The MIRACL Arabic-English Statistical Machine Translation System for IWSLT 2010

Ines Turki Khemakhem, Salma Jamoussi, Abdelmajid Ben Hamadou

MIRACL Laboratory, ISIM Sfax, Pôle Technologique, Route de Tunis Km 10,
B.P. 242 SFAX 3021, Tunisia
{ines_turki@yahoo.fr, salma.jammoussi@isimsf.rnu.tn,
abdelmajid.benhamadou@isimsf.rnu.tn}

Abstract

This paper describes the MIRACL statistical Machine Translation system and the improvements that were developed during the IWSLT 2010 evaluation campaign. We participated to the Arabic to English BTEC tasks using a phrase-based statistical machine translation approach. In this paper, we first discuss some challenges in translating from Arabic to English and we explore various techniques to improve performances on a such task.

Next, we present our solution for disambiguating the output of an Arabic morphological analyzer. In fact, The Arabic morphological analyzer used produces all possible morphological structures for each word, with an unique correct proposition. In this work we exploit the Arabic-English alignment to choose the correct segmented form and the correct morpho-syntactic features produced by our morphological analyzer.

1. Introduction

Translating two languages with very different morphological structures, such as English and Arabic poses a challenge to successful construction of statistical machine translation (SMT) used models [1]. Thus, the morphological preprocessing is a crucial step to converge with the morphological proprieties of the two languages.

Arabic words are often ambiguous in their morphological analysis. This is due to Arabic’s rich system of affixation and clitics and the omission of short vowels. The problem is that many words have different meanings depending on their diacritization. This leads to ambiguity when processing data for natural language processing applications such as machine translation.

In this paper, we present our SMT system used in the IWSLT2010 evaluation campaign. We first apply a morphological segmentation step for Arabic words where we identify syntactic class of each segmented word. Then we present a novel morphology preprocessing technique for Arabic. We exploit the Arabic-English morphology alignment to choose the correct segmented form and morpho-syntactic features produced by an Arabic morphological analyzer. Our goal is to improve the quality of our statistical translation system.

This paper is organized as follows: section 2 gives a brief description of some related works to the introduction of morphological analyzers and morpho-syntactic features in a machine translation process. In Section 3, an overview of the baseline SMT is given. 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 morphological disambiguation. We describe in section 6 our method of handling ambiguities on the Arabic morphological analyzer output. Section 7 gives a short overview of the data and tools used to build up our SMT system and gives its evaluation results, which are discussed in Section 8. Finally, section 9 concludes and suggests possible directions for future work.

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.

In prior work [2][3], on Arabic-to-English SMT it has been shown that morphological segmentation of the Arabic source benefits the performance of the SMT system. In [2], author uses a trigram language model to segment Arabic words. He then identifies functional morphemes to be merged or to be deleted in order to induce a symmetrical morphological structure. Habash and Sadat [3] compared the use of the BAMA [4] and MADA [5] toolkits to segment the Arabic source, able to improve translation for Arabic-English task. Sadat and Habash [6] also showed that it was possible to combine the use of several variations of morphological analysis both while decoding 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 [7] conducted an in depth 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 problem has 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. In [8], authors focus on incorporating morpho-syntactic features in the translation model for the English-Spanish machine translation process. They 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.

3. Phrase-Based Machine Translation

Statistical machine translation methods have evolved from using the simple word based models [1] to phrase based models ([9]; [10]; [11]).

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:

$$argmax_t P(t|s) = argmax_t P(s|t)P(t) \qquad (1)$$

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 a decoder. In our case, we use the open source Moses decoder described in [12].

4. Morphological segmentation

Arabic is a morphologically complex language. An Arabic word often corresponds to more than one independent word in English (Example: the Arabic word "أتتذكروننا" corresponds in English to the whole sentence: "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 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 this work, in order to handle the morphological ambiguities, we decide to use MORPH2 [13], an Arabic morphological analyzer developed at the Miracl laboratory[1]. 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 [14], 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 Arabic-English SMT

In this section, we briefly explore the challenges that prevent the construction of successful Arabic-English SMT system. In fact, the divergence of Arabic and English puts a rocky barrier in building a prosperous machine translation system. Thus, the morphological and syntactic preprocessing is an important step to converge with the morphological properties of the two languages.

Arabic is a highly agglutinative language with a rich set of suffixes. Its 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 across different categories. For example: the word "بعد" corresponds to many categories (table 1).

Table 1: Different meanings of the word "بعد"

meanings of a word "بعد"	Categories
After	Particule
Remoteness	Noun
Remove	Verb
go away	Verb

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

Arabic uses diverse prefixes, suffixes, and pronouns that can be attached to the words [15] and so correct

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

morphological analysis is required to resolve structural ambiguities among Arabic sentence. Identifying such particles is crucial for analyzing syntactic structures. So, there are multiple ways to segment a word as a list of morphemes. For example, the word "بعيد" can be segmented as presented in table 2.

Table 2: Ambiguity in segmenting one word

Segmented word	meanings	Categories
بعيد	Far	Noun
ب+عيد	By + Holiday	Particle + Noun

The previous example shows that disambiguating Arabic is a difficult task. This ambiguity can be resolved only within the phrase context. The segmentation is driven by the context of the word and by its structural dependencies in the sentence.

6. Disambiguation of morphological and syntactic analysis

6.1. Alignment step

The training corpus used in this work is the supplied Arabic-English BTEC training corpus, aligned at the sentence level. Each Arabic word, from Arabic data, is replaced by its first segmented form generated by MORPH2. In the other side, the English corpus is part-of-speech (POS) tagged by using treetagger tool [16] for annotating text with part-of-speech and lemma information.

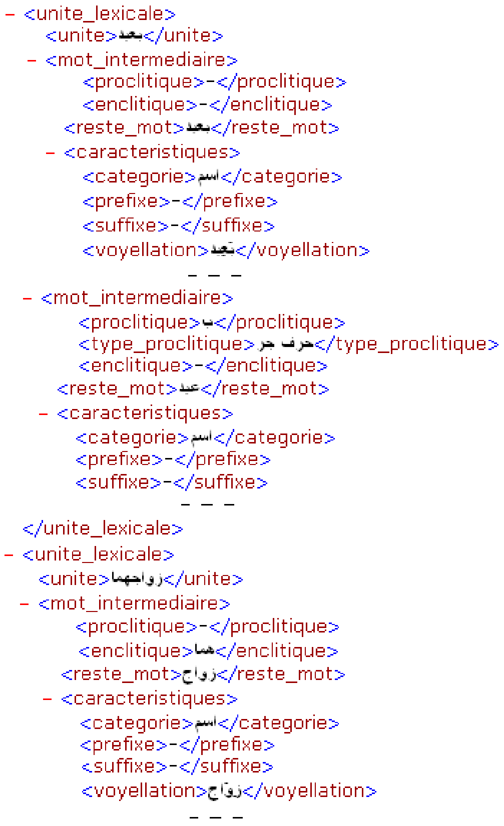
The disambiguation technique is implemented as a two-step morphological processing. We first apply word segmentation to Arabic. Arabic-English sentence alignment is illustrated in Figure 2, where each Arabic morpheme is aligned to one or zero English word. We then use the English word POS to identify the correct segmented form and POS for Arabic morphemes.

The alignment model was trained with GIZA++ [20] toolkit, which implements the most typical IBM and HMM alignment models for translation. The alignment models used in our case are IBM-1, HMM, IBM-3 and IBM-4.

6.2. Arabic-English morphological alignment for Arabic Word disambiguation

We pre-process Arabic data using the MORPH2 morphological analyzer, described in section 4. A sample of the morphological analyzer output is shown in Figure 1.

The obtained output consists of a set of all possible morphological analysis for each word. But only one proposition is correct. In figure 1, there exist two different analyses for the word "بعيد". This ambiguity can be resolved only by the phrase context. In a first attempt to Arabic-English SMT, we choose the first morphological analysis given by MORPH2. However by looking at the context, we notice in this example that the first morphological analysis is incorrect. Thus one needs to select the right segmented form and POS to ensure good SMT performance. Given the highly inflection nature of Arabic, resolving ambiguities is a hard task.



```
- <unite_lexicale>
    <unite>بعيد</unite>
  - <mot_intermediaire>
        <proclitique>-</proclitique>
        <enclitique>-</enclitique>
      <reste_mot>بعيد</reste_mot>
    - <caracteristiques>
        <categorie>اسم</categorie>
        <prefixe>-</prefixe>
        <suffixe>-</suffixe>
        <voyellation>بَعِيد</voyellation>
                    - - -
  - <mot_intermediaire>
        <proclitique>ب</proclitique>
        <type_proclitique>حرف جر</type_proclitique>
        <enclitique>-</enclitique>
      <reste_mot>عيد</reste_mot>
    - <caracteristiques>
        <categorie>اسم</categorie>
        <prefixe>-</prefixe>
        <suffixe>-</suffixe>
                    - - -
  </unite_lexicale>
- <unite_lexicale>
    <unite>زواجهما</unite>
  - <mot_intermediaire>
        <proclitique>-</proclitique>
        <enclitique>هما</enclitique>
      <reste_mot>زواج</reste_mot>
    - <caracteristiques>
        <categorie>اسم</categorie>
        <prefixe>-</prefixe>
        <suffixe>-</suffixe>
        <voyellation>زَوَاج</voyellation>
                    - - -
```

Figure 1: Possible analyses for the sentence "بعيد زواجهما" (in English: "by their wedding")

6.2.1. *Disambiguation of Arabic word segmentation*

Many of the ambiguities can be resolved by looking at the context. The example, illustrated in Figure 1, shows how difficult to disambiguate Arabic words. The non-voweled Arabic word "بعيد" have two possible morphological analyses. It can be a noun "بَعِيد" (in English: far), or a concatenation of the preposition "ب" (in English: by) with the noun "عيد" (in English: holiday).

A simple disambiguation technique consists of only taking the first morphological analysis generated by MORPH2. Each segmented Arabic word is given and its stem is associated with a morpho-syntactic feature (as verb "فعل" and noun "اسم" and particle "أداة" and proper noun "اسم علم"). In this example the Arabic word "بعيد" will be considered as noun "اسم". We then apply Arabic-English sentence alignment where the English corpus is part-of-speech (POS) tagged using TreeTagger [16]. Contrary to other probabilistic tagging methods, which have difficulties in estimating small probabilities accurately from limited amounts of training data, the TreeTagger avoids the sparse data problem by using a binary decision tree, which determines the appropriate size of the context used to estimate the transition probabilities.

GIZA++ outputs alignment for the sentence pair as depicted in (b) and (c) of figure 2 where we consider the Arabic sentence "بعيد زواجهما" (in English: "by their wedding").

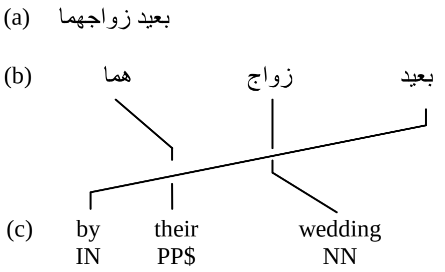


Figure 2: (a) Original Arabic sentence, (b) the segmented Arabic sentence as given by the first output of the morphological analyzer MORPH2, (c) English translation and its alignment with its POS.

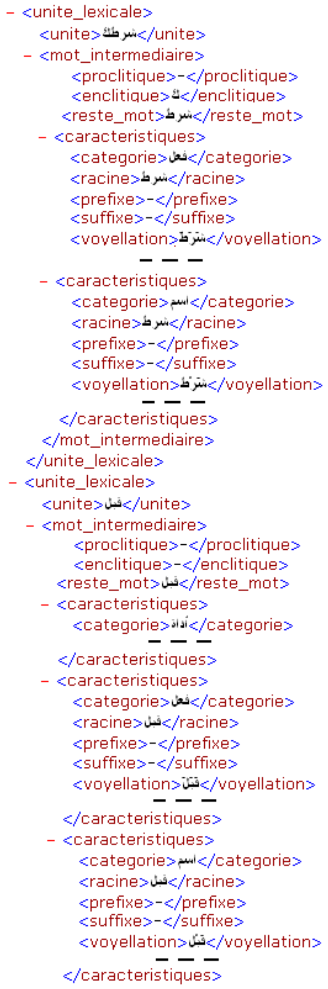
The morpheme "بعيد" is aligned to the English preposition: "by". Thus, we notice that these two words have two different part of speech ("بعيد" is a noun and "by" is a preposition). We can deduce that the segmented form of the word "بعيد" is incorrect. We then select, from the output of MORPH2, the correct segmented form where a preposition appears in the morphological analysis. In our example, the correct segmented form is "ب عيد" where "ب" is the preposition ("حرف جر"). So the Arabic sentence " بعيد زواجهما" will be segmented as:

"ب عيد زواج هما"

So, the use of the Arabic-English morphological alignment can help us to choose the correct segmentation produced by our morphological analyzer. More details are given in the next section to explain how to disambiguate the morpho-syntactic properties of Arabic words.

6.2.2. *Morpho-syntactic feature disambiguation*

Derivational, flexional and agglutinative aspects of Arabic yield prominent challenges in machine translation qualities. Thus, many morphological ambiguities have to be solved when dealing with Arabic language. In fact, many Arabic words are homographic: they have the same orthographic form, though the pronunciation is different [17]. In most cases, these homographs are due to the non vocalization of words. The example illustrated in figure 3 shows that non-vowelled Arabic word can be analyzed in multiple ways. The Arabic word "شرطك" is a concatenation of the stem "شرط" with the enclitic "ك" (for the possession pronoun). The stem "شرط" can be a verb (in English: stipulating) or a noun (in English: condition). We can also show in this example that non-voweled Arabic word "قبل" can be a verb (in English: accepted), or a noun (in English: kiss), or a particle (in English: before).



```
- <unite_lexicale>
    <unite>شرطك</unite>
  - <mot_intermediaire>
        <proclitique>-</proclitique>
        <enclitique>ك</enclitique>
      <reste_mot>شرط</reste_mot>
    - <caracteristiques>
        <categorie>فعل</categorie>
        <racine>شرط</racine>
        <prefixe>-</prefixe>
        <suffixe>-</suffixe>
        <voyellation>شَرَطَ</voyellation>
               — — —
    - <caracteristiques>
        <categorie>اسم</categorie>
        <racine>شرط</racine>
        <prefixe>-</prefixe>
        <suffixe>-</suffixe>
        <voyellation>شَرْط</voyellation>
               — — —
      </caracteristiques>
    </mot_intermediaire>
  </unite_lexicale>
- <unite_lexicale>
    <unite>قبل</unite>
  - <mot_intermediaire>
        <proclitique>-</proclitique>
        <enclitique>-</enclitique>
      <reste_mot>قبل</reste_mot>
    - <caracteristiques>
        <categorie>أداة</categorie>
               — — —
      </caracteristiques>
    - <caracteristiques>
        <categorie>فعل</categorie>
        <racine>قبل</racine>
        <prefixe>-</prefixe>
        <suffixe>-</suffixe>
        <voyellation>قَبِلَ</voyellation>
               — — —
      </caracteristiques>
    - <caracteristiques>
        <categorie>اسم</categorie>
        <racine>قبل</racine>
        <prefixe>-</prefixe>
        <suffixe>-</suffixe>
        <voyellation>قُبَل</voyellation>
               — — —
      </caracteristiques>
```

Figure 3: Possible analyses for the sentence "شرطك قبل" (in English: "your condition was accepted")

In Figure 3, we show that in the first output of the morphological analyzer of the Arabic sentence "شرطك قبل", the Arabic word "شرط" is considered as a verb: "فعل" and "قبل" is particle: "أداة". To improve this POS tagger results, we use the word alignment output of the segmented Arabic corpus to the part-of-speech tagged English corpus. In fact, each word from the English corpus was tagged using TreeTagger.

In a next step we used the GIZA++ toolkit to align Arabic and English sentences [18]. GIZA++ outputs alignment for a sentence pair in the corpus as depicted in (b) and (c) of figure 4. The illustrated example concerns the Arabic sentence "شرطك قبل" (in English: "your condition was accepted").

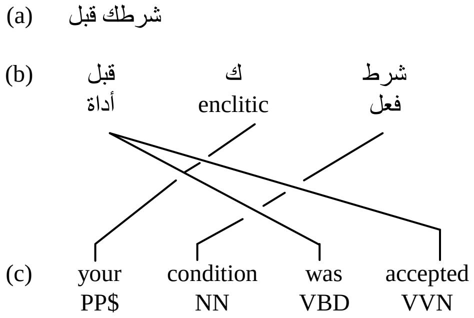


Figure 4: (a) Original Arabic sentence, (b) First output of morphological analysis with MORPH2: segmented Arabic sentence, (c) English translation and its alignment with a morphological analysis.

The morpheme "شرط" and "قبل" are aligned respectively to the English word: "condition" and "was accepted", where their syntactic classes are respectively noun: "اسم" and verb: "فعل". We can so select the correct morpho-syntactic feature produced by our morphological analyzer and in this case we will choose the second morphological analysis given by MORPH2. Thus, the part of speech of the stem "شرط" and "قبل" in the phrase context are respectively noun and verb.

The part-of-speech provided by TreeTagger is verb, proper noun, noun, adjective, adverb, conjunction, pronoun, preposition, etc. The morpho-syntactic feature of Arabic words aligned with English words tagged by adjective or noun will be replaced by the morpho-syntactic feature: noun: "اسم". While, the morpho-syntactic feature: adverb, conjunction, or prepostion will be replaced by the Arabic morpho-syntactic feature: particle: "أداة".

We can attest that the use of the Arabic-English morphological alignment can help us to choose correct morpho-syntactic features among those produced by our morphological analyzer. Thus we can use SMT alignment step to improve the POS tagging task, especially for agglutinative and inflectional languages like Arabic.

6.3. Disambiguation for translation

We first apply word segmentation to Arabic data. Each Arabic word from Arabic data is replaced by its first segmented form generated by MORPH2, where the stem is marked with its syntactic class. Then we apply our disambiguation technique to the Arabic data to choose the correct segmented form and morpho-syntactic feature produced by our morphological analyzer. The translation table was trained using the so obtained parallel data (no change was made on the English side). In decoding, the disambiguation technique was applied to the test input but manually using the Arabic-English morphological alignment of the training data.

7. Experiments

7.1. Used data

This is the first year that the Miracl laboratory participate to the evaluation campaign of the International Workshop on Spoken Language Translation (IWSLT). We submitted a run for the Arabic-English BTEC task[2]. We have used the data provided by the IWSLT10 organizers. For training the translation models, the train part of the IWSLT10 data was used (a training corpus of 19972 sentence pairs). As development data, we used provided subsets: the dev6 subset, made up of 489 sentences, which corresponds to the IWSLT07 development data, the dev6 have 6 English reference segments per source segment; the dev7 subset, made up of 507 sentences, which corresponds to the IWSLT08 development data (there were 16 English reference translations for each Arabic sentence). For testing datasets, we used provided subsets: the tst09, made up of 469 sentences; the tst10 subset, made up of 464 sentences, which corresponds respectively to the TWSLT09 and IWSLT10 test set. All BLEU scores presented in this paper are case-sensitive and include punctuations.

7.2. Baseline system

Our systems were trained on the 20k *train* bitext provided. The *moses* training script was used to build a phrase translation table from the bitext.

The Arabic-English baseline system is built upon the open-source MT toolkit Moses [3] [12]. Phrase pairs are extracted from word alignments generated by GIZA++ [20]. The phrase-based translation model provides direct and inverted frequency-based and lexical-based probabilities for each phrase pair. The English sides of the training corpora were used to generate 5-gram target language model for the translation task. For this purpose, the SRI language modeling toolkit [19] was used. The performances reported in this paper were measured using the BLEU score [21].

7.3. Experimental results

7.3.1. Arabic word segmenter

The Arabic part of the bitext was systematically segmented to train the phrase tables. We pre-process Arabic data and we present each word by its proclitic-prefix-stem-suffix-enclitic form, using our morphological analyzer MORPH2, as described in section 4. The category proclitic, prefix, suffix, enclitic encompasses function words such as conjunction markers, prepositions, pronouns, determiners and all inflectional morphemes of the language. 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 possession pronoun), and the rest of the word "عرف" as a verb: "فعل". A sample Arabic segmented word is given below, where clitic and affix are featured with their morphological classes (i.e. proclitic, prefix, suffix and enclitic) and stem is marked with its syntactic class (i.e. verb "فعل", noun "اسم", particle "أداة" and proper noun "اسم علم").

"enclitic_ هم suffix_نا عرف_فعل proclitic_ف"

The segmentations in this model are static in that all the occurrences of a word are assumed to be segmented in the same manner regardless of the context.

[2] Basic Travel Expression Corpus (BTEC)

[3] Moses open source project: http://www.statmt.org/moses

We used the supplied BTEC training corpus as input to MORPH2, and the algorithm converged to the model described above. Thus the output that defines the Arabic side of the training corpus consists of replacing each word by its segmentation according to the presented model. The resulting corpus was paired with the word-based English corpus to train the translation model. The translation table was trained using the so obtained parallel data (no change was made on the English side). In decoding, the same segmentation model was also applied to the test input.

The Arabic-English translation performance is reported in table3. So we can notice that the Arabic segmentation and the introduction of Arabic morpho-syntactic features heavily improves the translation quality. In fact, segmentation affects the translation models (alignments, phrase table) as well as the translation input.

Table 3: Comparison of BLEU scores with and without introducing MORPH2

System	Dev6	Dev7	Tst09	Tst10
Baseline	37.87	42.74	40.68	34.48
Using MORPH2	44.36	44.03	41.60	36.34

7.3.2. Disambiguation of Arabic word segmentation

Morphological analysis was carried out regardless of the word context and it is not enough to resolve the ambiguities. Therefore, we have exploited the Arabic-English alignment in order to find out which of the segmented forms, produced by our morphological analyzer, must be selected.

Table 4 shows the effect of disambiguation of Arabic word segmentation on the performance of our translation task.

Table 4: Effect of disambiguating Arabic word segmentation on the final translation BLEU scores (T1)

System	Dev6	Dev7	Tst09	Tst10
T1	44.82	44.50	41.86	35.99

Table 4 shows an improvement in BLEU score when used the Arabic-English alignment to disambiguate Arabic word segmentation. The result of Tst10 deteriorates by using disambiguation technique. This degradation could be explained by the differences of vocabulary between training and test set.

Due to Arabic's rich system of affixation and clitics, morphemes can have many surface forms and so correct segmentation is required to resolve structural ambiguities among Arabic sentence. By incorporating Arabic-English alignment, it was possible to improve Arabic word segmentation and so the SMT performance.

7.3.3. Morpho-syntactic feature disambiguation

In this section, we also investigated applying a similar Arabic-English alignment to identify correct morpho-syntactic feature produced by our morphological analyzer MORPH2. Given a word without any context, MORPH2 gives us the set of its all possible morphological tags.

In our experiments, using Arabic-English alignment for Morpho-syntactic feature disambiguation within the system training step provide a clear improvement of the performance of our translation system as shown in Table 5.

Table 5: Effect of Arabic-English alignment technique for morpho-syntactic feature disambiguation on the final translation BLEU scores

System	Dev6	Dev7	Tst09	Tst10
T1	44.82	44.50	41.86	35.99
T1+T2	45.67	46.23	43.35	35.86

8. Results evaluation and discussion

The first developed system was the baseline system enhanced with the segmentation MORPH2 tool, where each stem was marked with its morpho-syntactic feature. This system obtains better results than the baseline system. Thus, Arabic segmentation is very useful for statistical machine translation. In fact, an accurate alignment between the source and the target languages is an important criteria to obtain high quality translations. In addition, using the word category concatenated to the word can avoid the problem of homographics.

Due to Arabic's rich system of affixation and to improve results, we used the Arabic-English alignment to choose the correct segmented form and the right morpho-syntactic features among those produced by our morphological analyzer. We can notice an obvious improvement of results in term of the BLEU score.

Table 6 compares the performance of our three Arabic-English submissions on the 2010 test set. The official evaluation results of our submitted Arabic-English systems (primary and contrastives) are shown in Table 6.

9. Conclusion

This paper described the statistical machine translation systems developed by the MIRACL laboratory for the 2010 IWSLT evaluation campaign. We were interested in Arabic to English Statistical Machine Translation (SMT). Arabic is a morphologically rich language, and morphological analysis and disambiguation of Arabic is a difficult task which involves, in theory, thousands of possible tags.

We proposed to use segmentation for machine translation of Arabic, where stem is marked with its morpho-syntactic feature. Then, we presented a technique to the morphological disambiguation of Arabic text. We used an alignment step to exploit the target language POS tagging for Arabic disambiguation.

Our disambiguation technique is implemented as a two-step morphological processing. We first applied an Arabic word segmentation step using the Arabic morphological analyzer MORPH2, where the first morphological analysis generated by our Arabic morphological analyzer is chosen. We then proposed to exploit the Arabic-English alignment to choose the correct segmented form and its morpho-syntactic features among those produced by our morphological analyzer. In this case, one Arabic morpheme can be aligned to one or zero English word using an English corpus already part-of-speech (POS) tagged. The sentence alignment between Arabic and English corpus was trained with GIZA++ toolkit.

Experiments conducted in the framework of IWSLT evaluation campaign have shown the potential of the exploitation of the Arabic-English alignment to the morphological disambiguation of Arabic in a translation context.

Table 6: Summary of the obtained results by our submitted systems with the IWSLT09 and IWSLT10 evaluation test sets

Case+punct			
Arabic-to-English Systems		BLEU	
System	Features	Tst09	Tst10
Primary	After Arabic word segmentation and morpho-syntactic feature disambiguation	43.35	35.86
contrastive1	Baseline	40.68	34.48
contrastive2	Baseline with MORPH2	41.60	36.34
contrastive3	After disambiguation of Arabic word segmentation	41.86	35.99
No_case+no_punc			
Arabic-to-English Systems		BLEU	
System	Features	Tst09	Tst10
Primary	After Arabic word segmentation and morpho-syntactic feature disambiguation	44.57	35.23
contrastive1	Baseline	41.76	34.32
contrastive2	Baseline with MORPH2	42.14	35.67
contrastive3	After disambiguation of Arabic word segmentation	42.83	35.34

10. References

[1] Brown P., Della Pietra V., Della Pietra S., and Mercer R., "The mathematics of statistical machine translation: parameter estimation", *Computational Linguistics*, 19(1): 263–311, 1993.

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

[3] Habash N. and Sadat F., "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, 2006.

[4] Buckwalter T., "Buckwalter Arabic morphological analyzer version 1.0". *Linguistic Data Consortium, University of Pennsylvania*, 2002.

[5] Habash N. and Rambow O., "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, 2005.

[6] Sadat F. and Habash N., "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, 2006.

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

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

[9] Marcu D. and Wong W., "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, 2002.

[10] Koehn P., "Pharaoh: A Beam Search Decoder for phrase-based Statistical Machine Translation Models". *In R. Frederking & K. Taylor (eds.) Machine Translation: From Real Users to Research; 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, LNAI 3265,* Berlin/Heidelberg, Germany: Springer Verlag, 115–124, 2004.

[11] Och F. J., Ney H., "The alignment template approach to statistical machine translation", *Computational Linguistics*, 30(4): 417-449, 2004.

[12] 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., "Moses: Open source toolkit for statistical machine translation, *Proceedings of the ACL-2007 Demoand Poster Sessions*, Prague, Czeck Republic, 177–180, 2007.

[13] Belguith L., and Chaâben N., "Analyse et désambiguïsation morphologiques de textes arabes non voyellés", *Actes de la 13ème confrence sur le Traitement Automatique des Langues Naturelles*, Leuven Belgique, 493-501, 2006.

[14] Belguith L., Baccour L. and Mourad G., "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, 2005.

[15] Soudi A., Bosch A. and Neumann G., "Arabic Computational Morphology: Knowledge-based and Empirical Methods". In Arabic Computational Morphology, A. Soudi, A. van den Bosch and G. Neumann (eds.), Springer, 3-14, 2007.

[16] Schmid H., "Probabilistic Part-of-speech Tagging Using Decision Trees". *In Proceedings of International Conference on New Methods in Language Processing*, Manchester. 44-49, 1994.

[17] Attia M., "An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks". In The Challenge of Arabic for NLP/MT Conference, the British Computer Society Conference, London, 48-67, 2006.

[18] Turki Khemakhem I., Jamoussi S., and Ben Hamadou A., "Arabic morpho-syntactic feature disambiguation in a translation context". In Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation, COLIN*G 2010*, Beijing, 61–65, 2010.

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

[20] Och F. J., and Ney H., "A Systematic comparison of various statistical alignment models", *Computational Linguistic*s, 29(1): 19-51, 2003.

[21] Papineni K. A., Roukos S., Ward T., and Zhu W.J., "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, 2002.