Linguistic Factors in Statistical Machine Translation Involving Arabic Language

Islam Youssef[*], Mohamed Sakr[**] and Mohamed Kouta[***]

[*]Arab Academy for Science, Technology & Maritime Transport, Cairo, Egypt
[**]Al Shorook Academy High Institute for Computers and Information Systems, Cairo, Egypt
[***]Arab Academy for Science, Technology & Maritime Transport, Cairo, Egypt

Summary
Arabic is considered to have a rich morphology compared to English language. This fact adversely affects the performance of English-Arabic Statistical Machine Translation (SMT). Phrase-based SMT models have a limitation of mapping phrases or blocks from the source to the target languages without any use of linguistic information.
Incorporating linguistic tools, such as part-of-speech (POS) taggers can have an impact on translation quality.
In this paper, the use of POS tagging is incorporated as a linguistic feature in a factored translation model. The use of factored translation model and its impact on translation quality for English-Arabic machine translation is reported.

Key words:
Statistical Machine Translation, Phrase Based Model, Part of Speech Tagging, Factored Model, Decoding Algorithm.

1. Introduction

Phrase based statistical machine translation (PBSMT) systems are currently considered to be the state-of-the-art in SMT. They achieve top performing results according to the National Institute of Standards and Technology (NIST) [1]. Known limitations of PBSMT include bad performance when translating to morphologically rich languages as opposed to translating from them [2].
Arabic is considered to be a rich language with reference to inflection and derivation when compared to English.
Words are inflected for gender, number, and sometimes grammatical case, over more various clitics can be attached to word stems. An Arabic corpus will contain more surface forms than an English corpus of the same size, and will also be more sparsely populated. This morphological richness makes statistical machine translation from English to Arabic a challenging task.

In this paper, the use of factored translation model from English to Arabic is reported. First a brief overview of Phrase and Factored SMT models are presented, then a brief discussion about the proposed translation approach used, the data used in the experiments and finally the initial set of experiments and results showing improvements of translation quality.

1.1. Related Work

Niessen and Ney [3] make use of morphological information to improve word reordering before training and after decoding. Ueffing and Ney [4] used POS tags, in order to deal with the verb conjugation of Spanish and Catalan; POS tags were used to identify the pronoun + verb sequence and splice these two words into one term. Goldwater and McClosky [5] showed improvements in Czech to English word-based translation system when inflectional endings are simplified or removed entirely. Minkov [6] suggested a post-processing system which uses morphological and syntactic features, in order to ensure grammatical agreement for the output.
Koehn and Hoang [7] reported on experiments that showed incorporating factored translation models gains over standard phrase-based models, both in terms of automatic scores, as well as a measure of grammatical coherence. Avramidis and Koehn [8] have shown how SMT performance can be improved, when translating from English into morphologically richer languages, by adding linguistic information on the source, although the source language misses morphology attributes required by the target language, the needed information is inherent in the syntactic structure of the source sentence.
In Larkey [9] it was already shown that word segmentation for Arabic improves information retrieval. In Lee [10] a statistical approach for Arabic word segmentation was presented. It decomposes each word into a sequence of morphemes (prefixes-stem-suffixes), where all possible prefixes and suffixes are split from the original word. Diab [11] discussed a POS tagging method for Arabic.

Manuscript received November 5, 2009
Manuscript revised November 20, 2009

2. Overview of Factored SMT

Factored translation models are closely similar to phrase-based models, the main difference lies in the new linguistic factors and the training models gained from those new factors.

2.1 Phrase Based Model

The phrase translation model used here was formerly defined by Koehn [12]. The model is based on noisy channel that using Bayes rule to reformulate the translation probability. Thus translating an English sentence **f** into an Arabic **e** will be modeled as

$$\text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f}) = \text{argmax}_{\mathbf{e}}\, p(\mathbf{f}|\mathbf{e})\, p(\mathbf{e})$$

This allows for a language model probability $p(\mathbf{e})$ and a separate translation model probability $p(\mathbf{f}|\mathbf{e})$. During decoding, the input English sentence **f** is segmented into a sequences of consecutive words which are called phrases or chunks of I phrases f_1^I as in Figure 1. Phrases are segmented with a uniform probability distribution over all possible segmentations. Each English phrase f_i in f_1^I is translated into an Arabic phrase e_i. The Arabic phrases may then be reordered. Phrase translation is modeled by a probability distribution $\Phi\ (f_i|e_i)$.

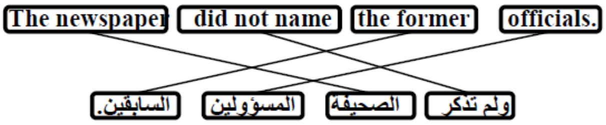

Figure 1: Example of phrase-aligned translation

Reordering of the Arabic output phrases is modeled by a relative distortion probability distribution $d(a_i - b_{i-1})$, where a_i denotes the start position of the English phrase that was translated into the i^{th} Arabic phrase, and b_{i-1} denotes the end position of the English phrase that was translated into the $(i-1)^{th}$ Arabic phrase.

A simple distortion model $d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1} - 1|}$ with an appropriate value for the parameter α is used.

In order to calibrate the output length, a new factor ω is introduced which is called the word cost for each generated English word in addition to the trigram language model p_{LM}.

The best Arabic output sentence $\mathbf{e}_{best}$ given an English input sentence **f** according to the model explained by [12] will be

$$\mathbf{e}_{best} = \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f})$$
$$= \text{argmax}_{\mathbf{e}}\, p(\mathbf{f}|\mathbf{e})\, p_{LM}(\mathbf{e})\, \omega^{\text{length}(\mathbf{e})}$$

where $p(\mathbf{f}|\mathbf{e})$ is decomposed into

$$p(f_1^I \mid e_1^I) = \prod_{i=1}^{I} \Phi(f_i \mid e_i)\, d(a_i - b_{i-1})$$

2.2 Word Alignment

Extracting a phrase translation table from a parallel corpus starts with word alignment. GIZA++ toolkit [13] is used to align words. First, the parallel corpus is aligned in a bidirectional way, Arabic to English and English to Arabic. This generates two word alignments that have to be reconciled. If we intersect the two alignments, we get a high-precision alignment of high-confidence alignment points. If we take the union of the two alignments, we get a high-recall alignment with additional alignment points as shown in Figure 2.

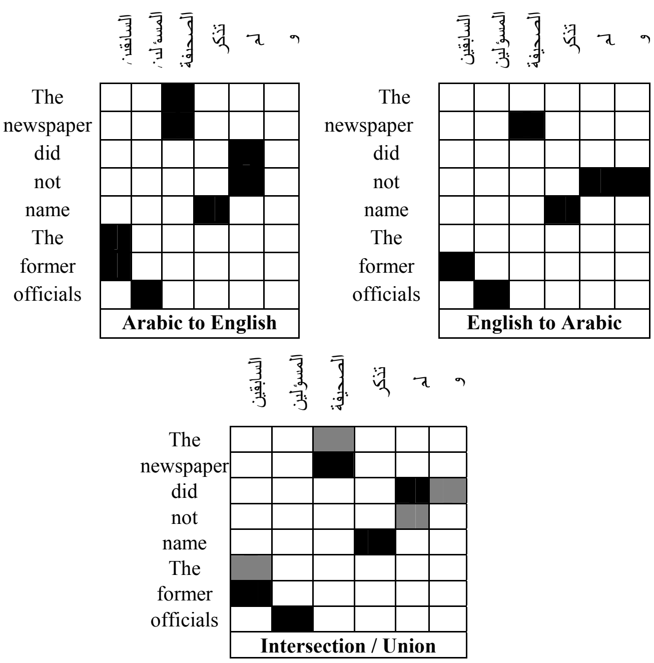

Figure 2: Example of Word Alignment

2.3 Decoder

The phrase-based decoder implements a beam search algorithm; the Arabic output sentence is generated left to right in the form of partial translations (or hypotheses). The decoder starts with an empty hypothesis. A new hypothesis is expanded by selecting a sequence of untranslated English words and a possible Arabic phrase translation for them. The Arabic phrase is attached to the existing Arabic output sequence. The English words are marked as translated and the probability cost of the hypothesis is updated. The highest probability final hypothesis with untranslated English words is the output of the search. The estimated phrase translation cost is calculated by multiplying its phrase translation probability with the language model probability for the generated Arabic phrase. Given the costs for the translation options, the estimated future cost can be computed for any sequence of consecutive foreign words by dynamic programming.

The beam size, e.g. the maximum number of hypotheses, is fixed to a certain number. The number of translation options is linear with the sentence length. And, the time complexity of the beam search is quadratic with sentence length, and linear with the beam size.

2.4 Factored Based Model

The factored models mainly add additional annotation at each word level. A word is not anymore only a token, but a vector of factors that represent different levels of annotation. It uses a log-linear approach, in order to combine the several components, including the language model, the reordering model, the translation models and the generation models. The model is defined mathematically by Koehn and Hoang [7] as follow

$$p(\mathrm{e}\,|\,\mathrm{f}) = \exp\sum_{i=1}^{n} \lambda_i h_i(e, f)$$

To compute the probability of a translation **e** given an input sentence **f**, an evaluation of each feature function h_i has to be performed and then multiplied by its feature weight λ_i. For instance, the feature function for a bigram language model component is

$$h_{\mathrm{lm}}(\mathbf{e}, \mathbf{f}) = p_{\mathrm{lm}}(\mathbf{e})$$

$$= p(\mathrm{e}_1)p(\mathrm{e}_2|\mathrm{e}_1)...p(\mathrm{e}_m|\mathrm{e}_{m-1})$$

where m is the number of words e_i in the sentence **e**

Considering the feature functions introduced by the translation and generation steps of factored translation models, the translation of input sentence **f** into the output sentence **e** breaks down to a set of phrase translations (f_j,e_j).

For a translation step component, each feature function h_t is defined over the phrase pairs (f_j,e_j) given a scoring function τ:

$$h_t(\mathrm{e}\,|\,\mathrm{f}) = \sum_j \tau(f_j, e_j)$$

For a generation step component, each feature function h_g given a scoring function γ is defined over the output words e_k:

$$h_g(\mathrm{e}\,|\,\mathrm{f}) = \sum_k \gamma(e_k)$$

The feature functions follow from the scoring functions (τ, γ) acquired during the training of translation and generation tables.

The feature weights λ_i in the log-linear model are determined with a minimum error rate training method.

3. Translation Approach

The approach followed in this paper incorporates the POS tagging as a linguistic factor in the Factored based translation model previously described. The data used, the translation process followed and some experiments are explained to show the effect of the added factor on the translation quality against a non factored system.

3.1 Data Used

Experiments were carried out on the Arabic English Parallel News Text Part 1, obtained from the Linguistic Data Consortium (LDC) catalog number LDC2004T18 and ISBN 1-58563-310-0. The corpus contains Arabic news stories and their English translations LDC collected via Ummah Press Service from January 2001 to September 2004. It totals 8,439 story pairs, 68,685 sentence pairs, 2 Million Arabic words and 2.5 Million English words. The corpus is aligned at the sentence level.

Another used corpus is the Corpora of the United Nations for the research purposes. The corpus is a paragraph-aligned six-language collection of resolutions of the General Assembly from Volume I of GA regular sessions 55-62. The corpus is described in [14] as a six-ways parallel public-domain corpus consisting of 2100 United Nations General Assembly Resolutions with translations in the six official languages of the United Nations, with an average of around 3 million tokens per language. The corpus is available in a preprocessed, formatting-normalized TMX format with paragraphs aligned across multiple languages.

3.2 Translation Process

First Training data was provided sentence aligned (one sentence per line), in two files, one for Arabic sentences, one for the English sentences. Then a cleanup process was conducted to remove empty lines, remove redundant space characters, drop lines (and their corresponding lines), that are empty, too short, or too long. A maximum phrase length of 40 words was used. The corpus was then tokenized and lowercased. The English and Arabic corpus are tagged with the Stanford Log-linear POS Tagger described by [15] and [16]. The tagger output uses the Penn Treebank tag set to identify each word POS. Table 1 & 2 represent output of the POS tagging process.

word	الإسلامية	إيران	جمهورية	في	الإنسان	حقوق	حالة
POS	DTJJ	NNP	NN	IN	DTNN	NN	NN

Table 1: Example of generated POS tagging for Arabic

word	Situation	of	human	rights	in	the	Islamic	republic	of	Iran
POS	NN	IN	JJ	NNS	IN	DT	NNP	NNP	IN	NNP

Table 2: Example of generated POS tagging for English

Table 3 below represents sample of POS tag meanings.

JJ	Adjective.	NNP	Proper Noun, singular.
NN	Noun, singular or mass.	IN	Preposition or conjunction, subordinating.
NNS	Noun, plural	DT	Determiner.

Table 3: Sample POS tags

In order to build a language model the SRILM toolkit [17] was used to generate a two tri-gram target Arabic language models containing the surface form and POS form.

The English source is aligned to the Arabic target using GIZA++, which generates vocabulary files and convert the parallel corpus into a numeric format. A training process then occurs generating the phrase alignments which start by word alignments taken from the intersection of bidirectional runs of GIZA++ with some additional alignment features from the union of the two runs.

The maximum likelihood lexical translation table is then estimated from the stored phrase translation pairs. The search decoding is done using the PBSMT system MOSES [18].

The translation process can be summarized by Table 4.

1	**Corpus Cleanup**
2	**Tokenization**
3	**Lowercasing**
4	**Tagging**
5	**POS Language Model & Surface Language Model Generation**
6	**Alignment & Translation Model Generation**
7	**Search Decoding**

Table 4 Translation process

4. Translation Evaluation

In order to evaluate translation performance BLEU (**bil**ingual **e**valuation **u**nderstudy) scoring tool proposed by Papineni [19] is used. It is based on the notion of modified *n*-gram precision, for which all candidate *n*-gram counts in the translation are collected and clipped against their corresponding maximum reference counts. These clipped candidate counts are summed and normalized by the total number of candidate *n*-grams. A sentence-aligned reference file and system generated output files wrapped into SGML format are used for evaluation as shown in Figure 3.

Another metric for measuring performance will be the NIST implementation of BLEU, with a different calculation of the brevity penalty as described in [20].

```
<refset setid="un-test" srclang="any" trglang="ar">
<DOC docid="un-test" sysid="ref">
…
<seg=767> وتصميما منها على تعزيز الاحترام الصارم للمقاصد والمبادئ
المكرسة في ميثاق الأمم المتحدة ،</seg>
…
</DOC>
</refset>
```

Figure 3: Arabic generated translation sample in SGML format

5. Experimental Results

In our results we compared the evaluation of translation quality obtained from the baseline system, which contains only the surface form of the words, with the morphologically extended system by the POS model. An input source English set has been prepared with its reference Arabic translation set; the two sets were composed of 2000 sentences (sentence level aligned) for evaluation.

The testing has been conducted by translating the input source English set on both MT systems. The output of the two systems was wrapped in SGML format for evaluation. The evaluation was carried by the BLEU and NIST scoring using N-gram co-occurrence scoring utility.

Model	BLEU	NIST
Baseline Surface Form	0.6095	9.9103
Surface Form + POS	0.6394	10.3957

Table 5: Translation Results

For the baseline system, we observed an average of 0.6095 score for BLEU and 9.9103 for NIST. Table 5 indicates that the baseline system resulting scores are enhanced in the system which was trained on both surface model and POS model which revealed scores of 0.6394, 10.3957 for BLEU and NIST evaluation respectively.

Figure 4 & 5 reveal the actual output of the MT evaluation script for different N-Gram chunks.

```
src set “test” (1 docs, 2000 segs)
NIST score = 9.9103  BLEU score = 0.6095 for system "Surface"
Individual N-gram scoring
        1-gram 2-gram 3-gram 4-gram 5-gram 6-gram 7-gram 8-gram 9-gram
        ------------------------------------------------------------
NIST:  7.2320 1.7391 0.5724 0.2443 0.1225  0.0661 0.0480 0.0373 0.0290 "Surface"
BLEU: 0.7767 0.6465 0.5592 0.4915 0.4355 0.3918 0.3553 0.3243 0.2969 "Surface"

Cumulative N-gram scoring
    1-gram 2-gram 3-gram 4-gram 5-gram 6-gram 7-gram 8-gram 9-gram
    ------------------------------------------------------------
NIST:  7.2320 8.9710 9.5435 9.7878 9.9103 9.9764 10.0243 10.0616 10.0907 "Surface"
BLEU: 0.7767 0.7086 0.6548 0.6095 0.5699 0.5354 0.5049  0.4777  0.4531  "Surface"
```

Figure 4: Evaluation Scores for Surface model

```
src set “test” (1 docs, 2000 segs)
NIST score = 10.3957  BLEU score = 0.6394 for system "POS"
Individual N-gram scoring
        1-gram 2-gram 3-gram 4-gram 5-gram 6-gram 7-gram 8-gram 9-gram
        ------------------------------------------------------------
NIST: 7.5491 1.8385 0.6120 0.2637 0.1325 0.0742 0.0532 0.0414 0.0331 "POS"
BLEU: 0.8092 0.6826 0.5950 0.5253 0.4670 0.4219 0.3842 0.3520 0.3250 "POS"

Cumulative N-gram scoring
        1-gram  2-gram  3-gram 4-gram  5-gram  6-gram 7-gram  8-gram 9-gram
        --------------------------------------------------------------
NIST: 7.5491 9.3876 9.9996 10.2633 10.3957 10.4700 10.5232 10.5646 10.5977 "POS"
BLEU: 0.8027 0.7372 0.6845 0.6394  0.5995  0.5646 0.5338  0.5062  0.4815 "POS"
```

Figure 5: Evaluation Scores for Surface & POS Model

The results shows that enriching the corpus with morphological factors especially for rich morphology language like Arabic have a beneficial impact on translation quality.

6. Conclusion

In this paper, it has been showed that the use of factored model for phrase-based English-Arabic machine translation tends to improve the morphological coherence of MT output. Our results on English-Arabic translation with the use of POS as a morphological feature showed to be beneficial with reference to translation quality.

The system with POS factor improved the translation quality with 0.0299 BLEU scores over the standard surface based system.

It will be interesting to see how other morphological features such as lemma, morphology and word class can be handled, in order to improve quality of translations into languages with a highly inflected morphology.

References

[1] The NIST machine translation scoring and tests is available at URL: http://www.nist.gov/speech/tests/mt

[2] P. Koehn, “Europarl: A Parallel Corpus for Statistical Machine Translation”, In Proceedings of MT Summit X, 2005.

[3] S. Niessen and H. Ney, “Morpho-syntactic analysis for reordering in statistical machine translation”, In Proceedings of MT Summit VIII, Santiago de Compostela, Galicia, Spain, 2001.

[4] N. Ueffing, and H. Ney, “Using pos information for statistical machine translation into morphologically rich languages”, In EACL ’03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, pages 347–354, Morristown, NJ, USA. Association for Computational Linguistics, 2003.

[5] S. Goldwater and D. McClosky. “Improving statistical mt through morphological analysis”, In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 676–683, Vancouver, British Columbia, 2005.

[6] E. Minkov, K. Toutanova and H. Suzuki, “Generating complex morphology for machine translation”, In ACL 07: Proceedings of the 45th Annual Meeting of the Association of Computational linguistics, pages 128–135, Prague, Czech Republic. Association for Computational Linguistics, 2007.

[7] P. Koehn and H. Hoang “Factored Translation Models”, In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 868–876, Prague, June 2007.

[8] E. Avramidis and P. Koehn, “Enriching Morphologically Poor Languages for Statistical Machine Translation”. In Proceedings of ACL-08: HLT, pages 763–770, Columbus, Ohio, USA, June 2008.

[9] L. S. Larkey, L. Ballesteros, and M. E. Connell. “Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. In Proceeding of the 25th annual of the international Association for Computing Machinery Special Interest Group on Information Retrieval (ACM SIGIR), pages 275–282, New York, NY, USA. ACM Press, 2002.

[10] Y. S. Lee, K. Papineni, S. Roukos, O. Emam, and H. Hassan. “Language model based Arabic word segmentation”. In E. Hinrichs and D. Roth, editors, Proceeding of the 41st Annual Meeting of the Association for Computational Linguistic, 2003.

[11] M. Diab, K. Hacioglu, and D. Jurafsky. “Automatic tagging of Arabic text: From raw text to base phrase chunks”. In D. M. Susan Dumais and S. Roukos, editors, HLT-NAACL 2004: Short Papers, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics, 2004.

[12] P. Koehn, F. Och, and D. Marcu, “Statistical Phrase-Based Translation”, in Proceedings of the Joint Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Edmonton, Canada, pp.127–133, 2003.

[13] F. Och, H. Ney, "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19-5, 1 March 2003.

[14] A. Rafalovitch, R. Dale, “United Nations General Assembly Resolutions: A Six-Language Parallel Corpus”, In Proceedings of the MT Summit XII, pages 292-299, Ottawa, Canada, August 2009.

[15] K. Toutanova and C. Manning, “Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger”, In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70, 2000.

[16] K. Toutanova, D. Klein, C. Manning, and Y. Singer. “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network”. In Proceedings of HLT-NAACL 2003, pp. 252-259, 2003.

[17] A. Stolcke, “SRILM - An Extensible Language Modeling Toolkit”, in Proceedings of the 7th International Conference on Spoken Language Processing, Denver, CO, pp.901–904, 2002.

[18] MOSES, “A Factored Phrase-based Beam search Decoder for Machine Translation”. URL: http://www.statmt.org/moses/, 2007.

[19] K. Papineni, S. Roukos, T. Ward and W-J. Zhu, “BLEU: A Method for Automatic Evaluation of Machine Translation”, in Proceedings of ACL 2002, Philadelphia, PA., pp.311–318, 2002.

[20] NIST Open Machine Translation Evaluation Plan (MT09), *NIST Open MT Evaluation, 2009.* **URL:** http://www.nist.gov/speech/tests/mt/2009/