Master’s Thesis Nr. 11
Systems Group, Department of Computer Science, ETH Zurich

in collaboration with
Google Inc.

Applying Morphology to English-Arabic Statistical Machine Translation

by
Soha Sultan

Supervised by
Dr. Keith Hall (Google Inc.)
Prof. Donald Kossmann

December 2010–May 2011
We introduce two approaches to augmenting English-Arabic statistical machine translation (SMT) with linguistic knowledge. The first approach improves SMT by adding linguistically motivated syntactic features to particular phrases. These added features are based on the English syntactic information, namely part-of-speech tags and dependency parse trees. We achieved improvements of 0.2 and 0.6 in BLEU score on two different data sets over the state-of-the-art SMT baseline system. The second approach improves morphological agreement in machine translation output through post-processing. Our method uses the projection of the English dependency parse tree onto the Arabic sentence in addition to the Arabic morphological analysis in order to extract the agreement relations between words in the Arabic sentence. Afterwards, classifiers for individual morphological features are trained using syntactic and morphological information from both the source and target languages. The predicted morphological features are then used to generate the correct surface forms. Our method achieves a statistically significant improvement over the baseline system according to human evaluation.

Abstract
Contents

List of Figures v
List of Tables vii

1 Introduction 1

2 Log Linear Phrase-Based Statistical Machine Translation 5
  2.1 Alignment . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
  2.2 Phrase Table . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
  2.3 System Training: MERT . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
  2.4 BLEU . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

3 Why is English-Arabic Machine Translation Hard? 9
  3.1 Arabic Morphology . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
    3.1.1 Morphology Function . . . . . . . . . . . . . . . . . . . . . . . . . . 10
    3.1.2 Morphology Type . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
  3.2 Arabic Inflectional Agreement . . . . . . . . . . . . . . . . . . . . . . . . . . 12
  3.3 Sparsity . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
  3.4 Syntactic Divergences . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14
    3.4.1 Word Order . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14
    3.4.2 Verb-less sentences . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
    3.4.3 Possessiveness . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
  3.5 Lexical Divergences . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
    3.5.1 Idiomatic Expressions . . . . . . . . . . . . . . . . . . . . . . . . . . 15
    3.5.2 Prepositions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
    3.5.3 Named Entities . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
## Contents

3.5.4 Ambiguity ........................................ 17  
3.5.5 Alignment Difficulties ............................ 17  
3.6 Error Analysis Summary ............................. 17

4 Related Work ...................................... 19

5 Adding Syntactic Phrase Constraints .............. 23  
5.1 POS Features ....................................... 23  
5.1.1 Personal Pronouns ............................... 24  
5.1.2 Possessive Pronouns ............................. 25  
5.1.3 Prepositions and Particles ....................... 26  
5.1.4 Determiners ..................................... 27  
5.1.5 Wh-nouns ........................................ 28  
5.2 Dependency Features ............................... 28  
5.3 Results ............................................. 29  
5.4 Limitations ......................................... 31

6 Fixing Inflectional Agreement through Post-processing 33  
6.1 System Description ................................ 33  
6.1.1 Algorithms for Inflection Prediction ............ 34  
6.1.2 Arabic Analysis and Generation ................ 36  
6.1.3 Syntax Projection and Relation Extraction ...... 37  
6.1.4 Feature Vector Extraction ....................... 40  
6.1.5 Training and Classification ...................... 42  
6.1.6 Language Model Incorporation .................. 42  
6.2 Evaluation .......................................... 42  
6.2.1 BLEU ............................................ 43  
6.2.2 Human Evaluation ............................... 45  
6.3 Experimental Setting ............................... 46  
6.4 Results ............................................. 47  
6.4.1 Prediction Accuracy .............................. 47  
6.4.2 BLEU Score ..................................... 48  
6.4.3 Human Evaluation Results ...................... 49  
6.5 Limitations ......................................... 49

7 Conclusion ........................................ 51

Bibliography .......................................... 53
List of Figures

5.1 Types of Arabic Pronouns ........................................ 25
5.2 Wrong translation of attached prepositions ....................... 26
5.3 Wrong translation of separate prepositions ....................... 27
6.1 System Diagram .................................................. 34
6.2 Example of Dependency Tree Projection .......................... 38
6.3 acomp relation .................................................... 40
6.4 acomp relation .................................................... 40
List of Tables
Introduction

The current state-of-the-art Phrase-based Statistical Machine Translation (PBSMT) system is limited to mapping chunks of words (phrases) from the source language to the target language. It only relies on lexical information, i.e., the surface forms of the words. The integration of other sources of linguistic information such as morphological or syntactic information is an active area of research. Works integrating such additional information in the translation process have shown significant improvements especially for translation between languages with large topological divergences. English-Arabic is an example of these language pairs. While Arabic-English translation is also difficult, it does not require the generation of rich morphology at the target side. Translation from English to Arabic is the main focus of this work; however, the used techniques are also useful for translation into other morphologically rich languages.

In this paper, we introduce two different techniques for improving English-Arabic translation through incorporating morphology and syntax in SMT. In the first part of the work, we experiment with adding syntactic features to the phrases in the phrase table. The features are based on the English syntactic information namely part-of-speech (POS) tags and dependency relations. POS features are suggested to penalize phrases which consist only of English words that do not correspond to Arabic words. These phrases are sources of error because they usually translate to Arabic words with different meaning or even with different POS tags. An example is a phrase containing only the English word "the" which should map in Arabic to the noun prefix "Al" and never appear as a separate word.¹

Dependency features are features that rely on the syntactic dependency parse tree of the sentences from which a certain phrase was extracted. These features are suggested because they can

¹This is assuming the used Arabic segmentation does not separate the "Al" from nouns. The choice of the set of POS features depends on the used Arabic segmentation. We run our experiments on a state-of-the-art PBSMT system and select the POS features based on the system’s segmentation.
solve a number of error categories, the main two of which are lexical agreement and inflectional morphological agreement. An example of lexical agreement is phrasal verbs where a verb takes a specific preposition to convey a specific meaning. When the verb and the preposition are in separate phrases, they are less likely to translate correctly. However, selecting a phrase containing both words in the same phrase guarantees their lexical agreement. Inflectional agreement is a syntactic-morphological feature of the Arabic language. Some words should have morphological agreement with other words in the sentence, e.g., an adjective should morphologically agree with the noun it modifies in gender, number, etc. Morphological agreement also applies to other related words such as verbs and their subjects, words connected by conjunction and others. To guarantee correct inflectional agreement of two syntactically related words, a phrase containing both words should be selected by the decoder. This guarantees their agreement since phrases are extracted from morphologically correct training sentences. The weights of the added features are then evaluated automatically using the Minimum Error Rate Training (MERT) algorithm. The results show an improvement in the automatic evaluation score BLEU over the baseline system.

In contrast to the first part of the work which aims at improving morphology by adding features and thus modifying the main pipeline of SMT, the second part introduces a probabilistic framework for morphological generation incorporating syntactic, morphological and lexical information sources through post-processing. While dependency features also aim at solving inflectional agreement, it has limitations that are overcome by post-processing. First, dependency features are added only for words which are at small distances in the sentence. This is because all phrase based SMT systems limit the length of phrases. Related words that have distances more than the maximum phrase length are not helped. Second, phrases that contain related words could be absent from the phrase table because they were not in the training data or were filtered because they were not frequent enough. Finally, other features that have more weight than dependency features could lead to selecting other phrases.

Using the decoder of the baseline system, the only component that can motivate selecting the correctly inflected words is the N-gram language model. In most SMT, 3- or 4-gram language models are used, which means only agreement between close words can be captured. The language model can fix the agreement issues assuming the following:

- the correct inflected word form is present in the phrase table.
- inflected phrases having the same semantics are clustered and all other translation feature values are normalized.

If both conditions apply, the correct inflected form of a word can be generated if the agreement relation is with a close word. However, the above two conditions are hard to apply because of the following reasons:

- Sparsity: The correct inflection of a word that agrees with the rest of the sentence might be absent from the phrase table because it was not in the training data or appeared very few times and subsequently got filtered.
- The lack of robust Arabic analysis and disambiguation tools leads to erroneous clustering of words. Because the units in the phrase table are actually phrases and not words, clustering becomes more difficult and more ambiguous. Clustering errors would hurt the
semantic quality of the SMT system, which should be avoided.

Therefore, we propose a different approach to solving agreement issues in SMT through post processing. Our approach avoids the above problems because it:

- relies on syntactic dependencies to identify potential agreements and therefore can handle agreement between largely distant words.
- generates inflected word forms that were never seen in the parallel training data, which helps in solving the sparsity problem.
- works with the output of any machine translation system.
- is language independent.

Our post-processing module improves the inflectional agreement of the Arabic translation output as proven by the automatic and human evaluations.

Chapter 2 is an introduction to state-of-the-art phrase-based statistical machine translation focusing on the background information needed for this work. Chapter 3 is an explanation of the problems that make English-Arabic translation specifically hard. The chapter concludes with the results of the manual error analysis conducted on the output of the baseline SMT system. A summary of the related research that targeted Arabic translation is provided in chapter 4. In the following chapter, we present our experimental study on the incorporation of syntactic features in SMT and report the results of a number of experiments. Finally, in chapter 6, we present our novel post-processing system that improves the morphological agreement of the output of the baseline SMT system. The last chapter concludes this work and suggests directions for future work.
1 Introduction
Log Linear Phrase-Based Statistical Machine Translation

The statistical machine translation system used as the baseline for this work is similar to that described in [ON04]. The log linear approach to SMT uses the maximum-entropy framework to search for the best translation of a source language text given the following decision rule:

$$
\hat{e}_1^T = \arg \max_{e_1^T} \sum_{m=1}^{M} \lambda_m h_m(e_1^T, f_1^T)
$$

where $e_1^T = e_1, e_2, e_3...e_I$ is the best translation for the input foreign language sentence $f_1^T = f_1, f_2, f_3...f_J$. $h_m(e_1^T, f_1^T)$ are the used feature functions including, for example, translation and language model probabilities. The unknown parameters $\lambda_m^M$ are the weights of the feature functions and are evaluated using development data as will be discussed in section 2.3.

Training the translation model starts by aligning the words of the sentence pairs in the training data using, for example, one of the IBM models [BPPM93]. To move from word-level translation systems to phrase-based systems which can capture context more powerfully, the phrase extraction algorithm of [ON00] is used. Subsequently, feature functions could be defined at the phrase level.

Using a translation model which translates a source-language sentence $f$ into a target-language sentence $e$ through maximizing a linear combination of features and weights allows easily extending it by defining new feature functions [ON04].

A brief description of the alignment algorithm follows in section 2.1. Section 2.2 briefly describes phrase tables and basic phrase features. Afterwards, section 2.3 provides an introduction of the MERT algorithm which is responsible for calculating the optimal feature weights of equation 2.1. Finally, the automatic evaluation metric BLEU is described in section 2.4.
2 Log Linear Phrase-Based Statistical Machine Translation

2.1 Alignment

For every sentence pair \((e_1^I, f_1^J)\), the Viterbi alignment is the alignment \(a_1^I\) such that:

\[
a_1^I = \arg\max_{a_1^I} Pr(f_1^I, a_1^I|e_1^I)
\] (2.2)

where \(a_j\) is the index of the word in \(e_1^I\) to which \(f_j\) is aligned.

Word alignments can be calculated using GIZA++ [ON00] which uses the IBM models 1 to 5 and the HMM alignment model, all of which do not permit a source word to align to multiple target words. GIZA++ allows many-to-many alignments by combining the Viterbi alignments of both directions: source-to-target and target-to-source using some heuristics.

2.2 Phrase Table

After alignment and phrase extraction, the phrases and associated features are stored in a phrase table. Given an aligned phrase pair: source-language phrase \(\bar{f}\) and a corresponding target-language phrase \(\bar{e}\), the most common phrase features are:

- The phrase translation probability:

\[
p(\bar{e}|\bar{f}) = \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{e}} \text{count}(\bar{e}, \bar{f})}
\] (2.3)

- The inverse phrase translation probability:

\[
p(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}} \text{count}(\bar{e}, \bar{f})}
\] (2.4)

- Additional features can include syntactic information, context information, etc.

The log linear model is very easy to extend by adding new features. After adding new features, feature weights need to be calculated using MERT which is described in the next section.

2.3 System Training: MERT

For log linear SMT translation systems, the output translation is governed by equation 2.1. In such systems, the best translation is the one that maximizes the linear combination of weighted features. Equation 2.5 shows a model using the feature functions discussed in section 2.2 in addition to a feature function representing the language model probability.

\[
\hat{e}_1^I = \arg\max_{e_1^I} \lambda_1 \log p(e_1^I|f_1^J) + \lambda_2 \log p(f_1^J|e_1^I) + \lambda_3 \log p(e_1^I) + ... + \lambda_M h_M(e_1^I, f_1^J)
\] (2.5)
where the translation and inverse translation probabilities are calculated as the multiplication of the separate phrase probabilities shown in equations 2.3 and 2.4, respectively. The third feature function is the language model probability of the output sentence.

These weights $\lambda_1^M$ can be calculated using, for example, gradient descent to maximize the likelihood of the data according to the following equation:

$$\lambda_1^M = \arg\max_{\lambda_1^M} \sum_{s=1}^{S} p_{\lambda_1^M}(e_s|f_s)$$

(2.6)

using a parallel training corpus consisting of $S$ sentence pairs. This method corresponds to maximizing the likelihood of the training data, but it does not maximize translation quality for unseen data. Therefore, Minimum Error Rate Training (MERT) is used instead [Och03]. A different objective function which takes into account translation quality by using automatic evaluation metrics such as BLEU score is used. MERT aims at optimizing the following equation, instead:

$$\lambda_1^M = \arg\min_{\lambda_1^M} \sum_{s=1}^{S} E(r_s, \hat{e}(f_s; \lambda_1^M))$$

(2.7)

where $E(r, e)$ is the result of computing a score based on an automatic evaluation metric, e.g., BLEU and $\hat{e}(f_s; \lambda_1^M)$ is the best output translation according to equation 2.1.

## 2.4 BLEU

The obvious way of evaluating the translation quality is by human evaluation. Because human evaluation is very time consuming and costly, the need for automatic evaluation heuristics arise. The most popular automatic evaluation metric is the BLEU score which is introduced by Papineni et al. [PRWZ02]. The idea behind the automatic methods is that a translation candidate is good if it is very similar to human translations. Therefore, to calculate the BLEU score of a translation output, one or more human translations are assumed to exist. BLEU score is mainly an N-gram precision metric. For the unigram score, the BLEU score is calculated as the number of words existing in the reference translations divided by the number of words in the candidate translation. In general, the N-gram precision component of the BLEU score for a specific N is calculated according to:

$$y = \frac{\text{Number of candidate n-grams found in the references}}{\text{Total number of candidate n-grams}}$$

$$= \frac{\sum_{C \in \text{Candidates}} \sum_{n-\text{gram} \in C} \sum_{R \in \text{References}} \sum_{n-\text{gram}' \in R} \mathbb{1}_{[n-\text{gram}=n-\text{gram}']}}{\sum_{C' \in \text{Candidates}} \sum_{n-\text{gram} \in C'} 1}$$

(2.8)

If BLEU score relies only on $p_n$, too short candidates will have high BLEU score. For example, a candidate translation consisting of one word that exists in one of the reference translation will have a BLEU score of 1 (the highest possible). Therefore, a brevity penalty BP is included in
BLEU score to penalize short translations and is calculated as:

$$BP = \begin{cases} 
1, & c > r \\
\frac{1 - r}{c(r - c)}, & c \leq r 
\end{cases}$$  \hspace{1cm} (2.9)$$

where $c$ is the length of the candidate translation corpus and $r$ is the sum of best matches of the references.

Finally, all N-gram scores and BP are combined to generate one BLEU score:

$$BLEU = BP \times \exp \left( \frac{1}{N} \sum_{n=1}^{N} \log p_n \right)$$  \hspace{1cm} (2.10)$$
Why is English-Arabic Machine Translation Hard?

This chapter mainly focuses on the reasons why English-Arabic translation is a very difficult task. First, an overview about Arabic morphology is given. Then, some information about Arabic inflectional agreement and its rules is provided in section 3.2. The problem of sparsity which is a consequence of the complex morphology is then discussed in section 3.3. Syntactic and lexical divergences and related problems are then explained in sections 3.4 and 3.5. Finally, a summary of the main errors as resulted from the manual error analysis is presented in section 3.6.

3.1 Arabic Morphology

Arabic morphology is complex if compared to English morphology. Similar to English, Arabic morphology has two functions: derivation and inflection, both of which are discussed in section 3.1.1. On the other hand, there are two different types of morphology, i.e., two different ways of applying changes to the stem or the baseword. These two types are the templatic morphology and the affixational morphology [Hab07a]. The functions and types of morphology are discussed in this section. As will be shown, Arabic affixational morphology is the most complex and constitutes the majority of English-Arabic translation problems.
3 Why is English-Arabic Machine Translation Hard?

3.1.1 Morphology Function

Derivational Morphology

is about creating words from other words (root or stem) while the core meaning is changed. An example from English is creating "writer" from "write". Similarly, generating كاتب "writer" from كتب "to write" is an example of derivational morphology in Arabic.

Inflectional Morphology

is about creating words from other words (root or stem) while the core meaning remains unchanged. An example from English is inflecting the verb "write" to "writes" in the case of third person singular. Another example is generating the plural of a noun, e.g., "writers" from "writer". An example in Arabic is generating آلبنات "the girls" from آلبنت "the girl".

Arabic inflectional morphology is much more complex than English inflectional morphology. English nouns are only inflected for number (singular/plural) and verbs are only inflected for number and tense (singular/plural, present/past/past-participle). English Adjectives are not inflected. On the contrary, Arabic nouns and adjectives are inflected for gender (feminine/masculine), number (singular/dual/plural), state (definite/indefinite). Arabic verbs are inflected also for gender and number besides tense(command/imperfective/perfective), voice(active/passive) and person (1/2/3).

3.1.2 Morphology Type

Templatic Morphology

In Arabic, the root morpheme consists of three, four or five consonants. Every root has an abstract meaning that’s shared by all its derivatives. According to known templates, the root is modified by adding additional vowels and consonants in a specific order to certain positions to generate different words. For example, the word كاتب "writer" is derived from the root كتب "to write" by adding alef "A" between the first and second letters of the root.
### 3.1 Arabic Morphology

This morphological type is common in most languages. It is about creating a new word from other words (roots or stems) by adding affixes: prefixes and suffixes. Affixes added to Arabic base-words are either inflectional markings or attachable clitics. Assuming inflectional markings are included in the BASEWORD, attachable clitics in Arabic follows a strict order as in: \([cnj+ [prt+ [art+ BASEWORD + pro]]]] [KH10b].

**Prefixes** include:
- \(cnj\): conjunctions such as \(w,f\) meaning and, then, respectively.
- \(prt\): some prepositions and particles such as \(b,l,k\) meaning by/with, to, as, respectively.
- \(art\): definite article \(Al\) meaning the.
- inflectional markings for tense, gender, number, person, etc.

**Suffixes** include:
- \(pro\) personal pronouns for verbs and possessive pronouns for nouns.

On the contrary, English affixational morphology is simpler because there is no clitics attached to the words. Affixational morphology in English is used for both inflection and derivation. Examples include,

- inflectional markings such as adding \(ed\) to the end of the verb to indicate past or past participle tense. Also, adding \(s\) to the end of the present form of a verb indicates it is singular. On the other hand, adding \(s\) to the end of a noun indicates that it’s plural.

- derivational morphology, for example, adding \(er\) to \(read\) generates a different word which is \(reader\). Examples of prefixes include \(mis, in\) and \(un\) for negation as in \(misrepresent, incorrect\) and \(undeniable\), respectively.

English words do not have attachable clitics like Arabic. Examples 3.1 and 3.2 shows how one Arabic word can correspond to 5 and 4 English words respectively.

(3.1)  
\[\text{wsyEtyhA} \]
\[w+ s+ yEty+ hA\]
\[\text{and will he gives her}\]
\[\text{conj prt BASEWORD prn}\]
\[\text{‘and he will give her’}\]
3 Why is English-Arabic Machine Translation Hard?

As shown above, Arabic affixational and inflectional morphology are very complex especially compared to English. The complex morphology is the main reason behind the problems of sparsity, agreement and lexical divergences as will be explained in more detail in this chapter.

3.2 Arabic Inflectional Agreement

In Arabic, there are rules that govern the inflection of words according to their relations with other words in the sentence. We call these rules throughout the work *agreement rules*. Agreement rules can involve information such as the grammatical type of the words, i.e., Part-of-Speech (POS) tags, the relationship type between the words and other specific variables. In this section, a few undetailed agreement rules are explained as examples.

**Verb-Subject**

Verbs should agree morphologically with their subjects. Verbs that follow the subject (in SVO order) agree with the subject in number and gender (see example 3.3). On the other hand, verbs that precede the subject (VSO) agree with the subject in gender only while having the singular 3rd person inflection (see example 3.4).

(3.3) الجزء: الرجال

*hbwA  AlrjAl

left.masc.pl  men.masc.pl

'The men left'

(3.4) الجزء

AlrjAl  *hb

men.masc.pl  left.masc.s

'The men left'

**Noun-Adjective**

Adjectives always follow their noun in determinism, gender and number. There are many other factors that add more rules, for example if the adjective is describing a human or an object, if the noun is plural but not in the regular plural form (i.e., broken plural), etc. Example 3.5 shows how the adjective *polite* follows the noun *sisters* in being definite, feminine and plural. This is an example where the noun refers to humans and is in the regular feminine plural form.

(3.5) الجزء

*hb  AlrjAl

men.masc.pl  left.masc.s

'The men left'
3.3 Sparsity

Sparsity is a result of Arabic complex inflectional morphology and the various attachable clitics. According to Kholy and Habash, while the number of Arabic words in a parallel corpus is 20%
3 Why is English-Arabic Machine Translation Hard?

less than English words, the number of unique Arabic words is over 50% more than the number of unique English words [KH10a].

Sparsity causes many errors in SMT output in a number of ways:

- Absence of word forms: the correct inflection of a word that agrees with the rest of the sentence could be absent from the phrase table because it was not in the training data or was infrequent and therefore was filtered.
- Poor Translation Probability Estimation: In SMT, translation probabilities are estimated through counting (refer to equation 2.3). Sparsity implies that words appear less frequently in the training data, which implies poor estimation of probabilities.
- Poor Language Model Estimation: Sparsity also causes poor estimation of language model probabilities.

3.4 Syntactic Divergences

3.4.1 Word Order

Subjects

The main sentence structure in Arabic is Verb-Subject-Object (VSO), while English sentence structure is Subject-Verb-Object (SVO). The order SVO also occurs in Arabic but is less frequent. Therefore subjects can be pre-verbal or post-verbal. Besides, subjects can be pro-dropped, i.e., subject pronouns do not need to be expressed because they are inferable from the verb conjugation [CMH10b]. Example 3.8 shows a case of a pro-dropped subject. The subject is a masculine third-person pronoun that is dropped because it can be inferred from the verb inflection.

(3.8) eěA ḍJÈ@ AltfAHp ḍklt.
the Apple he ate
’He ate the apple.’

Adjectives

In Arabic, adjectives follow the nouns that they modify as opposed to English where the adjectives precede the nouns. Example 3.9 shows the order of nouns and adjectives.

(3.9) ā
Ó B@
Al>myn Alrjl
the honest The man
’The honest man’
3.4.2 Verb-less sentences

In Arabic, verb-less sentences are nominal sentences which have no verbs. They usually exhibit the zero copula phenomenon, i.e., the noun and the predicate are joined without overt marking. These sentences are usually mapped to affirmative English sentences containing the copular verb to be in the present form. Example 3.10 shows an example of a nominal sentence.

(3.10) رايع
Algw rA}E.
wonderful The weather.
'The weather is wonderful'

One possible problem that can result from this syntactic divergence is when none of the three phrases "The weather is wonderful", "The weather is", or "is wonderful" exists in the phrase table, in which case "is" would be translated separately to an incorrect word. This results from the bad alignment of the word "is" to other Arabic words during training.

3.4.3 Possessiveness

The Arabic equivalent of possessiveness between nouns and of the of-relationship is called Idafa. The idafa construct is achieved by having the word indicating the possessed entity precede a definite form of the possessing entity. Refer to example 3.11 for an illustration of this construct.

(3.11) The child’s bag.
The bag of the child.

3.5 Lexical Divergences

Lexical divergences are the differences between the two languages at the lexical level. They result in translation problems, some of which are discussed in this section.

3.5.1 Idiomatic Expressions

Expressions are usually incorrectly translated as they are translated as separate words. Mapping each word or a few words to their corresponding meaning in Arabic usually results in a meaningless translation or at least a translation with a meaning that does not correctly match the English expression. Examples 3.12 and 3.13 illustrates the problem.
3 Why is English-Arabic Machine Translation Hard?

(3.12) Source: "brand new" Target:

\[
\text{ماركة جديدة}
\]
\[
jdydp mArkp
\]
\[
\text{new brand}
\]

'new brand’

(3.13) Source: "go all out" meaning "do your best"

Target:

\[
\text{الذهاب شاملة}
\]
\[
$Amlp Al*hAb
\]
\[
\text{totally going}
\]

'going totally’

3.5.2 Prepositions

Verbs that take prepositions cause problems in translation. Translating the verb alone to an Arabic verb and the preposition to a corresponding Arabic preposition usually results in errors. The same applies to prepositions needed by nouns. In example 3.14, although "meeting" is translated correctly to its corresponding Arabic word, the direct translation of the preposition leads to a wrong phrase.

(3.14) Source: meeting on

Target:

\[
\text{اجتماع على}
\]
\[
EIY AjtmAE
\]
\[
\text{on top of meeting}
\]

'meeting on top of’

3.5.3 Named Entities

Named entities cause a problem in translation. This is why a lot of work is being done in named entity recognition. Translating named entities word-by-word results in wrong Arabic output.
3.5.4 Ambiguity

Differences between the two languages sometimes cause translation ambiguity errors. For example, the word "just" can translate in Arabic to عادل EAdl as in "a just judge". It can also translate to فقط fqt: meaning "only". Therefore, sense disambiguation is required to achieve high quality translations.

3.5.5 Alignment Difficulties

Direct mapping of English words to Arabic words is not possible because of the lexical, morphological and grammatical differences. During alignment, this problem generates errors that are transferred to the phrase table. Some examples include:

- Auxiliaries: In English, auxiliaries can be added, for example, to express certain tenses or to express the passive voice. This is not the case in Arabic where different tenses are represented by inflecting the verbs or only by different diacritizations. This problem results in erroneous word mappings resulting from aligning an English sentence containing auxiliaries to an Arabic sentence. For example, sometimes "was" translates to في fy: meaning "in". Another example was found when "does" translates to لا IA: meaning "no". These cases result in extra prepositions which results in meaningless or ungrammatical Arabic sentences.

- Verb to be: Sentences with a present verb to be such as "The girl is nice" translates in Arabic to a nominal sentence (refer to section 3.4.2). If "is" is selected in a separate phrase, an extra incorrect word will be added to the sentence.

- Particles also usually result in extra Arabic prepositions, nouns or verbs breaking the semantic and grammatical structure of the Arabic sentence.

3.6 Error Analysis Summary

We conducted manual error analysis on a small sample of 30 sentences which were translated using a state-of-the-art phrase-based translation system similar to translate.google.com. Despite the small sample size, most errors mentioned in this chapter appeared in the output sentences. Morphological, syntactic and lexical divergences contributed to the errors. These divergences make the alignment of sentences from both languages very difficult and consequently result in problems in phrase extraction and mapping. Therefore, errors in the phrase table were very
3 Why is English-Arabic Machine Translation Hard?

Phrase table errors could directly lead to errors in the final translation output. They can result, for example, in missing or additional clitics in Arabic words and sometimes extra Arabic words. Besides, it is very common that English verbs map to Arabic nouns in the phrase table, which results in problems in the final grammatical structure of the output sentence. Ambiguity is also a phrase table problem. This is because the phrase table is based only on the surface forms not taking context into consideration. 17 sentences out of the 30 had errors because of these phrase table problems.

Morphological agreement is a major problem in the Arabic output. The main problems are the agreement of the adjective with the noun and the agreement of the verb with the subject. 9 sentences had problems with adjective-noun agreement, while 2 had problems with verb-subject agreement.

Named Entities and acronyms which were translated directly resulted in errors in 9 sentences.

In the next chapter, we introduce the previous works that has been done to overcome some of these problems. Afterwards, we explain how our work relates to and adds to previous research.
Related Work

A lot of work has been done to bridge the gap between source and target languages in SMT especially when one of the two languages is morphologically rich. Also, the greater the lexical and syntactic divergences between the two languages, the more the need for incorporating linguistic information in the translation process increases. Works that deal with divergences vary from preprocessing to post-processing to even modifying the statistical machine translation model.

Because Arabic is a polysynthetic language where every word has many attachments as explained in 3.1.2, segmentation of Arabic words is expected to improve translation from or to English which is an isolating language (i.e., each unit of meaning is represented by a separate word). Segmentation also helps the sparsity problem of morphologically rich languages. Therefore, Habash and Sadat experimented with many tokenization schemes [HS06]. They showed that Arabic tokenization through preprocessing achieved an improvement in BLEU score in Arabic to English translation especially with smaller corpora.

Most work on segmentation considered Arabic-English translation because this direction does not require post-processing to connect the Arabic segments into whole words. However, Badr et al. introduced a Rule based approach to Arabic segmentation and recombination [BZG08]. They also showed that the segmentation of Arabic helped achieving better translation quality.

Another work which emphasized on the sparsity problem of English-Arabic translation was presented by Kholy and Habash [KH10a] [KH10b]. They considered the tokenization and normalization of Arabic data to improve SMT. Similar to Badr et al., post-processing is needed to reconnect the morphemes into valid Arabic words, i.e., to detokenize and denormalize the previously tokenized and normalized surface forms. They compared three detokenization schemes which they called simple, rule-based and table-based. They also concluded that training SMT models on Arabic lemmas instead of surface forms helped translation quality.

To help with syntactic divergences, specifically word order (refer to section 3.4.1), reordering
4 Related Work

of the source sentence as preprocessing has been studied in many works. Because word order in Arabic is different from English, reordering is expected to help alignment as well as the order of words in the target sentence. Badr et al. studied a rule-based approach to reordering for English-Arabic translation [BZG09]. Also, Habash proposed improving Arabic-English SMT using pre-ordering rules [Hab07b]. Later, Carpuat et al. studied a reordering approach focused only on verb-subject constructions for Arabic-to-English SMT. They differentiated between main clauses and subordinate clauses and applied different rules for each case [CMH10b] [CMH10a]. Reordering based on Automatic learning has been also researched and has the advantage of being language independent as in [Gen10] and [XKRO09].

Most of the works mentioned above tried to decrease the divergences between the two languages through preprocessing and post-processing to make the two languages more similar. There have been also works incorporating linguistic information in the translation and language models. Gimpel and Smith explored augmenting SMT models with source-side context features [GS08]. They experimented with what they called lexical, shallow syntactic, syntactic and positional context features. Adding context features is expected to help sense disambiguation as well as other specification problems such as choosing whether a noun should be accusative, dative or genitive in German. They added the features to a log-linear translation model, then run MERT to estimate the mixture coefficients. They reported improvements in BLEU, NIST and METEOR scores for Chinese-English and English-German translations.

Another work which also considered source-side contextual features was done by Haque et al. [HNvdBW09]. In their work, they incorporated grammatical dependency relations in PB-SMT as a number of features. They achieved a 1.0 BLEU score improvement over their baseline system. More efforts towards improving target phrase selection included applying source-similarity features in PB-SMT [HNW+10]. They kept the source sentences along with the phrase pairs in the phrase table. In translating a source sentence, similarity between the source sentence and the sentences from which the phrase pair were extracted was considered as a feature in the log-linear translation model.

Language models that rely on morphological features in addition to lexical features were developed to overcome sparsity as well as inflectional agreement errors. The sparsity problem impacts SMT not only in the bilingual translation model but also in the used language model. Because Arabic is morphologically rich, i.e., most basewords are inflected by adding affixes that indicate gender, case, tense, number, etc., its vocabulary is very large. This leads to incorrect language model probability estimation because of the sparsity and the high Out-of-Vocabulary (OOV) rate. This is what motivated Sarikaya and Deng to present a Joint morphological-lexical language model (JMLLM) which combined the lexical information with the information extracted from a morphological analyzer [SD07]. Their results concluded that their model performed better than word based and morpheme based language models.

Research on predicting the correct inflection of words in morphologically rich languages has been done and evaluated on Russian and Arabic SMT outputs in [TSR08]. Their learning framework used a Maximum Entropy Markov Model \(k\)-MEMM \([MFP00]\) which is a structured prediction model where the current prediction is conditioned on the previous \(k\) predictions. They used \(k\) equals 2 and thus had a second order model where the prediction of the current word inflection was conditioned on the previous two predictions in the sentence. Their model is similar to the commonly used N-Gram language model but instead of using only the surface forms of
the words, their prediction relied on a large list of features which they categorized as monolingual and bilingual, each of which included syntactic, lexical and morphological features. They applied their model only as post-processing to machine translation output sentences.

The goal of our work is to solve some of the problems resulting from the large gap between English and Arabic. The work is divided into two parts, the first part applies changes to the statistical translation model while the second part is post-processing. In the first part, we experiment with adding syntactic features to the phrase table. The syntactic features include part-of-speech and dependency features. In contrast to the previous works which incorporated syntactic features as context features for disambiguation reasons, our work is motivated by the structural and morphological divergences between the two languages. The main two reasons behind adding these syntactic features are the complex affixation to Arabic words as well as the lexical and inflectional agreement.

In the second part of our work, we introduce a post-processing framework for fixing inflectional agreement in MT output. Our motivation is similar to that of Toutanova et al. [TSR08]. However, we focus our work on specific constructions, i.e., morphological agreement between syntactically dependent words. Their work is expected to also fix agreements but only locally because they use a left-to-right sentence decomposition instead of a top-down decomposition traversing the dependency tree of the sentence. Our framework is also a probabilistic framework which models each syntactically extracted morphological agreement relation separately. Also, the framework predicts each feature such as gender, number, etc. separately instead of predicting the surface forms, which decreases the complexity of the system and allows training with smaller corpora. The predicted features along with the lemmas are then passed to the morphology generation module which generates the correct inflections.

The next two chapters present the methods mentioned above. They discuss the methods in more detail and report the results compared to the baseline SMT system.
4 Related Work
Adding Syntactic Phrase Constraints

In this chapter, we introduce the proposed syntactic features. Section 5.1 presents the added POS features and the motivation of each of them. Afterwards, section 5.2 discusses the proposed dependency features. The features were added to phrases in a phrase-table from a state-of-the-art phrase based statistical machine translation system trained on billions of words. We conclude this chapter by discussing the results of a number of experiments and the limitations of this approach.

5.1 POS Features

In general, most POS features are added to penalize the incorrectly mapped phrase pairs. The English part of these phrase pairs usually does not have a corresponding Arabic translation (refer to section 3.5.5 for examples). Therefore, it is usually paired with incorrect Arabic phrases. We add the POS features to discourage these phrase pairs from being selected by the decoder. These features mark phrases that consist of one or more of personal and possessive pronouns, prepositions, determiners, particles and wh-words. The POS features used in our experiments are summarized in table 5.1. After adding the features, MERT is used to calculate their weights.
5 Adding Syntactic Phrase Constraints

<table>
<thead>
<tr>
<th>POS</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRP</td>
<td>Personal Pronouns: subject pronouns and object pronouns</td>
</tr>
<tr>
<td>PRP$</td>
<td>Possessive Pronouns</td>
</tr>
<tr>
<td>DT</td>
<td>Determiners: a, the, this, etc.</td>
</tr>
<tr>
<td>IN</td>
<td>Prepositions</td>
</tr>
<tr>
<td>RP</td>
<td>Particles</td>
</tr>
<tr>
<td>WDT</td>
<td>Wh-determiner: what, which</td>
</tr>
<tr>
<td>WP</td>
<td>Wh-pronoun: who, whether, which (head of a wh- noun phrase)</td>
</tr>
<tr>
<td>WRB</td>
<td>Possessive Wh-pronoun: whose</td>
</tr>
</tbody>
</table>

Table 5.1: Word classes for connectable phrases

5.1.1 Personal Pronouns

Personal pronouns in Arabic can be separate or attached. Similar to English, there are subject pronouns and object pronouns. In addition to the singular and plural pronouns, Arabic has also dual pronouns. Personal pronouns can attach to the end of verbs (subject or object) and to the end of prepositions. Figure 5.1 shows a diagram of the Arabic pronouns. It shows that pronouns are divided into separate and attached. Separate pronouns are the subject pronouns. On the other hand, pronouns could be attached to verbs, prepositions or nouns. Pronouns attached to verbs are either subject or object pronouns. Pronouns can attach also to prepositions or nouns. All tree leaves represent personal pronouns except for the pronouns attached to nouns which are possessive pronouns. The only leaf corresponding to possessive pronouns is highlighted. There are two reasons why personal pronouns should be penalized as separate phrases:

- Subject pronouns are the only separate pronouns. They are uncommon because pronominal subjects can also be attached or dropped.
- If a separate pronominal subject is generated in the target sentence, selecting a phrase containing the pronoun and verb together guarantee that they agree in gender, number, etc.
5.1.2 Possessive Pronouns

As shown in figure 5.1, possessive pronouns are always attached to the end of nouns. Example:

(5.1) *Her house* (منزلها)  
منزل ها  
+ha mnzl

Because possessive pronouns in English are separate words, there are entries for them in the phrase table. These entries usually map to Arabic words with different meanings. Table 5.2 shows some phrase table entries to show what are they mapped to in Arabic. Sometimes, these phrases are selected by the decoder, which usually results in erroneous translations. Therefore, penalizing those phrases should prevent them from being selected.
5 Adding Syntactic Phrase Constraints

<table>
<thead>
<tr>
<th>her</th>
<th>لَهَا</th>
<th>وَقَالَتْ</th>
<th>وَقَالَتْ إِنْ</th>
<th>صاحبة</th>
<th>لأطفالها</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAHbp</td>
<td>lhA</td>
<td>wqAlt</td>
<td>wqAlt</td>
<td>owner</td>
<td>! &gt; tfAlha</td>
</tr>
<tr>
<td>owner</td>
<td>'for her'</td>
<td>'and she said'</td>
<td>'and she said that'</td>
<td>'for her children'</td>
<td>'for her children'</td>
</tr>
</tbody>
</table>

his

<table>
<thead>
<tr>
<th>له</th>
<th>تَجْرِيه</th>
<th>تَقريره</th>
<th>بلده</th>
</tr>
</thead>
<tbody>
<tr>
<td>lp</td>
<td>tEryfp</td>
<td>tqryrp</td>
<td>bldp</td>
</tr>
<tr>
<td>'for him'</td>
<td>'his definition'</td>
<td>'his report'</td>
<td>'his country'</td>
</tr>
</tbody>
</table>

Table 5.2: Example of phrase table entries for possessive pronouns

5.1.3 Prepositions and Particles

In Arabic, there are attached and separate prepositions. Prepositions were discussed in section 3.5.2 as an example of lexical divergences. Translating prepositions separately can be harmful because sometimes they should be attached to Arabic words and sometimes context is needed in order to select the correct preposition. Figure 5.2 shows an example of an attached preposition. On the other hand, figure 5.3 shows an example for a separate preposition which was translated to a wrong preposition when translated as a separate phrase.

Figure 5.2: Wrong translation of attached prepositions
Therefore, selecting phrases containing only prepositions should be avoided. By adding a feature to mark these phrases, the feature is expected to get a negative weight and therefore penalized compared to other available phrases.

Particles when translated separately usually result in additional Arabic words because phrasal verbs including the verb and preposition can only map to an Arabic verb.

### 5.1.4 Determiners

The determiner class (DT) in English includes, in addition to other words, the definite and indefinite articles "the" and "a or an", respectively. In Arabic, the definite article corresponds to an "Al" attached as a prefix to the noun. There is no indefinite article in Arabic. Having them in separate phrases only introduces noise. Table 5.3 shows their entries in the phrase table. As shown, they correspond to prepositions which is very harmful to the adequacy and fluency of the output sentence.

---

**Figure 5.3: Wrong translation of separate prepositions**


5 Adding Syntactic Phrase Constraints

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>علي</td>
<td>في</td>
<td>من</td>
</tr>
<tr>
<td></td>
<td>Ely</td>
<td>fy</td>
<td>mn</td>
</tr>
<tr>
<td></td>
<td>on</td>
<td>in</td>
<td>from</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>في</td>
<td>علي</td>
<td>من</td>
</tr>
<tr>
<td></td>
<td>fy</td>
<td>Ely</td>
<td>mn</td>
</tr>
<tr>
<td></td>
<td>in</td>
<td>on</td>
<td>from</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 5.3: Example of phrase table entries for determiners: a, the*

5.1.5 Wh-nouns

Wh-nouns include wh-determiners, wh-pronouns and possessive wh-pronouns having the POS tags WDT, WP and WRB, respectively. We add features for these POS tags to discourage selecting separate phrases which contains only wh-nouns. The motivation for this is mainly gender and number agreement. When they are attached in one phrase with the word they refer to, they would probably be translated in the correct form.

5.2 Dependency Features

These features are based on the syntactic dependency parse tree of the English source sentence. They mark the phrases which has the two words of a specific set of relations in the phrase. For example, a feature amod(adjective modifier) is added to a phrase which contains both the adjective and the noun. These features are expected to get positive weights when trained by MERT and thus make the decoder favor these phrases over others. The suggested dependency features are summarized in table 5.4. The relations’ names follow the Stanford typed dependencies [dMM08].

The motivation behind these dependency features is mainly agreement: morphological or lexical. Assume that \( \alpha_1 \) and \( \alpha_2 \) are Arabic words that should have morphological agreement. Because phrases are extracted from the training data which are assumed to be morphologically correct, using a phrase that contains \( \alpha_1 \) and \( \alpha_2 \) assures that they agree.

As explained in 3.2, interesting morphological agreement relations include for example, noun-adjective and verb-subject relations. Lexical agreement relations include, for example, relations between phrasal verbs and their prepositions. For example, "talk about" is correct while "talk on" is not. Selecting a phrase where "talk" and its preposition "about" are attached guarantees their agreement.

Some of the dependency features are also motivated by the alignment problems discussed in
These problems arise from trying to align English sentences containing words that have no corresponding separate Arabic words in the Arabic sentences. For example, the *acomp* relation should favor selecting the phrase "is beautiful" over selecting the two separate phrase "is" and "beautiful". This is because the phrase "is" would translate to an incorrect Arabic word. Also, *aux* is motivated by the same reason, because most auxiliaries have no corresponding words in Arabic.

Relations *amod, nsubj, num, ref* and *conj* are all motivated by inflectional agreement. The relation *nsubj* is also specifically useful if the subject is a pronoun, in which case it will be most of the time omitted in the Arabic and only help in generating the correct verb inflection.

Adding *det* is beneficial in two ways. First, to discourage selecting a phrase with a separate "the" which would result in a wrong Arabic translation as shown in table 5.3. Second, attaching the determiner to its noun causes the Arabic word to have the correct form whether to have "Al" or not if the determiner in English is "the" or "a", respectively.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>aux</td>
<td>Auxiliary</td>
<td>Sam has died: r(died, has)</td>
</tr>
<tr>
<td>det</td>
<td>Determiner</td>
<td>The wall is high: r(wall, the)</td>
</tr>
<tr>
<td>conj</td>
<td>Conjunct</td>
<td>Sam is nice and honest: r(nice, honest)</td>
</tr>
<tr>
<td>acomp</td>
<td>Adjectival Complement</td>
<td>She is beautiful r(is, beautiful)</td>
</tr>
<tr>
<td>amod</td>
<td>Adjectival Modifier</td>
<td>Sam eats red meat: r(meat, red)</td>
</tr>
<tr>
<td>nsubj</td>
<td>Nominal Subject</td>
<td>Sam left: r(left, Sam)</td>
</tr>
<tr>
<td>num</td>
<td>Numeric Modifier</td>
<td>I ate 3 apples: r(apples, 3)</td>
</tr>
<tr>
<td>ref</td>
<td>Referent</td>
<td>I saw the book which you bought: r(book, which)</td>
</tr>
</tbody>
</table>

*Table 5.4: Dependency Relations Used as Features*

### 5.3 Results

We run our experiments on a state-of-the-art PBSMT system trained on standard parallel corpora such as the translated Arabic Treebank\(^1\) in addition to web data. The features are added to the phrases and stored in the phrase table during training. Afterwards, MERT is used to tune the feature weights. Finally, the system output is evaluated on two different data sets. WEB includes random sentences from the web and WIKI includes news articles from WikiNews\(^2\). Table 5.5 shows the number of sentences in each test set.

\(^1\)http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T07

\(^2\)http://en.wikinews.org
5 Adding Syntactic Phrase Constraints

<table>
<thead>
<tr>
<th>Test Data</th>
<th>Number of Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEB</td>
<td>6551</td>
</tr>
<tr>
<td>WIKI</td>
<td>510</td>
</tr>
</tbody>
</table>

*Table 5.5: Test Data*

We run and compare five different experiments as shown in table 5.6. BL is the baseline system without any added features. The baseline system does not include any segmentation or preordering. POSS is an experiment where only POS features were added to phrases of length one. POSM has POS features added to single- and multi-word phrases as long as all the words in the phrase have POS tags of the set listed in table 5.1. DEP is an experiment with only dependency features added to any phrase which contain the two words of the relation in the phrase. Finally, ALL includes the features added in POSS, POSM and DEP. Table 5.6 include the BLEU scores of both test sets for all five experiments.

As shown in the table, the experiment ALL outperforms the rest of the experiments. However, for the WEB data set, the POSS has a higher BLEU score by only 0.01. The baseline has the lowest BLEU score compared to the four experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>WEB</th>
<th>WIKI</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL</td>
<td>11.27</td>
<td>15.21</td>
</tr>
<tr>
<td>POSS</td>
<td><strong>11.47</strong></td>
<td>15.73</td>
</tr>
<tr>
<td>POSM</td>
<td>11.39</td>
<td>15.37</td>
</tr>
<tr>
<td>DEP</td>
<td>11.38</td>
<td>15.61</td>
</tr>
<tr>
<td>ALL</td>
<td><strong>11.46</strong></td>
<td><strong>15.81</strong></td>
</tr>
</tbody>
</table>

*Table 5.6: BLEU Scores*

The weights calculated by MERT are shown in tables 5.7 and 5.8. POS features are expected to be negative to penalize the phrases which have them. As shown in table 5.7, the resulting weights are noisy and are not all negative. On the other hand, dependency features are expected to have positive weights to favor the phrases which have them. As shown in table 5.8, although most of the weights are positive, some features have negative weights. In the next section, the limitations of this approach and possible improvements are discussed.
5.4 Limitations

Although the added features achieve an improvement in BLEU score, the improvement is slight. The reasons behind having small effect on BLEU score could be:

- Features are only added to a small number of phrases. Therefore, their effect on decoding is expected to be small. For example, in POSS, only 0.17% of the phrases are affected, while in POSM, only 0.5% are affected.

- MERT is limited to optimizing a small number of feature weights. For a large number of sparse features, MERT is unreliable [CMR08].

- MERT optimizes BLEU score which does not capture syntactic or morphological constraints well (refer to section 6.2.1 for more details).

Overcoming the MERT training problems is important for improving this approach. Setting the weights manually to positive or negative weights according to our linguistic knowledge is a bad idea because it would hurt the translation quality in general. One possible improvement is to use an alternative training algorithm which can optimize the weights of a large number of sparse features more reliably, similarly to Chiang et al. [CMR08].

In the next chapter, we introduce our post-processing system with details about the system pipeline and the used algorithms.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>DT</th>
<th>IN</th>
<th>PRP</th>
<th>PRP$</th>
<th>RP</th>
<th>WDT</th>
<th>WP</th>
<th>WRB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>-0.91</td>
<td>-1.25</td>
<td>0.26</td>
<td>1.29</td>
<td>0.03</td>
<td>-0.55</td>
<td>-7.45</td>
<td>-</td>
</tr>
<tr>
<td>POSS</td>
<td>-</td>
<td>0.11</td>
<td>0.65</td>
<td>-</td>
<td>2.17</td>
<td>-</td>
<td>0.97</td>
<td>3.60</td>
</tr>
<tr>
<td>POSM</td>
<td>-0.77</td>
<td>-0.86</td>
<td>-</td>
<td>-2.22</td>
<td>-13.43</td>
<td>0.13</td>
<td>1.20</td>
<td>-</td>
</tr>
</tbody>
</table>

*Table 5.7: POS Feature Weights As Evaluated by MERT*

<table>
<thead>
<tr>
<th>Experiment</th>
<th>acomp</th>
<th>amod</th>
<th>aux</th>
<th>conj</th>
<th>det</th>
<th>nsubj</th>
<th>num</th>
<th>ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>-</td>
<td>-1.04</td>
<td>0.36</td>
<td>-0.52</td>
<td>-</td>
<td>0.59</td>
<td>2.19</td>
<td>3.73</td>
</tr>
<tr>
<td>DEP</td>
<td>.90</td>
<td>-</td>
<td>0.44</td>
<td>-0.35</td>
<td>-</td>
<td>0.54</td>
<td>-</td>
<td>0.54</td>
</tr>
</tbody>
</table>

*Table 5.8: Dependency Features Weights As Evaluated by MERT*
5 Adding Syntactic Phrase Constraints
Fixing Inflectional Agreement through Post-processing

In this chapter, we introduce our post-processing framework. The goal of this system is to fix inflectional agreement between syntactically related words in machine translation output. We introduce the pipeline of the system in section 6.1. In the following section, our algorithm is compared to other possible algorithms. Afterwards, the main modules of the system will be explained in more detail. Finally, the evaluation methods, the experimental settings and the results are discussed.

6.1 System Description

Our post-processor is based on a learning framework. Given a training corpus of aligned sentence pairs, the system is trained to predict inflections of words in MT output sentences. The system uses a number of multi-class classifiers, one for each feature. For example, there is a separate classifier for gender and a separate classifier for number. Figure 6.1 shows the system pipeline.

In the training phase, the reference aligned parallel corpus is used. The morphology analyzer analyzes the Arabic sentences. It specifies the lemma, the part-of-speech(POS) tag and the morphological features (e.g., gender, number, person, etc.) for every word in the sentence. The syntax projector is the module which projects the dependency relations from the English sentence to the Arabic sentence using the alignments and the POS tags of both sentences. Subsequently, it extracts the agreement relations using the projected syntax. The feature vector extractor is responsible for extracting the feature vectors out of the available lexical, morpho-
6 Fixing Inflectional Agreement through Post-processing

Figure 6.1: System Diagram

Logical and syntactic information. The feature vectors as well as the correct labels which are extracted from the reference data are then used to train the classifiers.

For the prediction of the correct features, the MT translation output as well as the source sentence and the alignments are required. The data goes through the same steps as in the training phase. The extracted feature vectors are then used to make predictions for each feature separately.

The correct features are then used along with the lemmas of the words to generate the correct inflected word by the morphology generator. Finally, the LM filter uses an N-gram language model to add some robustness to the system against the errors of alignment, morphology analysis and generation and classification. If the generated sentence has a lower LM score than the baseline sentence, the baseline sentence is not updated.

6.1.1 Algorithms for Inflection Prediction

As mentioned above, we propose a system that predicts the correct inflections of specific words, i.e., words whose inflection is governed by agreement relations. We train a number of separate multi-class classifier, one for each morphological feature. In this section, we compare our algorithm to other possible algorithms.
6.1 System Description

Manual Rules

The way certain parts of a sentence should be inflected in correspondence to the inflection of other parts, e.g., the inflection of a verb based on its subject inflection or the the inflection of an adjective based on the noun can be encoded in a finite set of rules. However, such rules can be very difficult to enumerate. The rules could differ from a "part of speech" to another and from one language to another. The difficulty of writing manual rules also arise from the existence of exceptional cases to all rules. Therefore, taking this approach not only requires writing all set of POS and language dependent rules, but also requires handling all the special cases.

For example, consider the inflection of an adjective in agreement with the modified noun.

The general rule: an adjective should follow the noun in gender, number, case and state.

Some Exceptions:

- If the noun is a broken plural representing objects (no persons), the adjective should be feminine and singular no matter what the gender of the noun is.
- If the noun is a broken plural representing persons (only masculine), the adjective could be in a broken plural or a regular plural form.
- If the noun is a feminine plural representing objects, the adjective can be and is preferred to be singular.

Therefore, a learning approach which could be easily extended to different agreement relations and different languages is preferable.

Probabilistic Models

If all the dimensions affecting the correct word inflection could be encoded in a feature vector, many state-of-the-art probabilistic approaches can be used to predict the correct inflections. For example, Minkov et al. used a structured probabilistic model based on sentence order decomposition [MT07]. Their system has limitations in modeling agreement because their probabilistic model does not use the dependencies effectively. Although the prediction of a word inflection strongly depend on the inflection of the parent of the agreement relation, the feature vector in their system includes only the stem of the parent.

A tree-based structured probabilistic model such as k-MEMM or CRF that use the dependency tree is theoretically very effective. However, dependency trees for Arabic sentences are of poor quality and would result in a very noisy model that might degrade the MT output quality.

In this work, predicting the inflection of each word according to its agreement relation separately shows to be very effective. As will be explained in section 6.1.3, the relations are independent, for example, fixing the inflection of the adjective in an adjective-noun agreement relation is independent from fixing the inflection of the verb in a verb-subject agreement. Therefore, separating prediction adds robustness to the system and allows training with smaller corpora.
6 Fixing Inflectional Agreement through Post-processing

6.1.2 Arabic Analysis and Generation

For Arabic analysis, we use the Morphological Analysis and Disambiguation for Arabic (MADA) system developed by Habash and Rambow [HR05]. The system is built on the Buckwalter analyzer which generates multiple analyses for every word [Maa10]. MADA uses another analyzer and generator tool ALMORGEANA to convert the output of Buckwalter from a stem-affix format to a lexeme-and-feature format [Hab07a]. Afterwards, it uses an implementation of support vector machines which includes Viterbi encoding to disambiguate the results of ALMORGEANA analyses. The result is a list of morphological features of every word taking the context (neighboring words in the sentence) into consideration. The morphological features that are evaluated by MADA are illustrated in table 6.1. The last four rows of the table represent the attachable clitics whose positions in the word are governed by [prc3 [prc2 [prc1 [prc0 BASE-WORD enc0] ] ] ]. For more details about those clitics and their functions, the user is referred to the MADA+TOKAN Manual [HRR10]. In addition to the features listed in the table, the analysis output includes the diacrititzed form (diac), the lexeme/lemma (lex), the Buckwalter tag (bw) and the English gloss (gloss).

For generation, the lexeme, POS tag and all other known features from table 6.1 are input to the MORGEANA tool. The system searches in the lexicon for the word which has the most similar analysis. The analysis and generation tools where used in this work to change the declension of a word. For example, to change a word $w$ from the feminine to the masculine form, the following steps are taken:

- Input to MADA the surface form of the word $w$. MADA will output the best lexeme $l$ and list of features $f$ for this word.

- Change the gender feature in $f$ from f(eminine) to m(asculine): $f[\text{gen}] = m$. We call the modified feature list $\bar{f}$.

- Input to the MORGEANA generator the lexeme $l$ and the modified feature list $\bar{f}$. MORGEANA will generate the word $\bar{w}$ which is the masculine surface form of $w$. 


<table>
<thead>
<tr>
<th>Label</th>
<th>Name</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>pos</td>
<td>part of speech</td>
<td>verb, noun, (adjective, prep(osition), part(icle) and others (total: 34)</td>
</tr>
<tr>
<td>asp</td>
<td>aspect</td>
<td>c(ommand), i(mperfective), p(erfective), na(not applicable)</td>
</tr>
<tr>
<td>cas</td>
<td>case</td>
<td>n(ominative), a(ccusative), g(entitive), u(ndefined), na</td>
</tr>
<tr>
<td>gen</td>
<td>gender</td>
<td>f(eminine), m(asculine), na</td>
</tr>
<tr>
<td>mod</td>
<td>mood</td>
<td>i(ndicative), j(ussive), s(subjunctive), u(ndefined), na</td>
</tr>
<tr>
<td>num</td>
<td>number</td>
<td>s(ingular), p(lural), d(ual), u(ndefined), na</td>
</tr>
<tr>
<td>per</td>
<td>person</td>
<td>1, 2, 3, na</td>
</tr>
<tr>
<td>stt</td>
<td>state</td>
<td>i(ndefinite), d(efinite), c(onstruct/possesive/idafa), u(ndefined), na</td>
</tr>
<tr>
<td>vox</td>
<td>voice</td>
<td>a(ctive), p(assive), u(ndefined), na</td>
</tr>
<tr>
<td>prc0</td>
<td>proclitic level 0</td>
<td>0, na, &gt;a_ques (interrogative particle &gt;a)</td>
</tr>
<tr>
<td>prc1</td>
<td>proclitic level 1</td>
<td>0, na, fa, wa</td>
</tr>
<tr>
<td>prc2</td>
<td>proclitic level 2</td>
<td>bi, ka, la, li, sa, ta, wa, fi, IA, mA, yA, wA, hA</td>
</tr>
<tr>
<td>enc0</td>
<td>enclitic</td>
<td>0, na, pronouns, possessive pronouns and other particles</td>
</tr>
</tbody>
</table>

Table 6.1: Morphological Features Resulting from MADA analysis

6.1.3 Syntax Projection and Relation Extraction

To extract the morphologically dependent pairs (agreement pairs), syntax relations are needed. Although Arabic dependency tree parsers exist, for example the Berkeley and Stanford parsers which are compared in [GM10], they have poor quality. We benefit from the parallel aligned data to project the syntax tree from English to Arabic. The English parse tree is a labeled dependency tree following the grammatical representations explained in [dMM08]. Two approaches to projection were considered.

Direct Projection

Given a source sentence consisting of a set of tokens $s_i...s_n$, a dependency relation is a function $h_s$ such that for any $s_i$, $h_s(i)$ is the head of $s_i$ and $l_s(i)$ is the label of the relation between $s_i$ and $h_s(i)$.

Given an aligned target sentence $t_j...t_m$, $A$ is a set of pairs $(s_i, t_j)$ such that $s_i$ is aligned to $t_j$. Similarly, $h_t(j)$ is the head of $t_j$ and $l_t(j)$ is the label of the relation between $t_j$ and $h_t(j)$.

Similar to the unlabeled tree projection suggested by Quirk et al. [QMC05], projection can be done according to the following rule:

$$h_t(i) = j \Leftrightarrow \exists(s_m, t_i), (s_n, t_j) \in A \text{ such that } h_s(m) = n \quad (6.1)$$
Labels can be also projected using:

\[ l_i(i) = x \iff \exists (s_m, t_i), (s_n, t_j) \in A \text{ such that } l_s(m) = x \quad (6.2) \]

Although this approach is helpful for identifying some Arabic dependency relations, it has a number of limitations.

- Errors in the English parse tree are also projected to the Arabic parse tree.
- Many-to-many alignments introduce ambiguities that are difficult to resolve.
- The algorithm projects a dependency link as long as two pairs of words are aligned. Therefore, alignment errors result in projection errors. Also, the algorithm does not take into consideration the difference in structure between the two languages. For example, an Arabic noun might align to an English verb. In this case, the Arabic sentence can have a relation of "nsubj" with a noun head.

Figure 6.2 shows an example of the direct projection approach. In this example, the alignment between source and target sentences is one-to-one, therefore, there is no ambiguity problem. However, as can be seen, errors by the English parser were projected to the Arabic parse tree.

Because of the above limitations, we propose a different approach to partial tree projection. Our approach makes use of the Arabic analysis for robustness. It also takes syntactic divergences between the two languages into account.

Figure 6.2: Example of Dependency Tree Projection
Our approach

The end goal of the dependency tree projection in our work is the extraction of dependencies between pairs of words that should have morphological agreement, i.e., agreement links. Therefore, there is no need to first project the English tree to an Arabic tree from which we can extract agreement links. Obviously, the extra step would introduce more errors. Therefore, we extract agreement links directly using the lexical and syntactic information of both the English and Arabic sentences taking into consideration the typological differences between the two languages. The projection of some of the interesting relations is explained in this section.

ADJECTIVE RELATION (amod)

For an amod relation, an Arabic agreement relation is extracted if the English adjective aligns to an Arabic adjective, while the English noun aligns to an Arabic noun.

In the case when the English word aligns to multiple Arabic words, selecting the noun for the amod relation is based on the heuristic that the first noun after a preposition is marked as the noun of the relation. The motivation behind this rule is illustrated by example 6.3. If the first word of the multiple word alignment was selected as the noun of the relation, a link amod(AlfwtwgrAfy, mwad) would be extracted although amod(AlfwtogrAfy, IltSwyr) is the correct link. Therefore, linguistic analysis of erroneous agreement links lead to the mentioned rule.

Example 6.4 illustrates the ambiguity problem in a case of one-to-many alignment from English to Arabic. The word airline aligns to two Arabic words. We select the first word as the word being described by the adjective. However, in some cases, this rule introduces error.

Example 6.4 illustrates the ambiguity problem in a case of one-to-many alignment from English to Arabic. The word airline aligns to two Arabic words. We select the first word as the word being described by the adjective. However, in some cases, this rule introduces error.

(6.3) Photographic chemicals

Example 6.4 illustrates the ambiguity problem in a case of one-to-many alignment from English to Arabic. The word airline aligns to two Arabic words. We select the first word as the word being described by the adjective. However, in some cases, this rule introduces error.

(6.4) Saudi airlines

Example 6.4 illustrates the ambiguity problem in a case of one-to-many alignment from English to Arabic. The word airline aligns to two Arabic words. We select the first word as the word being described by the adjective. However, in some cases, this rule introduces error.

(6.4) Saudi airlines

Example 6.4 illustrates the ambiguity problem in a case of one-to-many alignment from English to Arabic. The word airline aligns to two Arabic words. We select the first word as the word being described by the adjective. However, in some cases, this rule introduces error.
6 Fixing Inflectional Agreement through Post-processing

the English syntactic dependency parse tree indirectly as shown in figure 6.3. The predicate *Ityf* has an agreement relation with the subject of the sentence *Alwld*.

![Figure 6.3: acomp relation](image)

**VERB-SUBJECT RELATION**
The agreement link that we are interested in is the link from the verb to the subject, which is only the reverse of the link in the English dependency parse tree.

**RELATIVE WORDS (REF)**
One way to extract the noun to which a relative word refers is through the English dependency parse tree. Figure 6.4 shows the projection of the ref relation.

![Figure 6.4: acomp relation](image)

### 6.1.4 Feature Vector Extraction

The feature extractor as shown in the framework diagram in figure 6.1 follows the morphology analysis and syntax projection. The selected features are from many sources of information: lexical, morphological and syntactic. Table 6.2 summarizes the features used in the feature vectors used by the classifiers.

For Arabic Features, morphological features include the features returned by the morphology analyzer. After analysis of classification errors, we noticed regular errors caused by the morphology analyzer, for which we added the extra two features: feminine ending and number of English gloss. Although all nouns and adjectives in Arabic that have particular endings, namely,
"ء"، "ؤ"، and "ي" are feminine, the analyzer confuses them frequently. The analyzer does not make use of these endings since it does not analyze the surface forms of the words; however, it uses prefix, suffix and stem lexicons to generate the features. Incorrect or missing labels to these words in the used lexicons would result in these errors. By adding this feature, we achieve a significant improvement in classification accuracy.

The number of the English gloss is also added to overcome the persistent error of the analyzer to analyze broken plurals as singular. The reason is because it actually identifies a plural by whether the stem is attached to a clitic for plural marking. However, in the case of broken plurals, there is no affix added to the stem, however, it is derived from the singular form, a case of derivational morphology. As a solution to this problem, a feature is added to indicate whether any of the English glosses of the word is plural or not.

The feature "Plural Type" is added because it significantly affects the decision about the correct inflection. For example, a regular masculine plural noun has its modifying adjective in masculine plural form, while an irregular plural noun usually has its modifying adjective in feminine singular form.

Syntactic features for Arabic include part of speech tags of the current and head words. Lexical features include the stem of the head word and the English gloss.

English features only include the part of speech tags and the surface forms of the aligned and head words. On the other hand, general features include the dependency relation type and whether the head comes before or after the current word in the sentence. The latter feature is useful for example in the case of verbs where the verb inflection rules are different for the SVO order versus the VSO order.

<table>
<thead>
<tr>
<th>Arabic Features</th>
<th>Morphological</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>asp, cas, gen, mod, num, per, stt, vox, prc0, prc1, prc2, enc0 (refer to table 6.1)</td>
</tr>
<tr>
<td></td>
<td>Feminine Ending: yes, no</td>
</tr>
<tr>
<td></td>
<td>Gloss Number: singular or plural</td>
</tr>
<tr>
<td></td>
<td>Plural Type: regular, irregular (broken plurals)</td>
</tr>
<tr>
<td></td>
<td>part-of-speech</td>
</tr>
<tr>
<td></td>
<td>stem(head), English gloss(head)</td>
</tr>
<tr>
<td>Syntactic</td>
<td></td>
</tr>
<tr>
<td>Lexical</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>English Features</td>
<td>Syntactic</td>
</tr>
<tr>
<td></td>
<td>part-of-speech (aligned, head)</td>
</tr>
<tr>
<td></td>
<td>surface form(aligned, head)</td>
</tr>
<tr>
<td>General</td>
<td>Relation type</td>
</tr>
<tr>
<td></td>
<td>amod, verb, acomp, etc.</td>
</tr>
<tr>
<td></td>
<td>Head position</td>
</tr>
<tr>
<td></td>
<td>before, after</td>
</tr>
</tbody>
</table>

Table 6.2: The features used by the classifiers
6 Fixing Inflectional Agreement through Post-processing

6.1.5 Training and Classification

To train the classifiers, the extracted feature vectors as well as the correct labels are needed. The data used for training is a set of parallel aligned sentences. The agreement relations are extracted according to 6.1.3. For every relation pair, prediction is done for one word given the inflection of the parent and other bilingual and lexical information that are encoded in the feature vectors. The result of the morphological analyzer for a specific feature is then used as the label for this feature classifier. As mentioned earlier, the accuracy of the analysis can vary between 95.5% and 99.9% according to [HR05]. Erroneous labels not only result in noisy training data, but also in imprecise classification accuracy evaluation.

For training the classifiers, we use an automatic tool for selecting the best classification model for each feature and also for selecting the best parameters for this model using cross-validation. The reported accuracy is actually the mean accuracy of the folds of the cross validation.

In classification, after agreement relations and then feature vectors are extracted, prediction is done separately for each feature using the corresponding classifier.

6.1.6 Language Model Incorporation

An N-Gram Language model is a probabilistic model specifically a model which predicts the next word in a sentence given the previous N words based on the Markov assumption [JM00]. The N-Gram language model probability of a sentence is approximated as a multiplication of N-Gram probabilities as shown in the second part of equation 6.5.

\[
P(w_1, ..., w_m) = \prod_{i=1}^{m} P(w_i|w_1, ..., w_{i-1}) \approx \prod_{i=1}^{m} P(w_i|w_{i-(n-1)}, ..., w_{i-1}) \tag{6.5}
\]

\[
P(w_1, ..., w_m) = \prod_{i=1}^{m} P(w_i|w_{i-1}, w_{i-2}) \tag{6.6}
\]

In this work, we use the language model probability as an indicator to the correctness and fluency of a modification. We compare \(P(\text{output sentence})\) and \(P(\text{post-processed sentence})\). If the post-processed sentence has much less probability (i.e., less by a difference more than a certain threshold) than the output translation, changes to the sentence are canceled. Change filtering using a language model is expected to provide some robustness against all sources of errors in the system. However, the language model is not very reliable. A simple example would be that the generated inflected word is out of vocabulary (OOV) for the language model, although it is morphologically the correct one.

6.2 Evaluation

To evaluate the system performance, we run three sets of evaluation experiments. First, accuracy of the classifiers is evaluated and compared to two other prediction algorithms. Prediction accuracy does not, however, measure the performance of the whole system. To evaluate the final
output of the system, BLEU score is used. The BLEU score of the output is compared to BLEU score of the baseline MT system output. Because of the BLEU score limitations in evaluating morphological agreement, human evaluation experiments are also used. In the section, the BLEU score limitations is explained in 6.2.1 and a brief description about the human evaluation is provided in 6.2.2.

6.2.1 BLEU

BLEU proved to be unreliable for evaluating morphological agreement in this work because of the following:

- In the evaluation data, every sentence has only one reference human translation.
- Because BLEU is based on merely counting which words (inflected surface forms) exist in the reference translations, two problems arise:
  - There are cases where the updated words do not exist at all in the reference translation. The example in 6.3 shows a sentence whose agreement problem was fixed but this did not result in any change in BLEU score. Although the gender was corrected to be masculine in both words, this resulted in zero difference in BLEU score because the reference translation contained a synonym "Almd Ary" of the corrected word and not the word itself "Al<stwA’y".
  - There are cases where the original word inflection scored higher in BLEU score because the original word simply existed in the reference translation, although the agreement was wrong. The example in table 6.4 shows how the SMT output sentence received a higher score than the post-processed one although the agreement was corrected and the whole sentence is grammatically and morphologically correct. The words in bold local governments can be noticed to disagree in definiteness in the translation output because local is definite and governments is indefinite. Although they were corrected to be both indefinite in the post-processed sentence, the post-processed sentence received a lower BLEU score. The reason is that the reference translation contains the definite form of the word governments and the word local. It is known that BLEU score is based on counting the words in the candidate translation whose surface forms exist in the reference translations. After correcting the agreement in post-processing, both the indefinite words local and governments were considered to be absent in the reference translation and thus BLEU score decreased.
6 Fixing Inflectional Agreement through Post-processing

<table>
<thead>
<tr>
<th>Source</th>
<th>Tropical Weather</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Output</strong></td>
<td></td>
</tr>
<tr>
<td>الطقس الدارية</td>
<td></td>
</tr>
<tr>
<td>AlmdAryp AlTqs</td>
<td></td>
</tr>
<tr>
<td>tropical.def.fem weather.def.masc</td>
<td></td>
</tr>
<tr>
<td><strong>Post Processed</strong></td>
<td></td>
</tr>
<tr>
<td>الطقس الداري</td>
<td></td>
</tr>
<tr>
<td>AlmdAry AlTqs</td>
<td></td>
</tr>
<tr>
<td>tropical.def.masc weather.def.masc</td>
<td></td>
</tr>
<tr>
<td><strong>Reference</strong></td>
<td></td>
</tr>
<tr>
<td>الطقس الاستوائي</td>
<td></td>
</tr>
<tr>
<td>Al&lt;stwA’y AlTqs</td>
<td></td>
</tr>
<tr>
<td>tropical.def.masc weather.def.masc</td>
<td></td>
</tr>
<tr>
<td>'The Tropical Weather’</td>
<td></td>
</tr>
</tbody>
</table>

*Table 6.3: Example about invalidity of BLEU*
6.2 Evaluation

<table>
<thead>
<tr>
<th>Source</th>
<th>Farms which are governed by local governments.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td></td>
</tr>
<tr>
<td>المزارع التي تحكمهاحكومات المحلية</td>
<td>المزارع التي تحكمهاحكومات المحلية</td>
</tr>
<tr>
<td>Al.mHlyp HkwmAt</td>
<td>Al.HkwmAt</td>
</tr>
<tr>
<td>local.def governments.indef</td>
<td>local.indef governments.indef</td>
</tr>
<tr>
<td>govern it</td>
<td>govern it</td>
</tr>
<tr>
<td>which farms.</td>
<td>which farms.</td>
</tr>
</tbody>
</table>

| Post Processed | 
| المزارع التي تحكمهاحكومات المحلية | المزارع التي تحكمهاحكومات المحلية |
| mHlyp HkwmAt | mHlyp HkwmAt |
| local.indef governments.indef | local.indef governments.indef |
| govern it | govern it |
| which farms. | which farms. |

| Reference | 
| المزارع الخاصة لإدارة الحكومات المحلية | المزارع الخاصة لإدارة الحكومات المحلية |
| Al.mHlyp Al.HkwmAt | Al.mHlyp Al.HkwmAt |
| local.def governments.def | local.def governments.def |
| the administration of | the administration of |
| that are under | that are under |
| farms. | farms. |
| 'Farms that are under the administration of the local governments’ | 'Farms that are under the administration of the local governments’ |

Table 6.4: Example about invalidity of BLEU

6.2.2 Human Evaluation

Side by side human evaluation experiments are used for the evaluation of our method. The goal of the experiments is to rate the translation quality. The human raters are provided with the source sentence and two Arabic sentence outputs, one is the output of baseline system and the other is the post-processed sentence. The sentences are shuffles; therefore, the raters score the sentences without knowing their sources. They give a rating between 0 and 6 according to meaning and grammar. 6 is the best rating for perfect meaning and grammar, while 0 is the lowest rating for cases when there is no meaning preserved and thus the grammar is irrelevant. Ratings from 5 to 3 are for sentences whose meaning is preserved but with increasing grammar mistakes. Ratings below 3 are for sentences which has no meaning, in which case the grammar becomes irrelevant and has minimal effect on the quality score.

Therefore, the human evaluation results are not expected to directly reflect whether the inflectional agreement, which is a grammatical feature, is fixed or not in the sentences. For sentences with high quality meaning, having the correct inflectional agreement should correspond to increasing the sentence score. However, sentences with no preserved meanings are not expected to receive higher scores for correct morphological agreement.
6 Fixing Inflectional Agreement through Post-processing

6.3 Experimental Setting

As a case study, we focus our experiments on the Adjective-Noun and Verb-Subject agreement relations. As mentioned earlier, the training data is a parallel corpus of aligned English-Arabic sentence pairs. The data contains surface forms of the English and Arabic words, i.e., no Arabic word segmentation is performed on the data. Training data is used for training the classifiers and also for the evaluation of their accuracy. Table 6.5 shows that the training data size as well as the number of occurrences of these relations in the training data. The adjective-noun and verb-subject relations are extracted according to the algorithm explained in section 6.1.3 and then used for training the classifiers.

<table>
<thead>
<tr>
<th>Number of Sentence Pairs</th>
<th>30412</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective-Noun Tokens</td>
<td>10087</td>
</tr>
<tr>
<td>Verb-Subject Tokens</td>
<td>3391</td>
</tr>
</tbody>
</table>

Table 6.5: Statistics about Training Data

The evaluation data is a set of sentence triplets extracted randomly from the web and translated by humans. Every triplets contains the English source, the MT Arabic output and an Arabic reference. In our evaluation data, only one Arabic reference sentence is available; however, the number of reference sentences can be and is preferred to be more than one. For BLEU score calculation, the post-processed Arabic output is evaluated against the reference sentences. For human evaluation, the MT output and the post-processed output are provided to the raters. Table 6.6 contains statistics about the evaluation data. Apart from only a small number of sentences, most adjective-noun tokens appear once per sentence. This means that about 38% of the sentences are considered for correction. Of course, some of these adjectives are generated correctly by the MT system, this is why they do not all encounter changes. Verb-subject relations appear in approximately 16% of the sentences. Table 6.10 shows the number of changed tokens as well as the number of changed sentences. It is clear from the table how the number of changed sentences is only different from the number of changed tokens by a small amount, especially per relation type.

<table>
<thead>
<tr>
<th>Number of Sentence Triplets</th>
<th>4993</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective-Noun Tokens</td>
<td>1914</td>
</tr>
<tr>
<td>Verb-Subject Tokens</td>
<td>806</td>
</tr>
</tbody>
</table>

Table 6.6: Statistics about Evaluation Data
6.4 Results

In this section, the results of the three evaluation methods are discussed. First, classification accuracy of our model is compared to two other baselines. Second, BLEU score is used for the evaluation of the final inflection generation. Finally, results of the human evaluation are discussed.

6.4.1 Prediction Accuracy

The feature prediction accuracy is the percentage of correct predictions out of the total number of predictions. We compare the accuracy of our classifiers to two other baselines. The comparison is done per feature.

The algorithms being compared are:

- **Classifier**: is our multi-class classification algorithm.
- **Majority**: makes prediction according to the most frequent label in the training data. For example, if the most common label for gender in the training data is *feminine*, this algorithm will predict all the gender features to be *feminine*. For high entropy features, this algorithm yields the worse prediction accuracy compared to low entropy features.
- **Copy**: copies the feature labels of the parent of the relation to the child. The motivation behind this algorithm is that the general rule in agreement is that the child follows the parent inflection.

Table 6.7 shows prediction accuracy for the features gen, num, prc0, stt, cas and per on adjective-noun relations. It is clear from the table how our classification algorithm produces the most prediction accuracy. The prediction for per is not applicable because these are adjectives, and per is the person feature and is defined only for verbs. For adjectives, the feature per has the same label of na.

The Majority algorithm has the worst performance on gen because it has the highest entropy. However, it yields good accuracy for the feature num because the majority of adjectives are in the singular form.

The Copy algorithm performs better than the Majority algorithm in all features except for the num feature.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>gen</th>
<th>num</th>
<th>prc0(AI)</th>
<th>stt</th>
<th>cas</th>
<th>per</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier</td>
<td>93.1</td>
<td>98.4</td>
<td>95.4</td>
<td>94.3</td>
<td>86.1</td>
<td>na</td>
</tr>
<tr>
<td>Majority</td>
<td>59.2</td>
<td>97.1</td>
<td>70.8</td>
<td>70.8</td>
<td>72.8</td>
<td>na</td>
</tr>
<tr>
<td>Copy</td>
<td>76.8</td>
<td>79.4</td>
<td>88</td>
<td>81.4</td>
<td>79.3</td>
<td>na</td>
</tr>
</tbody>
</table>

*Table 6.7: Prediction Accuracy: Adjective-Noun*
6 Fixing Inflectional Agreement through Post-processing

Table 6.8 shows the results of the three algorithms but on verb-subject relations. Also for verb-subject relations, our classification algorithm performs the best. However, for prc0, the accuracy is tied with the Majority algorithm. The reason is that in most verbs (99.9%), the label for prc0 is 0 meaning there is no attachments at this level. For per, the Majority algorithm yields prediction accuracy close to our algorithm because most verbs in the training data are in the form of 3rd person.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>gen</th>
<th>num</th>
<th>prc0</th>
<th>stt</th>
<th>cas</th>
<th>per</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier</td>
<td>82.6</td>
<td>96.2</td>
<td>99.9</td>
<td>na</td>
<td>na</td>
<td>97.9</td>
</tr>
<tr>
<td>Majority</td>
<td>57</td>
<td>95.4</td>
<td>99.9</td>
<td>na</td>
<td>na</td>
<td>97.7</td>
</tr>
<tr>
<td>Copy</td>
<td>77</td>
<td>83.5</td>
<td>41.22</td>
<td>na</td>
<td>na</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 6.8: Prediction Accuracy: Verb-Subject**

Finally, table 6.9 shows the classification accuracy of our algorithm when combining both adjective-noun and verb-subject relations together.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>gen</th>
<th>num</th>
<th>prc0</th>
<th>stt</th>
<th>cas</th>
<th>per</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier</td>
<td>91.04</td>
<td>97.8</td>
<td>96.5</td>
<td>95.6</td>
<td>89.4</td>
<td>99.5</td>
</tr>
</tbody>
</table>

**Table 6.9: Classification Accuracy: All**

6.4.2 BLEU Score

To evaluate the performance of the system, automatically, we computed the BLEU scores for six variations of the system. The label Adjectives is for the experiment when we only include adjective-noun relations, Verbs for the experiment with only verb-subject relations and finally All includes both relations. The other three experiments, are the same as above but with the language model filtering.

Table 6.10 shows statistics about the number of tokens whose features are corrected by the classifier in the first column. In the second column, the number of features whose inflected surface was updated. Column three lists the number of affected sentences for each experiment. The columns BLEU+ and BLEU- show the number of sentences where post-processing changes increases or decreases its baseline BLEU score, respectively. The last column lists the total BLEU score value for every experiment.

Table 6.10 shows statistics about the number of tokens whose features are corrected by the classifier in the first column. In the second column, the number of features whose inflected surface was updated. Column three lists the number of affected sentences for each experiment. The columns BLEU+ and BLEU- show the number of sentences where post-processing changes increases or decreases its baseline BLEU score, respectively. The last column lists the total BLEU score value for every experiment.

Although some features are changed for some of the tokens, their generated surface form was the same as the baseline. This can be an error by the generator, or the changed features only change diacritization and thus yield the same undiacritized surface form. By comparing the number of changed tokens to the number of changed sentences, the difference is very small especially for the separate relation experiments. This observation results in the conclusion that a relation type is usually found only once in a sentence.
One very important observation is that the number of sentences whose BLEU score was changed is on average only about 34%. This means that 66% of the changes have no effect on BLEU score. As clear from the table, the experiment which yields the maximum BLEU score value is All+LM with both relations and LM filtering. The improvement in BLEU score is not significant because, as shown in the fifth and sixth columns, only a small number of the sentences affect the BLEU score. Also, because BLEU is not a reliable metric for capturing morphological agreement, as explained in 6.2.1.

<table>
<thead>
<tr>
<th>System</th>
<th>Tokens(features)</th>
<th>Tokens</th>
<th>Sentences</th>
<th>BLEU+</th>
<th>BLEU-</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10.75</td>
</tr>
<tr>
<td>Adjectives</td>
<td>829</td>
<td>544</td>
<td>510</td>
<td>84</td>
<td>93</td>
<td>10.71</td>
</tr>
<tr>
<td>Adjectives+LM</td>
<td>-</td>
<td>309</td>
<td>297</td>
<td>56</td>
<td>41</td>
<td>10.76</td>
</tr>
<tr>
<td>Verbs</td>
<td>329</td>
<td>321</td>
<td>314</td>
<td>65</td>
<td>43</td>
<td>10.78</td>
</tr>
<tr>
<td>Verbs+LM</td>
<td>-</td>
<td>233</td>
<td>229</td>
<td>56</td>
<td>28</td>
<td>10.78</td>
</tr>
<tr>
<td>All</td>
<td>1165</td>
<td>878</td>
<td>799</td>
<td>145</td>
<td>131</td>
<td>10.75</td>
</tr>
<tr>
<td>All+LM</td>
<td>-</td>
<td>518</td>
<td>490</td>
<td>102</td>
<td>62</td>
<td><strong>10.80</strong></td>
</tr>
</tbody>
</table>

Table 6.10: BLEU Score Results

### 6.4.3 Human Evaluation Results

In this section, we show the results of the human evaluation explained in section 6.2.2.

The experiment was run to evaluate the system All+LM. The used sample size was 200 sentences obtained by randomly sampling the changed sentences. The mean of the difference between the scores of the two systems (Experiment - Baseline) was calculated to be 0.125 with a 95% confidence interval of ±0.1123. The T-statistic value is 2.2, which means that the post-processing system achieves a statistically significant improvement to the baseline system.

### 6.5 Limitations

The system suffers from the accumulation of errors through the system pipeline. The errors from alignment, English parsing, Arabic analysis, syntax projection, classification and finally Arabic generation are all reflected in the final result. Because every operation uses the output of previous steps, errors are cascaded and increased until the final stages of the system.

In syntax projection, besides the cascaded errors from alignment, parsing and Arabic analysis, imprecisions can also result for two more reasons. First, sometimes there are agreement relations in the Arabic sentence that do not correspond to relations in the English dependency parse tree. For example, there can be an adjective-noun relation in the Arabic sentence that does not
correspond to an *amod* relation in the English sentence, in which case it will not be recognized and consequently will not be corrected. Second, unresolvable ambiguity can result when an English word aligns to multiple Arabic words.

Classifiers’ training uses noisy feature vectors and labels. Classification of a word is built on a noisy analysis of the parent word. Finally, the generated surface form can be incorrect because of all these cascaded errors.

Efforts to add robustness to the system were exerted. For example, a relation extraction approach the makes use of syntactic information of both languages is used instead of the direct projection approach. Extracted feature vectors include information from both languages as well as features that overcome common errors by the Arabic analyzer. Although a language model is used to add robustness to the system against these cascaded errors, it has its own limitations. For a language with such large vocabulary, a strong language model that does not suffer from poor probability distribution and out of vocabulary words does not exist.

In the future, further improvements can be achieved by better refinement of the features included in training and classification. Besides, improving the analysis and generation tools is expected to produce a direct improvement to the final system output. The quality of the training data can be improved by using the output of forced-alignment filtering [HWVN10]. In forced-alignment, only Arabic sentences that can be generated from the English sentences using phrases present in the phrase table can be used. This can be helpful because phrases in the phrase table go through filtering and thus are more reliable to be considered correct mappings between the source and the target languages. Also, because the phrases are extracted after combining the alignments of both directions: source-target and target-source, the combined alignments are of higher quality than only using source-target alignment.

Experiments where the SMT system is trained on normalized Arabic words instead of the fully inflected surface forms can be conducted. In this case, the input to the post-processor will be uninflected text, which the post-processor will inflect. If a robust, high-quality normalization tool exists for Arabic, this approach is expected to improve SMT. However, high quality normalization tool is hard to find for Arabic mainly because of ambiguity.
Conclusion

We have shown that incorporating linguistic knowledge in SMT can achieve significant improvements even to a state-of-the-art SMT system. We have shown that English-Arabic SMT can be improved by methods which target the specific language pair. We proposed two approaches which were motivated by the manual error analysis conducted on the output of an English-Arabic SMT system similar to translate.google.com. Careful error analysis proved to be very effective in directing our research to the most important problems. Thorough understanding of the lexical, syntactic and morphological differences between the two considered languages was also necessary. The difficulties of English-Arabic translation in addition to a summary of the error analysis were provided in chapter 3.

In the first approach, we added linguistically motivated features to the phrases in the phrase table and trained the features using MERT. These features were based on the POS tags and the dependency parse tree relations of the English phrases. This approach achieved improvements in BLEU score over the state-of-the-art system.

In the second part of this work, we presented a novel post-processing system that targets morphological agreement in the machine translation output sentences. We have shown that a post-processing module that fixes morphological agreements can be successfully integrated to SMT systems. We have also shown that modeling agreement relations separately is very effective. Our system achieves a statistically significant improvement to the state-of-the-art SMT system, as proven by the human evaluation results.

Although our current system fixes only two types of morphological agreements: verb-subject and adjective-noun agreements, significant improvements to the SMT system were obtained. Extending the system to include other morphological agreements is expected to achieve even better results.

In the future, combining the two methods will be considered. Exploring other features and
developing better training algorithms that can reliably optimize for a large number of sparse features are possible research directions. Moreover, we can experiment with integrating the post-processing module to other machine translation systems and not only SMT systems.
Bibliography


Bibliography


[JM00] Daniel Jurafsky and James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and


Bibliography
