Pre-processing and Language Analysis for Arabic to French Statistical Machine Translation

Fatiha Sadat Emad Mohamed
Université du Québec à Montréal
201 Président Kennedy, Montréal
H2X 3Y7, QC, Canada
Sadat.fatiha@uqam.ca, emohamed@umail.iu.edu

RÉSUMÉ

(Traduction automatique statistique pour l'arabe-français améliorée par le prétraitement et l'analyse de la langue)

Dans cet article, nous nous intéressons au prétraitement de la langue arabe comme langue source à des fins de traduction automatique statistique. Nous présentons une étude sur la traduction automatique statistique basée sur les syntagmes, pour la paire de langues arabe-français utilisant le décodeur Moses ainsi que d'autres outils de base.

Les propriétés morphologiques et syntaxiques de la langue arabe sont complexes, ce qui rend cette langue difficile à maîtriser dans le domaine du TALN. Aussi, les performances d'un système de traduction statistique dépendent considérablement de la quantité et de la qualité des corpus d'apprentissage. Dans cette étude, nous montrons qu'un prétraitement basé sur les mots de la langue source (arabe) et l'introduction de quelques règles linguistiques par rapport à la syntaxe de la langue cible (français), permet d'obtenir des améliorations du score BLEU. Cette amélioration est réalisée sans augmenter la quantité des corpus d'apprentissage.

ABSTRACT

Arabic is a morphologically rich and complex language, which presents significant challenges for natural language processing and machine translation. In this paper, we describe an ongoing effort to build a competitive Arabic-French phrase–based machine translation system using the Moses decoder and other tools.

The results show an increase in terms of BLEU score after introducing some pre-processing schemes for Arabic and applying additional language analysis rules in relation to the target language. The proposed approach is completed using pre-processing and language analysis rules without increasing the amount of training data.

MOTS-CLÉS : Traduction automatique statistique, traduction arabe-français, pré-traitement de corpus, morphologie de l'Arabe.
KEYWORDS : Statistical machine translation, Arabic-French translation, Corpus pre-processing, Arabic morphology.

1 Introduction

Arabic is a morphologically rich and complex language, in which a word carries not only

inflections but also clitics, such as pronouns, conjunctions, and prepositions. This morphological complexity also has consequences for NLP applications, such as machine translation and information retrieval. On the one hand, developing an Arabic-French machine translation system is not an easy task, although there is a vast amount of training data nowadays. On the other hand, dealing with the complexity and ambiguity of the source language plays a major role in boosting the efficiency of the translation system.

In previous research, it was shown that morphological pre-processing of a morphologically rich language, such as Arabic does provide a benefit, especially in the case of limited volume of training data (Goldwater and McClosky, 2005 ; Sadat and Habash, 2006 ; Lee, 2004 ; El Ishibani et al., 2006 ; Hasan et al., 2003).

In Statistical Machine Translation (SMT) context, Habash et Sadat (2006) pre-processed Arabic texts using different segmentation schemes for translation into English and showed that the quality of translation is generally better than the baseline. Similar findings were reported by El Ishibani et al. (2006) on Arabic-English SMT.

In relation to Arabic-French SMT, few research and evaluations were reported, compared to Arabic-English SMT among other pairs of languages. One of the first statistically-driven machine translation systems for Arabic-French was reported by Hasan et al (Hasan et al., 2006) during the second Cesta evaluation campaign[1]. The proposed SMT system used a simple stemming algorithm based on finite-state automata to split Arabic words into prefixes, stem and suffixes. Nevertheless, this simple segmentation method showed a reduced OOV rate from 8.2% to 2.6% for the test data and thus a better quality of translation in terms of BLEU score (Papineni et al., 2001). Another research on Arabic-French SMT was focused on domain adaptation to the news domain and did not consider the pre-processing of the morphologically complex language such as Arabic (Schwenk and Senellart, 2009). An improvement of 3.5 BLEU points on the test set was realized.

In relation to improving an SMT system using some language analysis rules, such as re-ordering with Arabic as a source language, there was no reported research on Arabic-French SMT. However, Carpuat et al. (Carpuat et al., 2010) showed that post-verbal subject (VS) constructions are hard to translate because they have highly ambiguous reordering patterns when translated to English. They proposed to reorder VS construction into SV order for SMT word alignment only. This strategy significantly improves BLEU and TER scores of the SMT using Arabic and English language pair.

In this paper, we report some experiments related to our first participation in the 2012 TRAD evaluation campaign[2], that was coordinated by the *Laboratoire National de métrologie et d'Essais (LNE)* and CASSIDIAN (*the defence and security subsidiary of the EADS group*), and was funded by the French General Directorate for Armament (DGA). Our main interest at this stage is related to the pre-processing of the source language, in order to improve the quality of translation, rather than the radical changes that might improve the translation or training engines or the increase of the amount of training corpora.

This paper is organized as follows. The morphology of Arabic language is described in section 2. In section 3, we discuss the proposed solutions of pre-processing Arabic through

[1] http://www.technolangue.net/article.php3?id_article=199
[2] http://www.trad-campaign.org/

segmentation and language analysis. In section 4, we present the experiments on Arabic-French SMT with different evaluations. Section 5 concludes the present paper with a discussion and some perspectives.

2 The Morphology of Arabic Language

Before we delve into the methods, we need to discuss the nature of the Arabic language, which has a bearing on the text preparation stage. Figure 1 shows a white-space delimited word in Arabic.

The Arabic script is complicated in that each white-space-delimited unit may correspond to several syntactic units. The Arabic orthographic unit, a unit delimited by white space, usually carries more than one token. An example is a form like (*wsyktbwnhA*)[3] (Eng. and they will write it, depicted in Figure 1.). This grammatically complete sentence carries a conjunction w, a future particle s, a verbal token *yktbwn*, and a feminine singular third person object pronoun *hA*. The verbal token is made of a verb *ktb*, a masculine present third person inflection *y* and a plural indicative inflection *wn*. This nature entails that the type token ratio is much smaller than it is for a non-morphologically rich language like English for example. This means that the same word does not repeat often enough for the investigator to make valid observations. In order for any linguistic, especially lexical, investigation to be reliable, one needs to perform some sort of morphological analysis capable of reducing the word to its basic form. This has implications on Machine translation as it means that no matter how big the training corpus is; the Arabic side will always suffer from scarcity.

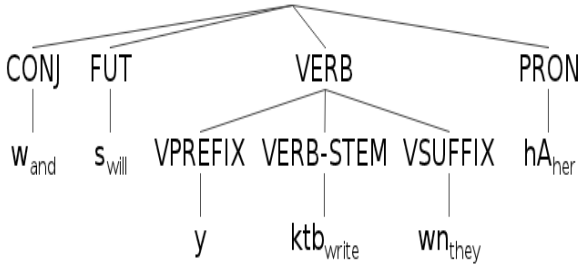

FIGURE 1 – The Morphology of an Arabic word

3 Pre-processing Arabic for SMT

With Arabic being morphologically complex and rich, lexical scarcity comes as a natural result. In such cases it helps to reduce this morphological complexity in order to obtain better alignments and decoding for Statistical Machine Translation (Habash et al., 2010).

Our goal at this stage is related to the pre-processing of Arabic as a source language, in order to improve the quality of translation. First, in order to perform Arabic pre-processing, we used a machine learning approach that performs word segmentation and POS tagging at the segment level. We then use rules to derive the different pre-processing schemes

[3] All Arabic transliterations are provided using the Buckwalter transliteration scheme (Buckwalter, 2002)

required for the machine translation experiments. Thus, instead of using MADA (Habash et al., 2010), the well known morphological analyzer for Arabic, we choose our own morphological analyzer that is memory-based learning for both word segmentation and part of speech tagging (Emad and Kübler, 2010).

The segmentation and POS tagging modules above give a rich representation with enough information for almost any further required transformation. Given an input sentence like (a), the system produces (b) as a segmented and annotated sentence, as described in the following example:

(a) وقد ارتبطت الاضطرابات بترحيل السلطات الفرنسية للعديد من المهاجرين غير الشرعيين

(In Buckwalter transliteration): *wqd ArtbTt AlADTrAbAt btrHyl AlslTAt Alfrnsyp llEdyd mn AlmhAjryn gyr Al$rEyyn*

(b) w/**CONJ**+qd/**VERB_PART** ArtbT/**PV**+t/**PVSUFF_SUBJ:3FS**

Al/**DET**+ADTrAb/**NOUN**+At/**NSUFF_FEM_PL** b/**PREP**+trHyl/**NOUN**
Al/**DET**+slT/**NOUN**+At/**NSUFF_FEM_PL** Al/**DET**+frnsy/**ADJ**+p/**NSUFF_FEM_SG**
l/**PREP**+l/**DET**+Edyd/**NOUN** mn/**PREP**
Al/**DET**+mhAjr/**NOUN**+yn/**NSUFF_MASC_PL_GEN** gyr/**NEG_PART**
Al/**DET**+$rEy/**ADJ**+yn/ **NSUFF_MASC_PL_GEN**

We set four different evaluations based on the variations on the output of the above example, as follows:

Basic. The Basic experiment is the baseline of all the work we are doing. In this experiment, the Arabic side undergoes minimal pre-processing in which we only separate the punctuation and remove the occasional diacritization (the short vowels). Short vowels do not normally occur in Arabic, but sometimes scattered ones are there mainly for disambiguation purposes; however since their use is not standardized and subjective, their removal usually leads to better agreement between the training and test sets.

Tokenized. In this context, tokenization means splitting the prefixes and suffixes that have a syntactic value and that usually stand as independent words in other languages. Examples of these include the possessive pronouns (-hm, -h, -y, -hA), conjunctions (w, f), and prepositions (l-, k-, t-). We have also chosen to split the Arabic definite article ***Al*** due to the perceived similarity in distribution between the Arabic and French definite articles.

The sentence above "wqd ArtbTt AlADTrAbAt btrHyl AlslTAt Alfrnsyp llEdyd mn AlmhAjryn gyr Al$rEyyn "

is thus tokenized as "**w/CONJ** qd/VERB_PART ArtbT/PV+t/PVSUFF_SUBJ:3FS Al/DET ADTrAb/NOUN+At/NSUFF_FEM_PL **b/PREP** trHyl/NOUN Al/DET slT/NOUN+At/NSUFF_FEM_PL Al/DET frnsy/ADJ+p/NSUFF_FEM_SG **l/PREP Al/DET** Edyd/NOUN mn/PREP Al/DET mhAjr/NOUN+yn/NSUFF_MASC_PL_GEN gyr/NEG_PART Al/DET $rEy/ADJ+yn/ NSUFF_MASC_PL_GEN".

Where the conjunction w, the prepositions b and l, and the definite article Al are no longer prefixes, but separate tokens. The process also normalized the definite article from ***l*** to ***Al***, which is the more frequent form.

MorpReduced. In the morphologically reduced experiment, we reduce the morphology of

Arabic to a level that makes it closer to that of the French language. An example of this is the dual form, which does not occur in French and has thus been transformed to the plural. The following table (Table 1) lists the most common examples of Arabic morphological reduction.

Rule	Example before applying the rule	Example after applying the rule
Regular Plural Nominative → Regular Plural Accusative	mstwTn***wn***	AlmstwTn**yn**
dual Nominative → Regular Plural Accusative	lAEb***An***	lAEb***yn***
Jussive Mood → Indicative Mood	hn lm ylEb***n*** hm lm ylEb***wA*** hmA lm ylEb**A**	hm lm ylEb***wn*** hn lm ylEb***wn*** hm lm ylEb***wn***

TABLE 1 – The most common rules for Arabic morphological reduction

Swapped. The swapped experiment tries to introduce some structural matching between the source language (Arabic) and the target language (French). Two structural changes have been attempted, as follows:

(a) While Arabic possessive pronouns follow the nouns, we have made them precede the nouns in order to match the French. For example ktAb -y (book -my) has now become (-y book) to match "mon livre" (in French).

(b) Arabic object pronouns, which follow the verb, have been made to precede it. *>nA >ryd h* (I want it) is now *>nA h >ryd* with the purpose of matching the French structure "*Je le veux*".

4 Experiments on SMT

Our SMT system was trained on 3.5 million words of French and their parallel text in Arabic (equivalent to 108 300 sentences) in addition to 9700 parallel sentences that were extracted from the essentially comparable UN corpus of 2009. Thus, the total number of sentences is 118 000 for the training corpora. The development corpus contains 20,000 words, namely 40,000 words with the reference. The evaluation corpus contains 15,000 words with 4 references.

The common practice of extracting bilingual phrases from the parallel data usually consists of three steps: first, words in bilingual sentence pairs are aligned using state-of-the-art automatic word alignment tools, such as GIZA++ (Och and Ney, 2003), in both directions; second, word alignment links are refined using heuristics, such as Grow-Diagonal-Final (GDF) method; third, bilingual phrases are extracted from the parallel data based on the refined word alignments with predefined constraints (Och and Ney, 2003).

The trigram language models are implemented using the SRILM toolkit (Stolcke, 2002).

Moses[4] (Koehn et al., 2007), an open source toolkit for phrase-based SMT system, was used as a decoder.

These steps of building a translation system are considered as a common practice in the state-of-the-art of phrase-based SMT systems. Our research for improving the Arabic-French SMT system was emphasized more on the pre-processing part of the SMT system.

We have measured the effect of the proposed pre-processing steps on data sparseness, based on the percentage of unknown unigrams (OOVs) on a development set (dev set). Table 2 summarizes the findings on the dev set. We give numbers in terms of tokens (the total number of words) and types (the number of unique words in the text, i.e. no-redundant words in the text).

Experiment	% OOV (Types)	% OOV (Tokens)	BLEU score
Baseline	10.74	4.81	17.69
Tokenized	7.99	2.00	25.84
MorphReduced	7.87	1.98	26.33
Swapped	7.87	1.98	25.48

TABLE 2 – Effect of pre-processing on the development set

It can be noticed that the tokenization has a major effect on combatting data sparseness and consequently improving the quality of translation as measured by the BLEU score. Morphological normalization, which is a layer on top of tokenization, improves things even further, and this is reflected in the difference between the baseline BLEU score and the MorphReduced BLUE score which is 8.6 absolute points.
The swapped experiment leads the system output to deteriorate; which leads to a review of the introduced rules for the structural matching between the source Arabic and the target French languages, in the future.
Table 3 compares the results, in term of BLEU scores, of the 4 experimental settings in 3 evaluations schemes, as follows:

(a) ***Standard***, which includes performing re-casing and removing white space before punctuation,

(b) ***Nopunct***, in which punctuation is stripped and evaluation is performed on the lexical text only, and

(c) ***Nopunctcase*** in which, in addition to removing punctuation, all words are lower-cased.

We can see from Table 3 that the Baseline experiment produces the lowest results, and that the tokenization scheme is a big leap with a 7.2 BLEU scores of improvement (25.9 vs. 33.1), which means that performing tokenization is a really a necessary step for translating

[4] Available on http://www.statmt.org/ moses/

from Arabic, an that the morphological complexity of Arabic could be a hindrance to quality automatic translation. While tokenization leads to considerable improvement, morphological reduction fares even better with a 7.4 BLEU score higher than the baseline. This could be due to the fact the morphological reduction reduces the number of unknown words even further than tokenization alone. Swapping elements to match the target language, which is built upon tokenization and morphological reduction, leads to a deterioration of the results a little as it cancels out the effect of the morphological reduction process. It is still an open question whether the positive effect of pre-processing will still carry over with increasing the amount of training data and to what extent this will help.

	Base	Tokenized	MorphReduced	Swapped
Standard	25.9	33.1	**33.3**	33.1
Nopunct	23.8	31.5	**31.7**	31.4
Nopunctcase	25.8	**34.1**	**34.1**	34

TABLE 3 – Results in terms of BLEU score

5 Conclusion

We have presented an ongoing project on developing a competitive Arabic to French machine translation, using the methods and data of the TRAD 2102 evaluation campaign.

We have introduced pre-processing schemes for the source language (Arabic) and some rules of language analysis related to the target language (French). Our method for POS tagging and segmentation of Arabic texts showed a significant improvement in terms of BLEU score; however it does not assume the best results. The introduced morphological rule that reduces the morphology of Arabic to a level that makes it closer to that of the French language, showed the best results. We have introduces extra swapping rules, that tries to introduce some structural matching between the source language (Arabic) and the target language (French); however there was no improvement in terms of BLEU score. Our future work is focused on the revision of these swapping rules and the introduction of more rule for the recognition and transliteration of named entities; which makes our translation system a hybrid rule-based and statistical SMT system. We will also investigate the integration of more training data such as comparable corpora to make our SMT system more competitive and reliable.

Références

BUCKWALTER, T. (2002). Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania. Catalog: LDC2002L49.

CARPUAT, M., MARTON, Y. et HABASH, N. (2010). Reordering Matrix Post-verbal Subjects for Arabic-to-English SMT. In proceedings of the 17th Conference sur le Traitement des Langues Naturelles (TALN 2010). Montreal, Canada.

DIAB, M., HACIOGLU, K. et JURAFSKY, D. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In *Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Boston, MA.

EMAD, M. et KÜBLER, S. (2010). Is Arabic Part of Speech Tagging Feasible Without Word Segmentation? In HLT/ACL 2010, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 705–708, Los Angeles, California, June 2010.

EL ISBIHANI, A., KHADIVI, S., BENDER, O., ET NEY, H. (2006). Morpho-syntactic Arabic Preprocessing for Arabic to English Statistical Machine Translation. In Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), Workshop on Statistical Machine Translation, New York City, pages 15-22.

GOLDWATER, S. et MCCLOSKY, D. (2005). Improving Statistical MT through Morphological Analysis. In *Proc. of Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver, Canada.

HABASH, N. et SADAT, F. (2006). *Arabic Preprocessing Schemes for Statistical Machine Translation*. In Proceedings of NAACL 2006, New York (USA). June 5-7.

HABASH, N., RAMBOW, O. et RYAN R. (2010). *The MADA and TOKAN Manual*.

HASAN, S., EL ISBIHANI, A. et NEY, H. (2006). Creating a Large-Scale Arabic to French Statistical Machine Translation System. In International Conference on Language resources and Evaluation (LREC), Genoa, Italy, pages 855-858.

KOEHN, P., SHEN, W., FEDERICO, M,. BERTOLDI, N., CALLISON-BURCH, C., COWAN, B., DYER, C., HOANG, H., BOJAR, O. ZENS, R., CONSTANTIN, A., HERBST, E., MORAN C. et BIRCH, A. (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of ACL 2007.

LEE, Y. (2004). Morphological Analysis for Statistical Machine Translation. In *Proc. of NAACL*, Boston, MA.

OCH, F., J. et NEY,H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational linguistics 29 (1), pages 19-51.

PAPINENI, K., ROUKOS, S., WARD, T. et ZHU, W. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176(W0109-022), IBM Research Division, Yorktown Heights, NY.

SADAT, F. et HABASH, H. *Arabic Preprocessing for Statistical Machine Translation: Schemes and Techniques*. In Proceedings of COLING-ACL 2006, Sydney, Australia. July 17-21 (2006).

SCHWENK, H. et SENELLART, J. (2009). Translation model adaptation for an Arabic/French news translation system by lightly-supervised training. In *MT Summit*.

STOLCKE, A. (2002). SRILM-An Extensible Language Modeling Toolkit. *In Proc. Of the International Conference on Spoken language Processing.*