Integrating morpho-syntactic features in English-Arabic statistical machine translation

Ines Turki Khemakhem
MIRACL Laboratory,
ISIM Sfax,
pôle Technologique,
Route de Tunis Km 10, B.P.
242 Sfax 3021, Tunisie
Ines_turki@yahoo.fr

Salma Jamoussi
MIRACL Laboratory,
ISIM Sfax,
pôle Technologique,
Route de Tunis Km 10, B.P.
242 Sfax 3021, Tunisie
Salma.jammoussi@isimsf.rnu.tn

Abdelmajid Ben Hamadou
MIRACL Laboratory,
ISIM Sfax,
pôle Technologique,
Route de Tunis Km 10, B.P.
242 Sfax 3021, Tunisie
abdelmajid.benh-
-amadou@isimsf.rnu.tn

Abstract

This paper presents a hybrid approach to the enhancement of English to Arabic statistical machine translation quality. Machine Translation has been defined as the process that utilizes computer software to translate text from one natural language to another. Arabic, as a morphologically rich language, is a highly flexional language, in that the same root can lead to various forms according to its context. Statistical machine translation (SMT) engines often show poor syntax processing especially when the language used is morphologically rich such as Arabic. In this paper, to overcome these shortcomings, we describe our hybrid approach which integrates knowledge of the Arabic language into statistical machine translation. In this framework, we propose the use of a featured language model SFLM (Smaïli et al., 2004) to be able to integrate syntactic and grammatical knowledge about each word. In this paper, we first discuss some challenges in translating from English to Arabic and we explore various techniques to improve performance on this task. We apply a morphological segmentation step for Arabic words and we present our hybrid approach by identifying morpho-syntactic class of each segmented word to build up our statistical feature language model. We propose the scheme for recombining the segmented Arabic word, and describe their effect on translation.

1 Introduction

Arabic is characterized by complex morphology and rich vocabulary. It is a derivational, flexional language. In addition, Arabic is an agglutinative language. In fact, most Arabic words are made up by the concatenation of certain morphemes together. An Arabic corpus will therefore have more surface forms than an English corpus of the same size.

On the other hand, many Arabic words are homographic: they have the same orthographic form, but they have not the same meaning. This property can reduce the size of the translation vocabulary and has an important implication for statistical modeling of the Arabic language. These factors affect the performance of English-Arabic Statistical Machine Translation (SMT).

To overcome these weaknesses of SMT, we propose a hybrid approach that seeks to integrate the linguistic information and enrich the lexical and syntactic resources in the statistical machine translation.

Arabic language translation has been widely studied recently. Most of the time, the rich morphology of Arabic language is seen as a serious problem that must be resolved to build up an efficient translation system. It has been proven that pre-processing Arabic data and integrating its morpho-syntactic features is useful to improve machine translation results. The use of similar techniques for English-to-Arabic SMT requires recombination of the target side into valid surface forms, which is not a trivial task.

In this paper, we describe an initial set of experiments on English-to-Arabic machine translation: we apply a morphological segmentation step for Arabic words and we present our hybrid approach by identifying morpho-syntactic class of each segmented word to build up our statistical feature language model. We propose the scheme for recombining the segmented Arabic, and describe their effect on translation.

Proceedings of the Second Workshop on Hybrid Approaches to Translation, pages 74–81,
Sofia, Bulgaria, August 8, 2013.

This paper is organized as follows: section 2 gives a brief description of some related works using hybrid approach to Machine Translation to introduce morpho-syntactic features in a machine translation process. Section 3 describes the baseline system. Then, section 4 presents the used morphological analyzer MORPH2 for Arabic texts, able to recognize word composition and to provide more specific morphological information about it. Next, we give information about Arabic syntax and morphology in section 5; in the remainder of this section, we discuss the complexity of the Arabic morphology and the challenge of recombining the translated and segmented Arabic words in to their surface forms. The Statistical Feature Language Model (SFLM) is explained in section 6, when used it aims to integrate morpho-syntactic knowledge about word in the language model. We propose in section 7 a scheme for recombining the translated and segmented Arabic words in to their surface forms. Section 8 gives a short overview of the data and tools used to build up our SMT system and shows the experimental details of our system using SFLM and the morphological analyzer MORPH2. Section 9 discusses the obtained results and, finally, section 10 presents some conclusions.

2 Related work

Arabic language translation has been widely studied recently. Most of the time, the rich morphology of Arabic language is seen as a serious problem that must be resolved to build up an efficient translation system. Research into machine translation hybridization has increased over the last few years particulary with the statistical approach for machine translation. Habash et al. (Habash et al., 2006) boost generation-heavy machine translation (GHMT) with statistical machine translation components. They use hybridization approach from the opposite direction by incorporating SMT components into rule-based systems. In (Sawaf, 2010), authors described a novel approach on how to deal with Arabic noisy and dialectal data. They normalize the input text to a commun form to be able to process it.

In recent years, the overall quality of machine translation output has been improved greatly. Still, SMT engines often show poor results in their syntactic forms. Hybrid approach try to overcome these typical errors by integrating knowledge of Arabic language. It has been proven that pre-processing Arabic data and integrating its features such as morphological information and syntactic structure is useful to improve machine translation results.

In the next, we review this body of research. Our own research differs in that how to integrate information into SMT components systems.

Most of the related work is on Arabic-to-English SMT. In prior work (Lee, 2004) (Habash and Sadat, 2006), it has been shown that morphological segmentation of the Arabic source benefits the performance of Arabic-to-English SMT. In (Lee, 2004), the author uses a trigram language model to segment Arabic words. He then proceeds to deleting or merging some of the segmented morphemes in order to make the segmented Arabic source align better with the English target. Habash and Sadat (Habash and Sadat, 2006) compared the use of the BAMA (Buckwalter, 2002.) and MADA (Habash and Rambow, 2005) toolkits to segment the Arabic source as well as simple pattern matching to do morphological analysis for Arabic-English SMT, and were able to improve translation for tasks with out-of-domain training corpora. Sadat and Habash (Sadat and Habash. 2006) also showed that it was possible to combine the use of several variations of morphological analysis both while decoding (combining multiple phrase tables) and rescoring the combined outputs of distinct systems.

Introducing morphological analyzers in Arabic machine translation process is very present in the literature. The recent work (Besacier et al., 2008) conducted in depth a study of the influence of Arabic segmenters on the translation quality of an Arabic to English phrase-based system using the Moses decoder. In this work, authors demonstrate that the use of the morphology information in the SMT has a great impact in improving results. They believe that simultaneously using multiple segmentations is a promising way to improve machine translation of Arabic.

Arabic is an inflected language with several homonyms words, consequently linguistic features are very useful to reduce statistical machine translation errors due to this phenomena. Some research works have been conducted in this area (Bilmes and Kirchhoff, 2003) (Schwenk and Déchelotte, 2007). The factored language model (FLM) approach of Bilmes and Kirchhoff (Bilmes and Kirchhoff, 2003) is a more linguisti-

cally-informed modeling approach than the n-gram one. FLM are an extension of standard language model where the prediction is based upon a set of features (and not only on previous occurrences of the predicted word). FLM addresses the problems of data-sparsity in morphologically complex languages by representing words as bundles of features, thus one can easily capture dependencies between subword parts of adjacent words. Some other works have been proposed to integrate linguistic information such as part-of-speech, morphology and shallow syntax in conventional phrase-based statistical translation (Koehn and Hoang. 2007). These translation models allow integrating multiple levels of information into the translation process instead of incorporating linguistic markers in either preprocessing or postprocessing steps. For example, in morphologically rich languages it may be preferable to translate lemma, part-of-speech and morphological information separately and combine the information on the target side to generate the output surface words. In this model the translation process is broken up into three steps. Translate input lemmas into output lemmas in a first step. Then, translate morphological and POS factors in a second step. Finally, generate surface forms given the lemma and the linguistic factors. These factored translation models have been used to improve the word level translation accuracy by incorporating the factors in phrase-based translation. In (Schwenk and Déchelotte. 2007), authors focus on incorporating morpho-syntactic features in the translation model for the English-Spanish machine translation process. In this work, authors propose the use of augmented units in the translation model instead of simple words. These units are composed by surface word forms combined with their morpho-syntactic categories. This method allows lexical disambiguation of words using their roles and their grammatical contexts.

Previous works on English-to-Arabic SMT using factored models were proposed in (Sarikaya and Deng. 2007) and (Badr et al., 2008). The first uses shallow segmentation, and does not make use of contextual information. In this work authors use Joint Morphological-Lexical Language Models to rerank the output. The second work shows that morphological decomposition of the Arabic text is beneficial, especially for smaller-size corpora, and investigates different recombination techniques. In this work, authors propose the use of factored translation models for English to Arabic translation. The factors on the English side are POS tags and the surface word. On the Arabic side, they use the surface word, the stem and the POS tag concatenated to the segmented clitics.

In (Kholy and Habash. 2010), authors emphasized on the sparsity problem of English-Arabic translation. They considered the tokenization and normalization of Arabic data to improve English-to-Arabic SMT.

3 Phrase-Based Machine Translation

Statistical machine translation methods have evolved from using the simple word based models (Brown et al., 1993) to phrase based models (Marcu and Wong, 2002; Och and Ney. 2003).

The SMT has been formulated as a noisy channel model in which the target language sentence, s is seen as distorted by the channel into the foreign language t. In that, we try to find the sentence t which maximizes the $P(t|s)$ probability:

$$\text{argmax}_t P(t|s) = \text{argmax}_t P(s|t)P(t) \quad (1)$$

Where P(t) is the language model and P(s|t) is the translation model. We can get the language model from a monolingual corpus (in the target language). The translation model is obtained by using an aligned bilingual corpus.

The translation model is combined together with the following six additional feature models: the target language model, the word and the phrase bonus and the source-to-target and target-to-source lexicon model and the reordering model. These models are optimized by the decoder[1]. In our case, we use the open source Moses decoder described in (Koehn et al., 2007).

4 Segmentation for Arabic translation

Arabic is a morphologically complex language. Compared with English, an Arabic word can sometimes correspond to a whole English sentence (Example: the Arabic word "أتتذكّروننا" corresponds in English to: "Do you remember us").

The aim of a morphological analysis step is to recognize word composition and to provide specific morphological information about it. For

[1] http://www.statmt.org/moses/

Example: the word "يعرفون" (in English: they know) is the result of the concatenation of the prefix "ي" indicating the present and suffix "ون" indicating the plural masculine of the verb "عرف" (in English: to know). The morphological analyzer determines for each word the list of all its possible morphological features.

In Arabic language, some conjugated verbs or inflected nouns can have the same orthographic form due to absence of vowels (Example: non-voweled Arabic word "فصل" can be a verb in the past "فَصَلَ" (He dismissed), or a masculine noun "فَصْلٌ" (chapter / season), or a concatenation of the coordinating conjunction " فَ" (then) with the verb "صل": imperative of the verb (bind)).

In order to handle the morphological ambiguities, we decide to use MORPH2, an Arabic morphological analyzer developed at the Miracl laboratory[2]. MORPH2 is based on a knowledge-based computational method. It accepts as input an Arabic text, a sentence or a word. Its morphological disambiguation and analysis method is based on five steps:

- A tokenization process is applied in a first step. It consists of two sub-steps. First, the text is divided into sentences, using the system Star (Belguith et al., 2005), an Arabic text tokenizer based on contextual exploration of punctuation marks and conjunctions of coordination. The second sub-step detects the different words in each sentence.
- A morphological preprocessing step which aims to extract clitics agglutinated to the word. A filtering process is then applied to check out if the remaining word is a particle, a number, a date, or a proper noun.
- An affixal analysis is then applied to determine all possible affixes and roots. It aims to identify basic elements belonging to the constitution of a word (the root and affixes i.e. prefix, infix and suffix).
- The morphological analysis step consists of determining for each word, all its possible morpho-syntactic features (i.e, part of speech, gender, number, time, person, etc.). Morpho-syntactic features detection is made up on three stages. The first stage identifies the part-of-speech of the word (i.e. verb "فعل", noun "اسم", particle "أداة" and proper noun "اسم علم"). The second stage extracts for each part-of-speech a list of its morpho-syntactic features. A filtering of these feature lists is made in the third stage.
- Vocalization and validation step: each handled word is fully vocalized according to its morpho-syntactic features determined in the previous step.

5 Challenges on English-Arabic SMT

In this section, we briefly explore the challenges that prevent the construction of successful SMT. The divergence of Arabic and English puts a rocky barrier in building a prosperous machine translation system. Morphological and syntactic preprocessing is important in order to converge the two languages.

Arabic is a highly agglutinative language with a rich set of suffixes. Inflectional and derivational productions introduce a big growth in the number of possible word forms. In Arabic, articles, prepositions, pronouns, etc. can be affixed to adjectives, nouns, verbs and particles to which they are related. The richness in morphology introduces many challenges to the translation problem both to and from Arabic.

In general, ambiguities in Arabic word are mainly caused by the absence of the short vowels. Thus, a word can have different meanings. There are also the usual homographs of uninflected words with/without the same pronunciation, which have different meanings and usually different POS's. For example: the word "ذهب", can correspond in English to: "gold" or to: "go". In Arabic there are four categories of words: noun, proper noun, verbs and particles. The absence of short vowels can cause ambiguities within the same category or cross different categories. For example: the word "بعد" corresponds to many categories (table 1).

meanings of a word "بعد"	**Categories**
after	Particule
remoteness	Noun
remove	Verb
go away	Verb

Table 1: Different meanings of the word "**بعد**"

[2] http://www.miracl.rnu.tn

In table 1, there exist four different analyses for the word "بعد". This ambiguity can be resolved only in the phrase context.

Due to the Arabic is an agglutinative language, the morphological decomposition is required. So as mentioned above, both training and decoding use segmented Arabic. The final output of the decoder must therefore be recombined into a surface form. This proves to be a non-trivial challenge for a reason that Arabic uses diverse systems of prefixes, suffixes, and pronouns that are attached to the words (Soudi et al., 2007). For example, the Arabic sentence "قبل ت عرض ك" can be recombined as presented in table 2.

Recombined sentence	meanings
قبل تعرضك	Before exposure
قبلت عرضك	Accepted the offer

Table 2: Ambiguity in recombining sentence

6 Statistical Feature Language Model

One of the problems of statistical language models is to consider that the word is depending only on its previous history (words or classes). But in fact, in natural language the appearance of a word depends not only on its history but also on some others features. The word "كتب" (write) and "كتب" (books) are two different words, but we can't predict them if we don't know their features and their contexts.

In order to settle such problem we are trying to introduce knowledge about the word features by using a featured statistical language model: Statistical Feature Language Model (Smaïli et al., 2004).

Arabic is an inflected natural language, linguistic features are very useful to reduce translation errors due to homonyms. By employing SFLM, each word is considered as an array of m features:

$$w_i^{1..m} = \begin{pmatrix} f_1^i \\ f_2^i \\ . \\ . \\ f_m^i \end{pmatrix} \qquad (2)$$

Each f_j^i is a linguistic characteristic of w_i. These characteristics or features could be the surface word, its syntactic class, its gender, its number, its semantic class, ...

(Smaïli et al., 2004) substitute in the classical n-gram language model, the words by their feature arrays which contain surface words and their linguistic characteristics. Thus, a SFLM model is built up by analogy with the classical n-gram model given by:

$$P(w_1, w_2, \dots, w_L) = \prod_{i=1}^{L} P(w_i | w_{i-1} \dots w_{i-n+1}) \qquad (3)$$

To define SFLM model it is enough to replace each word w_i by its feature array $(f_1^i, f_2^i, \dots, f_m^i)^t$ as follows:

$$P\left(w_1^{1..m}, w_2^{1..m}, \dots, w_L^{1..m}\right) = \prod_{i=1}^{L} P\left(\begin{pmatrix} f_1^i \\ f_2^i \\ . \\ . \\ f_m^i \end{pmatrix} \middle| \begin{pmatrix} f_1^{i-1} \\ f_2^{i-1} \\ . \\ . \\ f_m^{i-1} \end{pmatrix} \dots \begin{pmatrix} f_1^{i-n+1} \\ f_2^{i-n+1} \\ . \\ . \\ f_m^{i-n+1} \end{pmatrix}\right) \qquad (4)$$

Where $(f_1^i, f_2^i, \dots, f_m^i)^t$ is the feature array corresponding to the i^{th} word. This model is very simple to implement with classical language modeling toolkits like CMU (Clarkson and Rosenfeld, 1997) and SLRIM (Stolcke, 2002). In fact, we replace each word in the Arabic training and test corpora by its feature array. Thus the following notation is adopted:

$$w_i^{1\dots m} = f_1^i\_f_2^i, \dots, \_f_m^i \qquad (5)$$

The feature array $f_1^i\_f_2^i, \dots, \_f_m^i$ will be treated like only one string. In our experiments, we decided to employ a SFLM with two features. We choose to consider the word itself as first feature and its syntactic class (category) as second one. In this case, a word w_i is represented like the concatenation of the two strings w_i and $C(w_i)$ as follows:

$$w_i\_C(w_i) \qquad (6)$$

where $\mathbf{C(w_i)}$ represents the morpho-syntactic class of $\mathbf{w_i}$.

7 Arabic recombination

As mentioned in Section 1, Arabic is characterized by a rich morphology. In addition to being inflected for gender and number, words can be attached to various clitics for conjunction "و" (and), the definite article ال""the), prepositions) ب" .e.g)" (by/with), "ل" (for), "ك" (as)) and object pronouns (e.g. "هم" (their/them)).

We apply decomposition before aligning the training data, by splitting off each clitic and affix agglutinated to the word separately, such that any given word is split into at most five parts:

Proclitic + prefix+ stem +suffix + enclitic.

Then, the stem is associated with its morpho-syntactic feature. For example the word "أتعرفونهم" (in English: "do you know them" is replaced by:

أ ت عرف_ فعل ون هم

So in both training and decoding processes, segmented Arabic words are used. The final output of the decoder will be also a list of segmented words. Therefore this output must be recombined into a surface form to be able to evaluate the translation result by using the right surface words.

This proves to be a non-trivial challenge for a reason of order ambiguity: a segmented word can be recombined into two grammatically correct forms. Clitics can correspond to enclitic or proclitic. For example: in the segmented words: "سلم ت ك ذلك ال كتاب" the clitic "ك" can be recombined with the previous word ("ك": enclitic). So the segmented words "سلم ت ك ذلك ال كتاب" can b erecombined to "سلمتك ذلك الكتاب", in English: "I gave this book".

The clitic "ك" can be recombined also with the following word ("ك": proclitic), in this case, the segmented words "سلم ت ك ذلك" can be recombined to "سلمت كذلك الكتاب", in English: "I also gave the book".

Those two sentences have the same segmented form, but they have different meanings. By introducing morphological features (e.g. proclitic, prefix, stem, suffix and enclitic) for each segment, we may remove this ambiguity:

Therefore we apply reconstruction of the Arabic segmented words by agglutinating the morphological segments in the following order:

أ_ proclitic ت_ prefix عرف_ فعل ون_ suffix هم_ enclitic

8 Experiments

8.1 Used data

In this paper, we consider the translation task of texts from English into Arabic. We used IWSLT2010 data as a parallel corpus. For training the translation models, the train part of the IWSLT10 data was used which contains 19972 sentence pairs. For testing, we used a subset data made up of 469 sentences (there were 1 Arabic reference translation for each Arabic sentence). All BLEU scores presented in this paper are case-sensitive and include punctuations. For the Arabic language model we use trigrams to build up the baseline system and a 7-grams to build up our translation system. In fact, we use a 7-gram language model because in our system, each word in the training Arabic corpus is replaced by its list of morphological segments: proclitic, prefix, stem, suffix and enclitic.

8.2 Baseline system

The English-Arabic baseline system is built upon the open-source MT toolkit Moses (Koehn et al., 2007). Phrase pairs are extracted from word alignments generated by GIZA++ (Och and Ney. 2003). The phrase-based translation model provides direct and inverted frequency-based and lexical-based probabilities for each phrase pair. To train the trigram language models, SRILM (Stolcke, 2002) was used. The performances reported in this paper were measured using the BLEU score (Papineni et al., 2002).

8.3 Experimental results

- *Arabic word segmenter:*

In our method, each Arabic word, from the target training data, is replaced by its segmented form.

For example: the word "فعرفناهم" (in English: "and we have known them") is the result of the concatenation of the proclitic "فَ" (then): coordinating conjunction, the suffix "نا" for the present masculine plural, enclitic "هم" (for the masculine plural posession pronoun), and the rest of the word "عرف" indicating the stem. So, the word "فعرفناهم" will be replaced by:

"ف عرف نا هم"

- *SFLM for introducing Morpho-syntactic features:*

For introducing morpho-syntactic features into the English-Arabic translation system, we use part of speech tagging provided by MORPH2. We believe that using these features can improve

our language modeling when used with the SFLM model.

In our proposed method, each Arabic word, from the target Arabic training data, is replaced by the reduced word (obtained by removing its clitics and its affixes), combined with its syntactic class (category), where clitic and affix are featured with their morphological classes (e.g. proclitic, prefix, suffix and enclitic).

For example : the word "سيخبرهم" (in English: "he will notify them") is the result of the concatenation of the proclitic "س" indicating the future, the prefix "ي" for the present, enclitic "هم" (for the masculine plural posession pronoun), and the rest of the word "خبر" such as its syntactic class is verb: "فعل". So, the word "سيخبرهم" will be replaced by:

"enclitic_هم خبر_فعل prefix_ي proclitic_س "

In this notation, its morpho-syntactic feature (as verb "فعل", noun "اسم", particle "أداة" and proper noun "اسم علم"). The language model is then generated using the so obtained target Arabic training data, by the standard SRILM toolkit. The so obtained Arabic corpus in then used for training (without any change on the English side).

- Arabic post-processing

As mentioned above, both training and decoding phases use Arabic segmented words. The final output of the decoder will be also composed of segmented words. Therefore these words must be recombined into their surface forms. Therefore we apply reconstruction of the Arabic segmented words just by agglutinating the morphological segments in the following order:

Proclitic + prefix+ stem +suffix + enclitic.

The English-Arabic translation performance of this new system is reported in table3, and compared to the baseline system.

	Bleu
Baseline	12.58%
SMT hybrid	13.16%

Table 3: Comparison of the English-Arabic translation systems

Table 3 shows a significant improvement of the BLEU score when we use segmentation and introduce morpho-syntactic features into the English-Arabic translation system by using SFLM. The BLEU score increases from 12.58% to 13.16%.

These results attest that the use of morpho-syntactic features within SMT system can enhance translation performances, especially for agglutinative and inflectional languages, such as Arabic. Also, using the word category concatenated to the word, can avoid the problem of homographics and can improve language modeling efficacity.

9 Conclusion

English-to-Arabic machine translation has been a challenging research issue for many researchers in the field of Arabic Natural Language Processing. In this study, we have evaluated the effectiveness of morphological decomposition of the Arabic text and SFLM language modeling method to integrate morpho-syntactic features in English to Arabic machine translation. We also presented our method for recombining the segmented Arabic target. Our results suggest that morphological decomposition of the Arabic text is beneficial and that using morpho-syntactic features is a promising way to improve English to Arabic machine translation. The use of recombination of the target side technique is beneficial to overcome ambiguity in recombining Arabic text.

References

Badr I., Zbib R. and Glass J. 2008. Segmentation for English-to-Arabic statistical machine translation. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, Columbus, Ohio, 153-156.

Belguith L., Baccour L. and Mourad G. 2005. Segmentation des textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules". Actes de la 12éme Conférence annuelle sur le Traitement Automatique des Langues Naturelles, 451-456.

Besacier L., Ben-Youcef A. and Blanchon H. 2008. The LIG Arabic / English Speech Translation System. IWSLT08. Hawai. USA, 58-62.

Bilmes J. and Kirchhoff K. 2003. Factored language models and generalized parallel backoff". In Proceeding of Human Language Technology Conference, Edmonton, Canada. 4–6.

Brown P., Della Pietra V., Della Pietra S., and Mercer R. 1993. The mathematics of statistical machine

translation: parameter estimation, Computational Linguistics, 19(1): 263–311.

Buckwalter T. 2002. Buckwalter Arabic morphological analyzer version 1.0. Linguistic Data Consortium, University of Pennsylvania.

Carpuat M, Marton Y, and Habash N. 2010. Improving arabic-to-english statistical machine translation by reordering post-verbal subjects for alignment. In Proceedings of the Association for Computational Linguistics (ACL 2010) Conference Short Papers, Uppsala, Sweden, 178–183.

Clarkson P. and Rosenfeld R. 1997, Statistical language modeling using the CMU-Cambridge toolkit. In Proceedings of the European Conference on Speech Communication and Technology, Rhodes, Greece, 2707-2710.

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08, 224–233, Stroudsburg, PA, USA.

Habash N. and Rambow O. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), Ann Arbor, MI, 573–580.

Habash N. and Sadat F. 2006. Arabic Preprocessing Schemes for Statistical Machine Translation. In Proc. of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, New York City, NY, 49–52.

Habash N., Dorr B., and Monz C. 2006. Challenges in Building an Arabic-English GHMT System with SMT Components. In Proceedings of the 11th annual conference of the European Association for Machine Translation (EAMT-2006), Norway, 56-65.

Kholy A. and Habash N. 2010. Techniques for arabic morphological detokenization and orthographic denormalization. In Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC), Valletta, Malta.

Koehn P. and Hoang H. 2007. Factored translation models. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, 868–876.

Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowa B., Shen W., Moran C., Zens R., Dyer C., Bojar O., Constantin A., and Herbst E., 2007. Moses: Open source toolkit for statistical machine translation, in Proceedings of the ACL-2007 Demoand Poster Sessions, Prague, Czeck Republic, 177–180.

Lee Y. S. 2004. Morphological Analysis for Statistical Machine Translation. In Proceedings of HLT-NAACL: Short Papers on XX, Boston, Massachusetts, 57-60.

Marcu D. and Wong W. 2002. A Phrase-Based, Joint Probability Model for Statistical Machine Translation. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), Philadelphia, PA, 133-139.

Och F. J., and Ney H., 2003. A Systematic comparison of various statistical alignment models, Computational Linguistics, 29(1): 19-51.

Papineni K. A., Roukos S., Ward T., and Zhu W.J., 2002. Bleu: a method for automatic evaluation of machine translation. The Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, 311–318.

Sadat F. and Habash N. 2006. Combination of Arabic preprocessing schemes for statistical machine translation". In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL (Coling ACL'06), Sydney, Australia, 1–8.

Sarikaya R. and Deng Y. 2007. Joint Morphological-Lexical Language Modeling for Machine Translation. In Proc. of NAACL HLT, Rochester, NY, 145-148.

Sawaf H. 2010. Arabic Dialect Handling in Hybrid Machine Translation. In Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA 2010), Denver, Colorado.

Schwenk H., Déchelotte D. 2007. Bonneau-Maynard H. and Allauzen A., "Modèles statistiques enrichis par la syntaxe pour la traduction automatique". TALN 2007, Toulouse-France. 253-262.

Smaïli K., Jamoussi S., Langlois D. and Haton J. P. 2004. Statistical feature language model. INTERSPEECH, Korea, 1357-1360.

Soudi A., Bosch A. and Neumann G. 2007, Arabic Computational Morphology: Knowledge-based and Empirical Methods. In Arabic Computational Morphology, Springer, 3-14.

Stolcke A., 2002. SRILM an Extensible Language Modeling Toolkit. The Proc. of the Intl. Conf. on Spoken Language Processing, Denver, CO, USA, 901–904.