an effective technique to overcome the inherent problem of Statistical Natural Language Processing (SNLP), namely sparseness of data. Moreover, it is an essential intermediate task for many Natural Language Processing (NLP) applications such as Question Answering (QA) – discovering paraphrased answers may provide additional evidence that an answer is correct (Ibrahim et al. 2003) – and Machine Translation (MT) (Elghamry 2007).

Paraphrasing is classified into word-based, phrase-based, sentence-based, paragraph-based and text-based paraphrasing. Current experiments focus on phrase-based paraphrasing for two main reasons. First, it is more challenging than lexical paraphrasing (i.e. synonymy identification) which is relatively simple due to the widespread of machine-readable thesauri. Second, the performance of current Arabic paraphrasing systems on phrase-based paraphrasing still needs improvement. Experiments focus on two types of phrases: named entities (e.g. names of organizations, locations, persons ... etc.) and common noun phrases.

According to Callison-Burch (2007), bilingual paraphrasing approaches outperform monolingual ones for many languages including Arabic. However, being based on parallel and/or comparable corpora, these approaches might not be practical for languages with scarce resources like Arabic. Therefore, the proposed bilingual approach tries to dispense with such corpora, meanwhile go unsupervised and robust. Preliminary experiments show promising results about acquiring orthographic, lexical and syntactic phrase-based paraphrases.

The rest of this paper falls in four parts. The first part reviews related work to bilingual paraphrase acquisition. The second part explains the proposed approach, its tools and implementation. The third part shows the used evaluation methodology and results. Finally, the paper ends with a conclusion of the main findings of the preliminary experiments and future work for a full-scale application of the proposed approach.

II. Related Work

Previous bilingual approaches to paraphrasing relay on one of three resources: multiple translations, comparable corpora (Quirk et al. 2004) and parallel corpora (Callison-Burch 2007). Multiple translations approaches – which are applied to

English and French (Barzilay and McKeown 2001) and English and Chinese (Pang et al. 2003) – assume that different translations of the same source text paraphrase one another. In spite of the promising results achieved by such approaches, the scarcity of multiple translations and the fact that developing them manually is time and effort consuming are obstacles for a full-scale coverage.

Approaches using comparable and parallel corpora achieve better results than multiple translations in terms of coverage, especially for such languages with available corpora such as English (Quirk 2004, Callison-Burch 2007). Callison-Burch (2007) used parallel corpora for Arabic paraphrase acquisition using the only available source for Arabic parallel corpora, namely the LDC Arabic/English Parallel News Text[1]. No clear results are reported on applying this approach to Arabic; however, the approach is used to build the freely available Arabic paraphrase systems Linear B (http://linearb.co.uk/) and Lingo24 (www.lingo24.com).

Practical experience shows that these two systems perform better on the word-based paraphrasing than phrase-based paraphrasing for two reasons. First, parallel and comparable corpora for Arabic, though available, are still scarce. To the best of the authors' knowledge, the only ones available are LDC Arabic/English Parallel News Text[1] and ISI Arabic-English Automatically Extracted Parallel Text[2]. Second, using parallel and/or comparable corpora entails using alignment techniques, which pose

[1] A corpus of Arabic news stories and their English translations collected via Ummah Press Service from January 2001 to September 2004. It totals 8,439 story pairs, 68,685 sentence pairs, 2M Arabic words and 2.5M English words. The corpus is aligned at sentence level. It is available through Linguistic Data Consortium (LDC) catalog number LDC2004T18, URL: http://www.ldc.upenn.edu/

[2] An Arabic-English comparable corpus which is automatically extracted from news articles published by Xinhua News Agency and Agence France Presse. It is obtained using the automatic parallel sentence identification method described in Stefan, D. and Marcu, M. (2005). Machine Translation Performance by Exploiting Non-parallel Corpora, *Computational Linguistics*, Vol. 31. pp. 477-504. The corpus contains 1,124,609 sentence pairs; the word count on the English side is approximately 31M words.

another source of errors. Therefore, the proposed approach tries to avoid alignment and to find an alternative for both parallel and comparable corpora.

III. The Bilingual Paraphrasing Approach

The proposed approach is based on the same hypothesis previously used by Barzilay and McKeown (2001): different translations of the same source text are paraphrases of one another. However, instead of using corpora of multiple translations, our approach generates necessary multiple translations using current Machine Translation (MT) systems such as Microsoft Translator, Google and Golden Al-Wafi (ATA 2002). It is also assumed that different MT systems use different dictionaries and are trained on different corpora; thus they are likely to yield different translations based on their different dictionaries, corpora and rules.

The approach is straightforward; it does not require any corpus preprocessing tasks and it does not rely on intermediate NLP tools such as POS taggers, NP chunkers or parsers. Therefore, the authors save time and effort; and minimize the sources of errors to one source only, namely the problems of the MT systems used.

Due to lexical and syntactic MT problems, a necessary phase of the proposed approach is MT output validation; that is, to validate the output against documents originally written in the target language (here Arabic). However, even with using Web documents, many rare, yet correct, translations yield zero search hits. For instance, the "National Center for Environmental Research" is translated by Golden Al-Wafi as "المركز الوطني للبحث البيئي" /*Almrkz AlwTny llbHv Alby}y*/[3], which is a correct translation yet it gets zero search results on Google search engine. Therefore, relying on the regular Web validation technique, which uses the entire phrase as a search query, might not be helpful.

[3] Buckwalter's transliteration scheme (www.qamos.com)

Alternatively, we used a bigram-based term validation technique. We divide each translated phrase into consecutive bigrams and check the validity of each bigram independently on the Web. Each valid bigram is given a score of 1 and each invalid bigram a score of 0 (zero). The validity of the phrase is, therefore, measured as:

$$\textit{Phrase Validity} = \frac{\textit{Sum of valid bigrams}}{\textit{Total number of bigrams}}$$

A score of 1 is the maximum attained for an entirely valid phrase and a 0 score indicates an invalid translation. Intuitively, only phrases giving scores ≥ 0.8 are considered as valid.

Accordingly, given the aforementioned example of the "National Center for Environmental Research", Golden Al-Wafi translates it as "المركز الوطني للبحث البيئي" /*Almrkz AlwTny llbHv Alby}y*/,Microsoft Translator as"المركز الوطني للبحوث البيئية" /*Almrkz AlwTny llbHwv Alby}yp*/and Google as "المركز الوطنى لبحوث البيئة" /*Almrkz AlwTnY lbHwv Alby}p*/. The validation of each translation is measured as such:

	Bigram 1	**Bigram 2**	**Bigram 3**	**Phrase Validity Score**	**Result**
Al-Wafi	المركز الوطني /*Almrkz AlwTny*/	الوطني للبحث /*AlwTny llbHv*/	للبحث البيئي /*llbHv Alby}y*/	1	Valid Translation
	1	1	1		

	Bigram 1	**Bigram 2**	**Bigram 3**	**Phrase Validity Score**	**Result**
Google	المركز الوطني /*Almrkz AlwTny*/	الوطني لبحوث /*AlwTny lbHwv* /	لبحوث البيئة /*lbHwv Alby}p*/	1	Valid Translation
	1	1	1		

	Bigram 1	**Bigram 2**	**Bigram 3**	**Phrase Validity Score**	**Result**
Microsoft Translator	المركز الوطني */Almrkz AlwTny/*	الوطني للبحوث */AlwTny llbHwv/*	للبحوث البيئية */llbHwv Alby}yp/*	1	Valid Translation
	1	1	1		

Table (1): Example of Phrase Validation Process

Given three valid translations of the same source phrase – "National Center for Environmental Research", the three translations are considered as paraphrases of one another.

To sum up, the algorithm informally goes in four phases:

1. First, compiling source phrases: the source language for the present study is English whose resources – basically parsers and annotated corpora – are quite available.
2. Second, submitting source phrases to MT systems.
3. Third, implementing the bigram-based term validation.
4. Finally, selecting phrases with a score ≥ 0.8.

In spite of the problems of MT systems, using them is expected to achieve better coverage rates than parallel corpora, especially in terms of term variations. Meanwhile, using MT systems makes the proposed approach language independent and thus more applicable. The evaluation methodology and the results of our preliminary experiments and an error analysis are presented in the following subsections.

IV. Evaluation and Results

In order to test our approach, a list of 2000 named-entities (i.e. names of organizations and locations) is compiled using Google search engine. Another list of 2000 common NPs is extracted from the British National Corpus (BNC). Each list is submitted to each of the used MT systems: Microsoft Translator, Google and Golden Al-Wafi (ATA 2002).

For evaluation, a human rater is used for two purposes: first, to evaluate the MT output (i.e. to judge it as valid/invalid and) and to decide whether valid translations are paraphrasing; second, to measure the agreement rate between the human rater and the results of the bigram-based term validation according to the Kappa Coefficient.

Kappa Coefficient is a statistical measure for the agreement between two raters, taking into consideration the difference between actual or observed agreement and agreement given by chance. It is defined as:

$$Kappa = \frac{P(o) - P(e)}{1 - P(e)}$$

Where
P(o) is the probability of observed agreement
P(e) is the probability of expected agreement

The human rater and the bigram-based term validation achieve a good kappa rate of ~80%. The main differences between the two raters are among the phrases scoring around 0.8. For instance, the "National Center on Addiction and Substance Abuse" is translated given the following three translations:

1. "المركز القومي للادمان وتعاطي المخدرات" (Google)

   /*Almrkz Alqwmy llAdmAn wtEATy AlmxdrAt*/

2. "المركز الوطني على سوء إستخدام المادة والإدمان" (Al-Wafi)

   /*Almrkz AlwTny ElY sw' <stxdAm AlmAdp wAl<dmAn*/

3. "المركز الوطني بشأن والادمان تعاطي المواد" (Microsoft Translator)

   /*Almrkz AlwTny b$>n wAlAdmAn tEATy AlmwAd*/

According to the abovementioned bigram-based term validation techniques, these translations are given the scores of 1, 0.83 and 0.83, respectively. Thus they considered as valid by the bigram-based term validation, yet the second and third translations are invalid according to the human rater being semantically and syntactically incorrect.

Final results of our approach can be summarized as follows:

Paraphrases of score 1 each	**Recall**	**Precision based on human rater's evaluation**	**F-measure**
	~ 47.5%	~ 86%	~ 61%
Paraphrases of score ≥ 0.8 each	**Recall**	**Precision based on human rater's evaluation**	**F-measure**
	~ 5%	~ 50%	~ 9%

Table (2): Final Results

Generated paraphrases can be divided into three classes: orthographic, lexical and syntactic paraphrases. Orthographic paraphrases are paraphrases with the same lexical and syntactic structures yet with different orthographic forms for such letters as ء /'/ (*hamza*) and ة /*p*/ (*teh marbuta*). Examples of orthographic paraphrases are given in table (3) below:

Source Phrase	**Paraphrase 1**	**Paraphrase 2**	**Orthographic Difference bet. the Two Paraphrases**
National Center for Simulation	المركز الوطني **للمحاكاة** /*Almrkz AlwTny* ***llmHAkAp***/	المركز الوطني **للمحاكاه** /*Almrkz AlwTny* ***llmHAkAh***/	The last word is written with ة /*p*/ (*teh marbuta*) first and then with ه /*h*/ (*heh*)
National Center for Theoretical Sciences	المركز الوطني للعلوم **النظرية** /*Almrkz AlwTny llElwm* ***AlnZryp***/	المركز الوطني للعلوم **النظريه** /*Almrkz AlwTny llElwm* ***AlnZryh***/	The last word is written with ة /*p*/ (*teh marbuta*) first and then with ه /*h*/ (*heh*)
National Bank of Egypt	البنك **الأهلي** المصري /*Albnk* ***Al>hly*** *AlmSry*/	البنك **الاهلي** المصري /*Albnk* ***AlAhly*** *AlmSry*/	The second word is written with أ /*>*/ (*alef with hamza above*) first and

			then with ا /*A*/ (*bare alef: alef with no hamza*)
The Egyptian Organization for Human Rights	المنظمه المصرية لحقوق **الانسان** /*AlmnZmh AlmSryp lHqwq* ***AlAnsAn***/	المنظمة المصرية لحقوق **الإنسان** /*AlmnZmp AlmSryp lHqwq* ***Al<nsAn***/	The last word is written with ا /*A*/ (*bare alef: alef with no hamza*) and then with إ /*<*/ (*alef with hamza under*)

Table (3): Examples of Orthographic Paraphrases

The second category of the generated paraphrases is lexical paraphrases. These are paraphrases that contain synonymous words like the ones in table (4) below:

Source Phrase	**Paraphrase 1**	**Paraphrase 2**	**Lexical Difference bet. the Two Paraphrases**
National Center for Higher Education Management Systems	المركز الوطني **لأنظمة** إدارة التعليم العالي /*Almrkz AlwTny* ***l>nZmp*** *<dArp AltElym AlEAly*/	المركز الوطني **لنظم** إدارة التعليم العالي /*Almrkz AlwTny* ***lnZm*** *<dArp AltElym AlEAly*/	The two synonymous words are: أنظمة /*>nZmp*/ and نظم /*nZm*/; both of which mean systems
International Organization for Conservation of Cultural Heritage	المنظمة الدولية **لحماية** التراث الثقافي /*AlmnZmp Aldwlyp* ***lHmAyp*** *AltrAv AlvqAfy*/	المنظمة الدولية **للحفاظ على** التراث الثقافي /*AlmnZmp Aldwlyp* ***llHfAZ ElY*** *AltrAv AlvqAfy*/	The two synonymous words are: حماية /*HmAyp*/ and حفاظ /*HfAZ*/; both of which mean conservation
Egyptian Association against Torture	الجمعية المصرية **ضد** التعذيب /*AljmEyp AlmSryp* ***Dd*** *AltE*yb*/	الجمعية المصرية **لمناهضه** التعذيب /*AljmEyp AlmSryp* ***lmnAhDh*** *AltE*yb*/	The two synonymous words are: ضد /*Dd*/ and مناهضه /*mnAhDh*/; both of which mean against

Table (4): Examples of Lexical Paraphrases

The last category of the resulting paraphrases is the syntactic paraphrases. This means that the same phrase is given in different syntactic structures as in table (5):

Source Phrase	**Paraphrase 1**	**Paraphrase 2**	**Syntactic Difference bet. the Two Paraphrases**
National Center for Environmental Research	المركز الوطني **للبحوث البيئية** */Almrkz AlwTny* ***llbHwv Alby}yp****/*	المركز الوطنى **لبحوث البيئة** */Almrkz AlwTnY* ***lbHwv Alby}p****/*	The first ends with an Adjectival Phrase (ADJP) whereas the second with a NP
National Association of Social Workers	الجمعية الوطنية **لموظفي الخدمات الاجتماعية** */AljmEyp AlwTnyp* ***lmwZfy AlxdmAt AlAjtmAE****yp/*	الرابطه الوطنية **للاخصائيين الاجتماعيين** */AlrAbTh AlwTnyp* ***llAxSA}yyn AlAjtmAEyyn****/*	The first ends with a NP whereas the second with a ADJP
further information	**مزيد من المعلومات** */****mzyd mn AlmElwmAt****/*	**معلومات أخرى** */****mElwmAt >xrY****/*	The first is a compound NP which includes an ADJP whereas the second is a simple NP

Table (5): Examples of Syntactic Paraphrases

There are paraphrases that include more than one difference like the phrases included in table (6):

Source Phrase	**Paraphrase 1**	**Paraphrase 2**	**Types of Difference bet. the Two Paraphrases**
National Center for the Preservation of Democracy	المركز الوطني **لحفظ الديمقراطية** */Almrkz AlwTny lHfZ AldymqrATyp/*	المركز الوطني **للحفاظ على الديمقراطيه** */Almrkz AlwTny llHfAZ ElY AldymqrATyh/*	Orthographic & Syntactic
National Association for Retired Firefighters	الجمعية الوطنية **لرجال الإطفاء المتقاعدين** */AljmEyp AlwTnyp lrjAl Al<TfA' AlmtqAEdyn/*	الرابطه الوطنية **للمتقاعدين من رجال الاطفاء** */AlrAbTh AlwTnyp llmtqAEdyn mn rjAl AlATfA'/*	Orthographic & Syntactic

Table (6): Examples of Multiple Differences between Paraphrases

The 30% loss of the performance rate is attributed to two main reasons. First, the three MT systems yielded exactly the same translation for 7% of the tested phrases; and thus no paraphrases were available. Second, 23% of the output translations are linguistically unacceptable; that is, they include lexical and/or syntactic errors.

V. Conclusion and Future Work

This paper presented the initial experiments for an unsupervised bilingual approach for Arabic paraphrases acquisition. Managing to extract different term variations (i.e. term paraphrases) – orthographic, lexical and syntactic – for 71% of the tested phrases shows that it is a promising approach. It deals with phrase-based paraphrasing which is poorly handled by current Arabic paraphrasing systems and is not limited to the phrases present in parallel corpora. Meanwhile, it does not require much preprocessing or NLP tools.

The main problem of the present approach was the law recall rates. Approximately, 7% of input phrases were given the same translation by all the used MT systems like "World Health Organization"; it was translated by all systems as منظمة الصحة العالمية /*mnZmp AlSHp AlEAlmyp*/ and no system translates it as المنظمة العالمية للصحة /*AlmnZmp AlEAlmyp llSHp*/ which is a correct translation that gets 7,720 search hits on Google search engine. In order to get such paraphrases, the authors expect for future work to integrate the proposed bilingual algorithm with monolingual paraphrasing rules automatically acquired from the bilingually generated paraphrases. Such rules might also contribute to finding paraphrases for the terms mistakenly translated by MT systems.

Although the bigram-based term validation achieves a good kappa rate with human evaluation, there should be more variables to test. For example, trigrams are to be compared with bigrams in terms of precision. The threshold of ≥ 0.8 gets a rather poor precision rate and thus higher thresholds are to be tested together with their effect on recall rates.

References

ATA Software Technology Ltd. (2002). Golden Al-Wafi Translator Software. Version 1.12 [Online]. Available: www.**ata**soft.com Accessed: 21 April 2001.

Barzilay, R. and McKeown, K. (2001). Extracting Paraphrases from a Parallel Corpus. In *Proceeding of ACL-2001*.

Callison-Burch, C. (2007). Paraphrasing and Translation. Unpublished PhD. School of Informatics, University of Edinburgh, UK.

Elghamry, K. (2007). Machine Translation Oriented Syntactic Normalization of Noun Phrases in Arabic. *Proceedings of Information and Communication Technologies International Symposium (ICTIS07): Workshop on Arabic Natural Language Processing*, Morocco, 2007.

Ibrahim, A., Katz, B., and Lin, J. (2003). Extracting Structural Paraphrases from Aligned Monolingual Corpora. In *Proceedings of the Second International Workshop on Paraphrasing (ACL 2003).*

Pang, B., Knight, K., and Marcu, D. (2003). Syntax-Based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In *Proceedings of HLT/NAACL*

Quirk, C., Brockett, C., and Dolan, W. (2004). Monolingual Machine Translation for Paraphrase Generation. In EMNLP-2004.

Appendix: Sample Arabic Paraphrases & their Scores Generated by Our Approach

Source Phrase	Paraphrase 1	Score 1	Paraphrase 2	Score 2	Paraphrase 3	Score 3
National Center for Public Policy Research	المركز الوطني لبحوث السياسات العامة	1	المركز الوطني للبحوث في مجال السياسة العامة	1	-------	
National Center for Atmospheric research	المركز الوطني لابحاث الغلاف الجوي	1	المركز القومي لبحوث الغلاف الجوي	1	-------	-------
National Center for Higher Education Management Systems	المركز الوطني لنظم إدارة التعليم العالي	1	المركز الوطني لأنظمة إدارة التعليم العالي	1	-------	-------
National Center for Health Statistics	المركز الوطني للاحصاءات الصحية	1	المركز الوطني لإحصائيات الصحة	1	-------	-------
National Center for Public Productivity	المركز الوطني لمعدل الإنتاج العام	1	المركز القومي للانتاجيه العامة	1	-------	-------
The Mobilization of Muslim Women in Egypt	تعبئة النساء المسلمات في مصر	1	تعبئة المرأة المسلمة في مصر	1	-------	-------
Federation of Egyptian Chambers of Commerce	إتحاد غرف التجارة المصرية	1	اتحاد الغرف التجارية المصرية	1	-------	-------
European Bank for Reconstruction and Development	المصرف الأوروبي للإنشاء والتعمير	1	البنك الاوروبي للاعمار والتنمية	1	-------	-------

social workers	الاخصائيون الاجتماعيون	1	موظفو الخدمات اجتماعية	1	-------	-------
Whole World	كل عالم	1	العالم كله	1	-------	-------
European cup	الكأس الأوروبي	1	كأس اوروبا	1	-------	-------
National Center for Health Education	المركز الوطني للتثقيف الصحي	1	المركز الوطني لتعليم الصحة	1	المركز الوطني للتثقيف في مجال الصحة	1
National Center for Environmental Research	المركز الوطني للبحوث البيئية	1	المركز الوطنى لبحوث البيئة	1	المركز الوطني للبحث البيئي	