future internet

MDPI

Article

Improved Arabic–Chinese Machine Translation with Linguistic Input Features

Fares Aqlan [1], Xiaoping Fan [1,2,*], Abdullah Alqwbani [1] and Akram Al-Mansoub [3]

[1] School of Information Science and Engineering, Central South University, Changsha 410083, China; aqlanfares@csu.edu.cn (F.A.); alqwbani@csu.edu.cn (A.A.)
[2] Academy of Financial and Economic Big Data, Hunan University of Finance and Economics (HUFE), Changsha 410205, China
[3] School of Computer Science and Engineering, South China University of Technology (SCUT), Guangzhou 510006, China; almansoub555@gmail.com
* Correspondence: xpfan@csu.edu.cn

Received: 4 December 2018; Accepted: 16 January 2019; Published: 19 January 2019

Abstract: This study presents linguistically augmented models of phrase-based statistical machine translation (PBSMT) using different linguistic features (factors) on the top of the source surface form. The architecture addresses two major problems occurring in machine translation, namely the poor performance of direct translation from a highly-inflected and morphologically complex language into morphologically poor languages, and the data sparseness issue, which becomes a significant challenge under low-resource conditions. We use three factors (lemma, part-of-speech tags, and morphological features) to enrich the input side with additional information to improve the quality of direct translation from Arabic to Chinese, considering the importance and global presence of this language pair as well as the limitation of work on machine translation between these two languages. In an effort to deal with the issue of the out of vocabulary (OOV) words and missing words, we propose the best combination of factors and models based on alternative paths. The proposed models were compared with the standard PBSMT model which represents the baseline of this work, and two enhanced approaches tokenized by a state-of-the-art external tool that has been proven to be useful for Arabic as a morphologically rich and complex language. The experiment was performed with a Moses decoder on freely available data extracted from a multilingual corpus from United Nation documents (MultiUN). Results of a preliminary evaluation in terms of BLEU scores show that the use of linguistic features on the Arabic side considerably outperforms baseline and tokenized approaches, the system can consistently reduce the OOV rate as well.

Keywords: Arabic morphology; factored translation model; phrase-based machine translation; pre-processing; statistical machine translation

1. Introduction

With the rapid development of communication technology and economic globalization, translation between languages has become increasingly frequent, thereby drawing growing attention to machine translation (MT). Consequently, as a subtopic of artificial intelligence, MT has achieved considerable progress in recent decades. The primary goal of researchers in this field is to make the translation quality of machines closer to that of humans.

Several approaches have been demonstrated to be useful in MT. Statistical MT (SMT) is one of the most widely supported approaches, whereas phrase-based SMT (PBSMT) is considered state-of-the-art MT [1]. PBSMT uses a sequence of words, aptly called "phrases" or "blocks," rather than single words. This approach segments the source content into phrases translates phrases into the target translation

units provides more reliable quality than word-based approaches [2]. However, the traditional PBSMT models are distance-based that consider the reordering between both sides as equal, and therefore it still has many shortcomings which require in-depth investigations in various aspects. For instance, the weakness of controlling the reordering decisions [3], the data sparseness issue that becomes more prevalent while translating a highly-inflected language, and the lack of additional linguistic information of parallel corpus, which affects translation accuracy, particularly for a morphologically rich language, such as Arabic.

The authors in [4] presented a factored translation model as an extension of PBSMT. This model integrates translation corpora into linguistic information, such as syntactic and morphological features, thereby achieving better translation for many language pairs. However, one of the significant problems in PBSMT is the weakness in dealing with the new form of the surface after the integration of linguistic information or features. In this work, we attempt to address the integration of these features into the source side of the model to enhance the translation of the Arabic–Chinese language pair, given the importance and necessity of these languages in economic, cultural, and global aspects.

On one hand, deep learning and machine learning techniques have been widely applied and shown significant improvement in different tasks, such as speech recognition, transfer learning, load monitoring, and others. An example of this is the pipeline presented by the authors of [5] to overcome the limitations raised by the scarcity of large amounts of training data as well as time-consuming and complicated back-propagation systems. The framework is based on extreme learning machines that can predict the unknown relationship between factors of the input and output sides by using the collection of hidden units and output weights; it has reported top performance on the UK domestic appliance-level electricity (UK-DALE) dataset.

As for speech recognition, the authors in [6] used Hermite polynomials to adapt the activation functions for audio modeling showing a powerful ability to learn nonlinear and more complex feature representations. The results of experiments showed that the proposed method provides a solution for data scarcity. However, the fact that computing the nonlinear transformations by using a recurrence formula can highly increase the computational load.

One of the recent work on transfer learning is by the authors of [7], who proposed a transfer learning algorithm for remaining useful life prediction of a turbofan engine. This algorithm is based on bi-directional long short-term memory neural networks to deal with the difficulty of obtaining an appropriate number of samples in data-driven prognostics. The results showed that transfer learning provides an efficient performance in most cases, however, for transferring from multiple operating conditions to single operating conditions, the transfer learning led to a worse result.

Another line of work in low-resource scenarios is speech-to-text translation (ST) that has been proved valuable, for instance in the documentation of unwritten or endangered languages. Traditional ST presents a pipeline of automatic speech recognition (ASR) and MT where transcribed source audio and parallel text are required to train ASR and MT, respectively. In [8], the authors proposed a pipeline to extract information from speech signals to produce ST alignments where the speech of source language (as an endangered language) was annotated with translations than with transcriptions. This pipeline aids in Computational Language Documentation (CLD), and hence it provides valuable resources for the training of ASR systems.

As the number of unwritten or endangered languages is so large, the documenting of these languages is becoming an interesting area for research. One of the recent efforts in this area is the project breaking the unwritten language barrier (BULB) [9] which focuses on collecting spoken translations along with a speech by analyzing and annotating audio recordings using automatically discovered linguistic units (phones, morphs, words, grammar, etc.). An example of a speech translation dataset is the Fisher and Callhome Spanish–English corpora [10] that contains transcribed and translated telephone conversations of Spanish to English. Despite the medium size and low-bandwidth recordings of this corpora, it has been used in many previous works dealing with Spanish as a low-resource language which makes it an appropriate language to work with.

However, to make ASR systems perform reasonably well, a large amount of annotated data is needed. Many languages lack available speech data or the annotations required to train ASR; according to [9] only about 1% of the world's languages have the minimum amount of available data. Furthermore, using a system trained for a different or related language (as cross-language) to build ASR system for low-resource languages leads to quite poor performance [11].

Due to the scarcity of available documentation scenario that matches Arabic and Chinese as well as the expensive process and the required knowledge of domain to obtain hand-labeled training data, we leave the investigation of ASR technology between this language pair for future work, since this area is bringing more questions than answers. In these scenarios, a direct MT system is appealing.

Arabic–Chinese MT is a challenging task because they are low-resources languages that belong to different families. Arabic is a morphologically rich language with many morphological features, such as affixes, stems, roots, and patterns [12]. Also due to the highly-inflected words in Arabic, the well-known issues of missing words and data sparsity arise; and hence affect the accuracy of translation and leads to more out-of-vocabulary (OOV) words [13]. By contrast, Chinese is written using logograms and lacks linguistic morphology [14].

In this study, we propose an approach that takes advantage of the rich linguistic information of Arabic to improve translation quality between Arabic (as the source side) and Chinese (as the target side). First, we introduce a baseline model of our experiments using the default settings of PBSMT. To solve the sparse data problems, we apply multiple optimizing tokenization schemes to the preprocessing steps in later experiments and then compare their performance. Finally, we inject several linguistic factors into Arabic to create an optimized factored model. The factors used in our experiments include surface form, lemma, part-of-speech (POS), and morphological features (morph). The morph features such as case, state, gender, gloss, person, number, voice, and mood. We investigate the effects of these factors individually and on combination models.

Translation models with general representations such as lemma form provide improved generalization based on the inflectional variants for the same word. While POS can draw more statistics that help in disambiguation, morph features obtain better data representation and linguistic knowledge that improve generalization ability. The best combination and alternative paths models based on these features can consistently overcome data sparseness and reduce OOV rates caused by limited training data. Several advantages of these linguistic features can enrich the training pipeline by proper coherent sentences, and thus improve the translation quality and accuracy. While neural machine translation (NMT) shows state-of-the-art performance on the plentiful training data, PBSMT still competitive or superior in the case of the low resource we focus on [15]. Furthermore, PBSMT and NMT achieve comparable performance on Arabic–English, and Arabic–Hebrew MT systems [16,17].

The rest of the paper is organized as follows. Section 2 reviews previous works on Arabic–Chinese MT and factored translation systems. Section 3 presents the challenges in building MT between Arabic and Chinese and the primary methods used in our work, such as preprocessing techniques and input linguistic features. Section 4 discusses the factored MT approach. Section 5 describes the data preparation and experiment setting for the phrase-based and factored models. Section 6 explores the evaluation results in terms of BLEU scores and OOV rates, which exhibit considerable improvements by using linguistic features. Section 7 concludes this work and provides suggestions for future work.

2. Related Works

Previous work on MT between Arabic and Chinese is unexpectedly scarce, despite the economic and global prevalence of these languages. Chinese has approximately 1.2 billion native speakers, whereas Arabic has nearly 400 million. Such values make Arabic–Chinese MT an important and attractive field in linguistics. Two aspects may explain the limited research on MT on this language pair: (a) the scarcity of freely available parallel corpora; and (b) the morphological differences and complexity that require linguistic knowledge.

In the current work, we discuss related studies in this field. The first work was proposed by [18] who compared between the direct approach of SMT and the pivot-based approach using English as the pivot language in translating Arabic into Chinese. The result showed that the pivot-based approach performed better than the direct translation. Although the pivoting strategy provides a solution to the lack of parallel data, it causes information loss when using a morphologically poor language, such as English, as a pivot language [19]. In [20], the authors collected a corpus from United Nations (UN) documentation and used it to build standard PBSMT systems of Arabic–Chinese and Chinese–Arabic translations. The authors in [21] compared between SMT and NMT across 30 translation directions including Arabic–Chinese but did not conduct any individual preprocessing for Arabic. In [22], the authors evaluated SMT performance using several tokenization schemes of MADAMIRA for Arabic (as source language) translated into Chinese and other languages, the results of this framework showed that using Arabic pre-processing enhanced the translation and helped to overcome the data sparsity problem. However, all previous studies on this language pair have presented without using any linguistic features, which have been proven valuable in the MT field. We address this limitation in this research.

Studies that used linguistic annotation have been performed in different languages (A large amount of studies have been done on this topic. Here we only review some example works) with annotation factors such as lemma, POS, morph, and others. The authors in [23] explored the difficulties of integrating linguistic features into PBSMT model. In [24], the authors achieved improvements by adding linguistic annotations to English (as the source side) translating into Greek and Czech, where the proposed models reduced the error of noun case agreement and verb conjugation. The authors in [25] discussed the benefits of enriching the input with morphological features to enhance the translation from English into Hindi and Marathi; the results showed that the integrated models reduced the number of OOV words and improved the fluency of the translation. In [26], the authors created factored translation models from English to Slovenian and compared the results with non-factored approaches; their experiments showed that the factored approach is better than the non-factored while translating complex texts. In [27], the authors discussed the benefits of annotation features for several European language pairs, in which the experiments showed that factored models have better results in enhancing the translation quality.

Studies on factored translation models between Arabic or Chinese and other languages are few. For instance, the authors in [28] used additional features to create English–Arabic factored models and compare the performance with segmented models, the factored models achieved better quality. In [29], the authors developed a factored approach that improved Arabic–English translation by injecting POS and combinatory categorial grammar (CCG) into English, where factored models gained better quality enhanced the grammatical performance. The authors in [30] presented the results of PBSMT and a factored translation model between Korean and Chinese, in which the factored model performed a better MT quality compared to the standard PBSMT model.

Some previous works explored the effects of factors on NMT as a new state-of-the-art MT. The authors in [31] investigated the benefits of using linguistic input features for English↔German and English→Romanian NMT systems which showed improvement upon the baseline system. In [32], the authors discussed the effects of pre-reordering and input linguistic knowledge into Japanese–English, and Chinese–English NMT system, in which the features used including POS tags, word class, and reordered index. According to the results of this framework, linguistic features are useful in enhancing the translation quality, and demonstrated a better performance than the pre-ordering approach. In [33], the authors presented NMT models using factors on the target side of English-French, the results showed that factored NMT reduced the OOVs rates and gained a better performance over the baseline. However, in this work, we create a PBSMT system that remains competitive or superior under low-resource conditions [15].

In contrast to previous work, the present study intends to take advantage of the linguistic information of Arabic to build the first factored Arabic → Chinese translation model to improve

the quality of direct translation between this language pair as well as overcome the issues of missing and OOV words. This model extends PBSMT by adding well-known features of the Arabic side. Then, the results are compared with multiple standard PBSMT systems using different preprocessing techniques.

3. Challenges and Approach

In this section, we explore the challenges and motivation in building an Arabic → Chinese MT model and the primary methods used in this framework.

3.1. Linguistic Issues

Arabic and Chinese belong to different language families. Arabic is a morphologically rich and complex language with a typical verb–subject–object (VSO) order. Although VSO is the typical word order of Arabic, the corpora used show a mixed distribution of VSO, subject-verb-object (SVO), and verb-object-subject (VOS) clauses (VOS order is admitted in specific contexts, for instance using a pronoun to express the object). By contrast, Chinese lacks morphology and exhibits a systematic word order of SVO. Word order in Chinese is considerably closer to that in English compared with that in Arabic. Figure 1 shows an example of Arabic typical order and Chinese order.

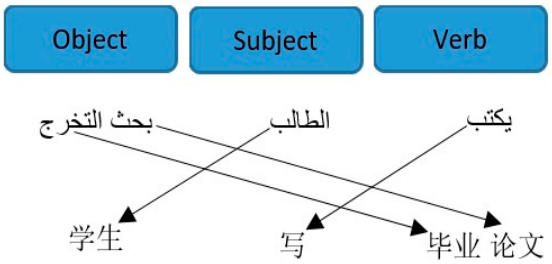

Figure 1. Example of the sentence order in Arabic and Chinese.

The different word order of this language pair makes the accuracy of word alignment in MT difficult to achieve. This issue causes data sparseness problems, given that numerous words are misaligned cannot be found in the training corpus, thereby leading to low-quality translation results.

Furthermore, dealing with morphology-rich languages is difficult in SMT. A word in Arabic may have various morphologies to represent different meanings (affixes, stems, and clitics) with several types of morphological features, such as person, number, case, and tense. Moreover, optional diacritical marks in Arabic words cause further ambiguity. These features of Arabic cause more challenges, as they worsen the data sparsity issue, increase the OOV words, and consequently complicates the alignment of word-level in translation between Arabic and another language.

In contrast, Chinese is considered a morphology-poor language. A Chinese word can be composed of a single character or multiple characters without word boundaries, and it has a complicated system of quantifier formation and verbal aspects. The authors in [18] discussed in detail the challenges of building a direct MT approach between Arabic and Chinese.

Consequently, the morphology analysis of word units becomes increasingly necessary to improve the translation results between Arabic and Chinese. In this work, we present solutions for these issues by using preprocessing tokenization schemes and injecting linguistic factors into the source side of the translation corpora.

3.2. Arabic Preprocessing

The morphological complexity of Arabic language makes it difficult to achieve high quality and accuracy translation between Arabic and various languages. To facilitate the translation process, we first need to analyze the morphological and orthographic features contained in the Arabic word

by using preprocessing techniques; it is aptly called the preprocessing schemes. The advantages of morphological preprocessing have been proven useful for Arabic [12], this procedure helps reduce data sparseness, thereby improving translation quality. In this work, we use three preprocessing schemes to compare effects across different techniques on Arabic-to-Chinese translation.

For the baseline approach, we test the tokenization script *tokenizer.perl* of Moses [34], which simply separates the punctuation marks of Arabic side using the same default as English tokenization.

We also experiment with MADAMIRA [35], a state-of-the-art morphological analyzer for the Arabic language to implement the Penn Arabic Treebank (ATB) scheme as morphology-aware tokenization. Except for the definite article, the ATB scheme tokenizes all clitics on the Arabic side. It also has a default normalization for '(' and ')' characters into "-RRB-" and "-LRB-", in addition to the well-known Alif/Ya Arabic normalization. This scheme performs better in Arabic–English MT compared with other schemes [36].

Furthermore, we use a new tokenization scheme of MADAMIRA, the D3* scheme proposed by [22], which tends to exhibit better performance in Arabic-to-Chinese translation and other multiple languages. Compared with the ATB scheme, D3* first requires tokenization according to the D3 scheme, which tokenizes all clitics, splits ال(Al) (Arabic transliteration is presented according to the Habash-Soudi-Buckwalter scheme [37]) "the" definite article, and then removes the definite article from the Arabic context.

3.3. Linguistic Input Features

Many previous works have demonstrated the benefits of the factored model as an extension of PBSMT between different languages, which annotates the source side and/or target side with linguistic features. In this work, we focus on the annotation of the source side (i.e., Arabic) using some popular linguistic features to determine whether these features will improve the quality of Arabic → Chinese translation. We use MADAMIRA, which analyzes the Arabic context by applying rule-based and supervised learning techniques to implement the morphological features of the input word. Several types of linguistic features can be obtained by MADAMIRA for Arabic. Here, we discuss additional details about the individual factors we use in our experiments.

Lemma:

Arabic is a highly inflected language, which makes Arabic lemmatization important in the preprocessing step to enhance the information extraction results. The inflections of Arabic words are generated by adding prefixes, suffixes, and vowels to the root. For example, the word وسيخبرونها(wsyxbrwnhA) "and they will tell her" has the prefix (ws) "and will" and suffix (wnhA) "they her", which are both attached to the root (xbr) that basically means "telling." This word also has the stem (yxbr) "tell" and the lemma (xbr) "the concept of telling".

Given its language complexity, lemma exhibits better performance in Arabic information retrieval compared with the root and stem. The root leads to low precision, e.g., differences exist between the stem for broken plurals and their singular patterns. Moreover, imperfect verbs have different stems with their perfect verbs.

Several natural language processing (NLP) tools can be used to output lemmas for Arabic words. In this work, we use MADAMIRA morphological lemmatizer. The reported lemmatization accuracy for modern standard Arabic by the MADAMIRA system is 96.2%, and it has been evaluated on a dataset extracted from the Penn Arabic Treebank. Each Arabic token in our corpora has a diacritized lemma, as shown in the example of Table 1.

POS:

POS tags play an essential role in different NLP tasks, such as MT. These tags provide the linguistic knowledge and syntactic role of each token in the context, which helps in information extraction and reduces data ambiguity.

The basic POS for Arabic words has many categories: noun, e.g., كتاب(ktAb) "book" or كتب(ktb) "books" in plural form and مكتوب(mktwb) "writing or message;" verb, e.g., يكتب(yktb) "write" or كتب(ktb) "wrote" in past form; adjective, e.g., مكتوب(mktwb) "written or fated;" and particle, e.g., من(mn) "from" or إلى(Ălý) "to."

From the preceding examples, we can see that the word كتب(ktb) "books" can be used as a plural noun and as the verb "wrote" in past form. Meanwhile, the word مكتوب(mktwb) can be used as a noun "message" or an adjective "fated". In such situation, POS tagging helps analyze and distinguish word meaning, which is optically called "word sense" in the translation corpora.

In this work, we extract POS tags for the input tokens of Arabic by using the MADAMIRA morphological analyzer, which has been considered a state-of-the-art Arabic tagger with a POS accuracy of 96.91% [38]. As shown in Table 1, the results of the MADAMIRA tagger annotate each word by its POS tag.

Morph Features:

MT approaches suffer from data sparseness problems when translating into or from morphologically rich and complex languages, such as Arabic. Thus, morphology analysis is necessary to handle data sparseness and improve translation quality.

Different word types in the Arabic language have various sets of morph features. For example, verbs have person, gender, number, voice, and mood, whereas nouns have case, state, gender, gloss, number, and the attached proclitic DET. Concatenative speech includes affixes and stems, whereas templatic speech has root and patterns.

To enable our approach to utilize the advantage of linguistic knowledge, we use the MADAMIRA analyzer to annotate the Arabic input with morph features because it provides the structure and form of each word in the corpus (see Table 1).

Table 1. An example of the Arabic features in our approach.

Word	Transliteration	Lemma	POS	Morph. Features
تقرر	tqrr	تَقَرَّر	verb	PV+PVSUFF_SUBJ:3MS
أن	Ân	أنْ	conj_sub	SUB_CONJ
تدرج	tdrj	أدْرَج	verb	IV3FS+IV_PASS
في	fy	فِي	prep	PREP
جدول	jdwl	جَدْوَل	noun	NOUN+CASE_DEF_GEN
القمة	Alqmħ	قِمَّة	noun	DET+NOUN+NSUFF_FEM_SG+CASE_DEF_GEN
.	.	.	punc	PUNC

4. Factored Translation Model

The factored model represents the extension of standard PBSMT by using a log-linear approach to combine language, reordering, translation, and generation models. The factored model is based on the integration of rich linguistic features into the translation model, where the word form not only becomes a token but also a vector of factors that provide various knowledge levels.

The following example is extracted from our factored corpus to show how the Arabic word الوثائق(AlwθAŷq) "documents" is integrated and aligned using the format of surface | factor1 | factor2 | factor3 | ...

الوثائق | وَثِيقَة | noun | DET+NOUN+CASE_DEF_NOM

Although these factors do not have particular meanings within the model, they enrich the translation model with general representations to overcome the problems of data sparseness in limited training data. Moreover, these factors allow the direct modeling of many translation aspects, such as morphological, semantic, and syntactic levels [12]. For instance, words with the same lemmas allow their inflectional variants to share better representation in the model, which helps reduce the data sparseness problems. The POS tags are also beneficial for disambiguation.

Figure 2 shows the diagram of our factored model approach, which we report on experiments using features, such as lemma, POS, and morph features as additional annotations apart from surface form. We compare the effects of several configurations based on these input factors.

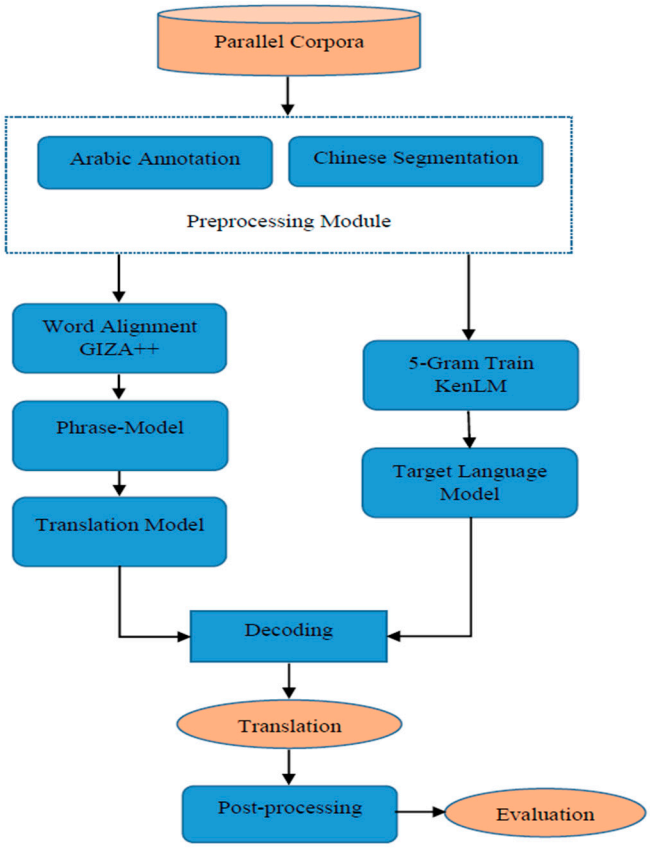

Figure 2. Diagram of the proposed approach.

Although the generation step can be used in factored models to generate the probabilities of translation components based on target factors, this framework does not use a generation step because we focus on source factors; instead, the words with their factors assign the same probability. The authors in [39] indicated that sparse data arise when no mapping (1–1) exists between the surface+factor on the source side and the inflected word on the target side. To bridge this gap and reduce data sparseness, we test alternative paths in the decoding step using multiple decoding paths and backoff configurations to cover occurrences wherein a word appears with a factor that has not been trained with. Additional details about these configurations are provided in Section 5.

5. Experiments Settings

5.1. Data Preparation

For our experiments, we extract the Arabic–Chinese corpus from the Multi-UN corpus [40], which used in many MT studies. We remove suspicious sentences with wrong language text or considerable non-alphanumeric symbols. We use the Moses script *remove-non-printing-char.perl* to remove non-printing characters, and *clean-corpus.perl* to eliminate sentences that are longer than 80 tokens.

To best capture the effects of the preprocessing and linguistic features, we experiment with the relatively medium dataset sizes, where the data sparsity problem becomes of more relevance [22]. For training, we utilize 200,000 lines, 1600 lines as development data, and 2000 lines for evaluation. Table 2 summarizes the statistics of our parallel corpus in words and sentences.

Table 2. Statistics of the Arabic-Chinese parallel corpus. For Chinese, size refers to the segmented words.

Corpus	Sentences	Arabic Words	Chinese Words
Train	200,000	5,306,635	5,509,739
Tune	1600	34,618	34,091
Test	2000	44,646	48,709

5.2. Chinese Segmentation

Given that Chinese words are composed of one or multiple characters without spaces between words, segmentation is an essential task in the preprocessing steps of MT [41]. Chinese segmentation separates words in a sentence to make words 1-to-1 mapping between the parallel phrases, and therefore the segmentation is necessary for the word alignment. However, the segmentation faults caused by segmentation schemes lead to word-mismatching problems which, likewise, affect the translation quality. In this work, we use the Stanford Word Segmenter (http://nlp.stanford.edu/software/segmenter.shtml) for Chinese [42], which is a conditional random field-based segmenter that splits the Chinese context into words separated by spaces according to Chinese Treebank segmentation. An example of a segmentation output is as follows:

Sentence: 我相信，这些努力正逐渐地取得成果。
Output: 我 相信 ， 这些 努力 正 逐渐 地 取得 成果 。

5.3. Machine Translation (MT) Systems

In all models, we build translation systems using Moses, which is an open source MT toolkit. Using GIZA++ [43] to extract phrase pairs with word alignment, the alignment symmetrization follows grow-diag-final-and and msd-bidirectional-fe for lexical reordering. On the target side of the parallel corpora, we use KenLM [44] to create a 5-gram language model and memory mapping. In the tuning process, minimum error rate training (MERT) is used to perform the tuning. The experiments are discussed in detail for the standard PBSMT and factored MT systems.

5.3.1. Phrase-Based MT Models

In this approach, we run three different tokenizations for the Arabic side to evaluate the translation output through preprocessing schemes in the standard PBSMT system.

Baseline: This experiment is the baseline of all our work, in which we use minimal preprocessing on the Arabic side by applying Moses default tokenizer to separate punctuation marks between Arabic words.

Tokenized-ATB: The Arabic corpus is tokenized and normalized using the ATB scheme of the MADAMIRA morphological analyzer. Except for the definite article, this process tokenizes all clitics added to an Arabic word. It also normalizes according to the default (Alif/Ya) normalization; an example is shown in Figure 3.

Tokenized-D3*: The tokenization scheme is the same as that of the D3 scheme, which tokenizes all clitics and splits the definite article because this article is attached to an Arabic word. However, this case does not occur in Chinese. Instead, we use the MADAMIRA scheme to remove all the definite articles in the Arabic corpus to make it closer to Chinese.

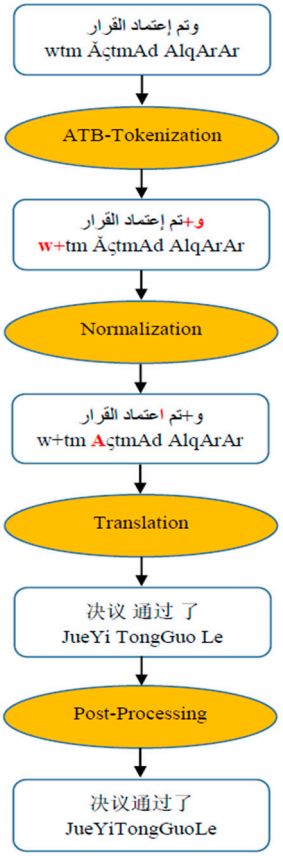

Figure 3. Step-wise result of the Penn Arabic Treebank (ATB) approach, the red letters are normalized and tokenized by the approach, Chinese pinyin is presented to ease readability.

5.3.2. Factored Models

This part discusses the core of this work. We annotate the Arabic side by adding linguistic features on top of the surface forms in the translation model when translating Arabic into Chinese to optimize the translation results. These features provide rich representation and sufficient knowledge for nearly any required transformation. In this framework, we use three features (lemma, POS, and morph) for Arabic to conduct several experiments using different settings based on these features. Each word in Arabic factored corpus is a vector of features that represent information levels. For Chinese, we use Chinese segmented words.

Surface+Lemma: This model incorporates diacritized lemma into the input side to reduce data sparseness and provide improved generalization based on the inflectional variants for the same word. Mapping lemma and surface onto surface enhances performance.

Surface+POS: The translation option of this model is based on adding POS tags to the input corpus. POS tags help in disambiguation and in extracting additional information about the data.

Surface+Morph: To obtain better data representation and linguistic knowledge, we inject an input word using several morph features which are case, state, gender, gloss, number, person, voice, and mood. These features improve generalization ability.

Surface+Lemma+POS: In this integrated model, we incorporate lemma and POS features to solve the data sparseness and disambiguation problems and evaluate whether multiple factors also increase translation performance.

Surface+Lemma+Morph: Here, we combine two features (lemma and morph) on top of the surface form. The motivation is to obtain more flexible notions and improve the precision of the translation output.

Surface+POS+Morph: The word in this model represents the vectors of the POS tags and morph features to enrich the input through additional knowledge, thereby solving the data sparseness problem and enhancing generalization.

Surface+All Features: We also test the effect of integrating all available features (lemma, POS, morph) to determine whether the addition of such rich information will improve the translation quality of this language pair.

Multiple Decoding Paths Lemma/Morph: The factored model allows the use of multiple paths in parallel; that is, translation options originate from different phrase tables. We set two models in this model. First, the surface-level model maps the surface onto the surface. Second, the lemma/morph model provides morphological analysis. Translation options from multiple tables compete in the decoding step. When the same translation is found in different tables, varying scores are used to create translation options for each occurrence. The translation model that uses the multiple decoding path (MDP) strategy becomes more robust; hence the input sentence is translated with higher probability.

Lemma Backoff: For the translation of morphologically rich and complex languages, such as Arabic, into simpler languages, such as Chinese, translating lemmas instead of the words that have not observed in the training corpus is useful. This strategy is called the backoff model. In contrast to MDP, the decoder in the backoff model finds one phrase in different phrase tables.

The first table is used as a priority table, whereas the second table is a backoff table for translations that are not found in the first one. In this framework, the surface level is a priority phrase table, whereas the lemma-level phrase table is used as a backoff phrase table, as shown in Figure 4. This model helps decrease out-of-vocabulary rates and enhances translation quality.

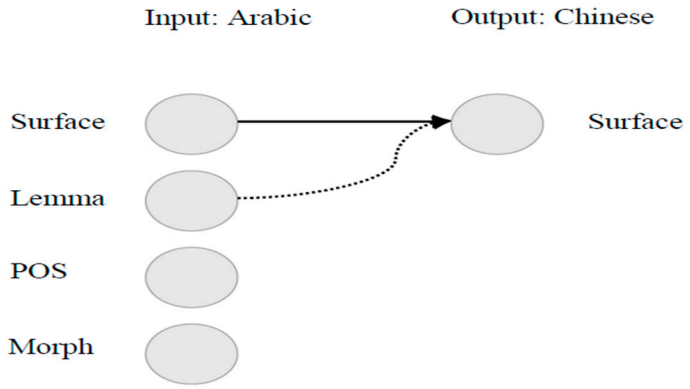

Figure 4. Lemma backoff model for the Arabic-Chinese factored MT.

5.4. Automatic Evaluation

We conducted 12 experiments on a subset of the Multi-UN corpus to evaluate the performance of standard phrase-based MT and the factored models for the Arabic–Chinese language pair. For evaluation, we used a test set with one reference throughout all our experiments.

To make the evaluation as fair as possible, minimize the effects of Chinese segmentation and understand the effects of Arabic pre-processing and features, the Chinese output was post-processed before evaluation by using the script *deseg.py* (https://github.com/EdinburghNLP/wmt17-scripts/blob/master/en-zh/deseg.py) of WMT17 for Chinese desegmentation. This process merges the Chinese output by removing all spaces (except for texts with an ASCII letter on both sides) and then converts ASCII commas and periods to their equivalent CJK Unicode.

We measured the translation results with BLEU scores [45] by using the *multi-bleu-detok.perl* (https://github.com/EdinburghNLP/nematus/blob/master/data/multi-bleu-detok.perl) script, which provides the exact same case-sensitive results as *mteval-v13a.pl* on untokenized text. To compare the results of different PBSMT and factored systems versus the baseline system, statistical significance tests were carried out on BLEU with paired bootstrap resampling method

(http://www.cs.cmu.edu/ark/MT/paired_bootstrap_v13a.tar.gz) [46], which measures p-value statistical confidence using 1000 iterations.

To evaluate the effect of the pre-processing schemes and linguistic factors on data sparsity, we measured the rates of OOVs unigrams on the test set in terms of tokens and types. The tokens refer to the total number of words, while the types indicate the number of unique words in the text. Table 3 summarizes the findings.

Table 3. BLEU scores with the improvement over the baseline model, along with the effects of pre-processing schemes and linguistic factors on the test set according to OOV rates. Besides, results of statistical significance test where ✓indicates a significant improvement over the baseline with the specified p-value. The best result in terms of each metric (the highest BLEU and lowest OOVs) is highlighted in bold.

System	Model	BLEU		OOV		Sign.	
				Tokens (%)	Types (%)	$p < 0.05$	$p < 0.01$
PB-SMT	Baseline	16.13	–	5.74	22.80	–	–
	Tok-ATB	17.32	(1.19)	1.52	12.53	✓	
	Tok-D3*	17.67	(1.54)	1.40	11.20	✓	
Factored-MT	Surface+Lemma	17.96	(1.83)	1.72	14.51	✓	✓
	Surface+POS	**18.30**	**(2.17)**	1.70	14.33	✓	✓
	Surface+Morph	17.90	(1.77)	1.71	14.42	✓	✓
	Surface+Lemma+POS	17.48	(1.35)	1.80	13.92	✓	
	Surface+Lemma+Morph	17.79	(1.66)	1.72	14.16	✓	
	Surface+POS+Morph	17.43	(1.30)	1.75	14.20	✓	
	Surface+All Features	17.82	(1.69)	1.69	13.19	✓	✓
	Multiple Decoding Paths	17.66	(1.53)	0.81	6.72	✓	
	Lemma Backoff	17.59	(1.46)	**0.70**	**6.71**	✓	

6. Results and Analysis

Table 3 shows the results for Arabic-to-Chinese translation on the extracted Multi-UN corpus. In PBSMT, we observe that tokenization strategies (ATB and D3*) have a major effect in alleviating the data sparseness problem, improve the translation quality and achieve higher BLEU scores than the baseline with an advantage to the D3* scheme. The differences in terms of BLEU scores between baseline and both tokenization schemes are statistically significant (p-value < 0.05). This result confirms that Arabic preprocessing helps address data sparseness.

All the factored systems considerably outperform the baseline and achieve better or comparable performance with tokenized PBSMT. The differences in BLEU scores between tokenized phrase-based MT and the factored systems are not statistically significant, whereas the differences between the baseline and all the factored systems are statistically significant (p-value < 0.05). Moreover, using a single feature provides a better result than a combination of models with different features, as we notice that the POS features outperform all the systems. The results confirm that the translation model is benefited from the existence of additional linguistic information.

MDP Lemma/morph as well as lemma Backoff models exhibit the best results in the reduction of OOV rates compared to all models. Although the tokenized models show better results in terms of OOV rates compared to some of the factored models, the effects of OOV reduction are not always well reflected by BLEU score [32]. Another important observation is that most OOV words are related to proper nouns that were not found in the training corpus. In this case, transliteration of named entities would be helpful to improve the translation quality. According to these results, factored models are the clear winners in all the scenarios we have presented.

To perform a manual analysis, we randomly selected sentences from the system output over the baseline, tokenized models and best-factored models as shown in Figure 5, and observed the following:

(1) Alleviate issues of dropping translations

The baseline and both tokenized models ignore or drop the translation of some words in the dataset, whereas this condition occurs at a lower rate in the factored models. This inclusion is due to the poor performance of SMT system for unobserved words in the training corpus, and worse performance (dropping to 48.6%) on words that were seen once [15].

Although Arabic morphological analyzer is helpful in MT, it provides complicated pre-processing making the system training cumbersome which may increase the incidence of the dropping problem.

Factored models choose to include most of these words taking advantage of the syntactic structures that enrich the training pipeline by good coherent sentences. Sentence 3 in Figure 5 shows an example where the words "eliminate those obstacles" were dropped by the baseline system, and in sentence 2 the words "have evolved over time" are dropped by both tokenized models. Unlike the baseline and tokenized models, factored models take those words into consideration.

(2) Overcome the sparseness problem

The tokenized and factored models help decrease the OOV rates, while the baseline model suffers from this problem thereby affects the performance of the translation results. We explain this to the morphological complexity and lexical diversity of Arabic, where the morphological analysis tool helps by providing a list of word-level analyses as well as splitting prefixes and suffixes, and removing diacritization that should increase the matching accuracy.

The use of linguistic features provides us new word forms such as lemma, improves the generalization ability and allows the decoder to use multiple translation options based on MDP or backoff model.

As an example, in sentence 3 of Figure 5 the diacritized word "plans" mapped to OOV by the baseline model, whereas removing the diacritization in the tokenized models helped the decoder to find the translation. In the lemma backoff model the decoder had two options to provide the output, which are surface-level and lemma-level phrase tables, during the translation, if the decoder didn't find the surface form, it goes back to its lemma form to gain the translation, and therefore decrease OOV rates.

(3) The improvement of D3* scheme

Since that Chinese doesn't have the definite article "the"; removing the definite article by the D3* model makes Arabic as a source language more similar to Chinese as a target language that improves the alignment and decreases the OOV rates which, likewise, gains a better performance. As in the examples of Figure 5, where removing "the" from the words "the state, the necessary" in sentence 1 and "the government" in sentence 3 enhanced the translation performance of D3* model.

(4) The advantages of linguistic factors

From the selected examples, we notice that the factored models obtained better quality and grammatical performance, due to the strength of the linguistic features which helps to characterize the words in sentence perfectly.

POS tags help to best exploit the data which considerably reduce the number of possible options, and promotes the ability of disambiguation for a token that has a different meaning in the context. Morph features improve generalization ability taking benefits from the rich knowledge for each token in the context. As for lemma Backoff and multiple decoding paths lemma/morph, both models show considerable results in handling the OOV issues, especially for the translation from a morphologically complex language as Arabic.

The evaluation results answer our empirical question that using linguistic factors on the Arabic side improves the quality of Arabic-to-Chinese translation. Further, different configurations of the factored model provide better translation quality compared with the baseline and better or comparable performance with both tokenized models.

Sentence 1	
Arabic	وتقوم الدولة بتهيئة الظروف اللازمة لممارسة هذا الحق. The State creates the necessary conditions for the exercise of this right.
Chinese-Ref	国家为每位公民创造必要的条件来行使该项权利。 The State creates the necessary conditions for every citizen to exercise this right.
Baseline-Output	该国行使其权利创造必要的条件。 This State exercises its rights to create the necessary conditions.
ATB Model-Output	国家和创造必要的条件下行使这种权利。 The State and the creation of the necessary conditions to exercise this right.
D3* Model-Output	该国所必需的条件下行使这种权利。 This right is exercised under the conditions necessary for this State.
POS Model-Output	国家正在创造必要的条件下行使这一权利。 The State is creating the necessary conditions to exercise this right.

Sentence 2	
Arabic	تطورت هذه الآليات بمرور الوقت. These mechanisms have evolved over time.
Chinese-Ref	这些机制随着时间推移而演变。 These mechanisms evolve over time.
Baseline-Output	随着时间的推移而这些机制。 These mechanisms over time.
ATB Model-Output	从这些机制。 From these mechanisms.
D3* Model-Output	这一机制。 This mechanism.
Morph Model-Output	随着时间的推移而发展这些机制。 These mechanisms are developed over time.

Sentence 3	
Arabic	ونتيجة لذلك، أعدت الحكومة خططاً وبرامج وطنية واتخذت تدابير للتغلب على هذه العقبات والقضاء عليها. As a result, the Government had prepared national plans and programs and had taken measures to overcome and eliminate those obstacles.
Chinese-Ref	因此，为克服并消除这些障碍，政府制定了相关的计划与规划，并采取了一些措施。 Therefore, in order to overcome and eliminate these obstacles, the Government has formulated relevant plans and programs, and has taken some measures.
Baseline-Output	因此, OOV 国家政府采取的措施和方案，以克服这些障碍。 Therefore, measures and programmes adopted by the OOV governments to overcome these obstacles.
ATB Model-Output	因此，政府和国家计划和方案采取了措施，克服这些障碍和消除。 Therefore, the Government and national plans and programmes have taken measures to overcome these obstacles and eliminate them.
D3* Model-Output	因此，政府制订了国家计划和方案措施，以克服这些障碍并消除。 Therefore, the Government has formulated national plans and measures to overcome these obstacles and eliminate them.
Lemma Backoff Model-Output	因此，政府制订了国家计划和方案采取了措施，以克服这些障碍并消除。 Therefore, the Government has formulated national plans and programs to take measures to overcome these obstacles and eliminate them.

Figure 5. Examples of Arabic-to-Chinese MT output over the baseline model, tokenized models, and best-factored models. The English glosses are presented to ease readability.

7. Conclusions and Future Work

Integrating additional linguistic knowledge is one of the core problems in the PBSMT model, which we addressed in this work. To investigate the benefits of linguistic input features for Arabic

→ Chinese MT, we compare a linguistically augmented MT model and PBSMT. We performed a preliminary evaluation of several deep linguistic features for Arabic, including lemmas, POS, and morph features. Several configurations were applied to evaluate the results of the factored model approach. We also empirically tested the contribution of various tokenization schemes to the PBSMT system, in addition to the effects of all models on reducing the data sparsity.

Our results show that using tokenization schemes for Arabic pre-processing helps to deal with the major issue of data spareness in the translation from Arabic as a morphologically rich and complex language. Linguistic factors, which we utilized to annotate the Arabic corpus, improved things even further. Factored systems achieved better performance compared with the baseline and tokenized PBSMT. The best system (POS model) yielded a BLEU score of 2.17 over the baseline and 0.63 over the tokenized phrase-based model, while lemma backoff model reduced the OOV rates from 5.74% to 0.70% for tokens, and from 22.80% to 6.71% for types.

To the best of our knowledge, this work is the first to test factored MT on the Arabic–Chinese language pair. Considering the following aspects can help improve this project: (1) using a big dataset to explore the linguistic input features on Arabic-to-Chinese with neural machine translation that has been proven useful for multiple languages; (2) adding factors to the target side language (Chinese); (3) testing other preprocessing tools that may perform better in Arabic–Chinese translation; and (4) conducting experiments on Chinese-to-Arabic translation that may reveal new insights.

Author Contributions: F.A. designed, wrote the paper and did the experiments; X.F. supervised the work and offered financial support; A.A. provided ideas to enrich the research, did the review and editing of the original draft; A.A.-M. did data analysis and surveyed the related works.

Funding: This research was supported by the National Natural Science Foundation of China under Grant No. 61876190.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Castilho, S.; Moorkens, J.; Gaspari, F.; Calixto, I.; Tinsley, J.; Way, A. Is Neural Machine Translation the New State of the Art? *Prague Bull. Math. Linguist.* **2017**, *108*, 109–120. [CrossRef]
2. Och, F.J.; Ney, H. The Alignment Template Approach to Statistical Machine Translation. *Comput. Linguist.* **2004**, *30*, 417–449. [CrossRef]
3. Mehay, D.N.; Brew, C. CCG syntactic reordering models for phrase-based machine translation. In Proceedings of the Seventh Workshop on Statistical Machine Translation, Montreal, QC, Canada, 7–8 June 2012; pp. 210–221.
4. Koehn, P.; Hoang, H. Factored translation models. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 28–30 June 2007; pp. 868–876.
5. Salerno, V.M.; Rabbeni, G. An Extreme Learning Machine Approach to Effective Energy Disaggregation. *Electronics* **2018**, *7*, 235. [CrossRef]
6. Siniscalchi, S.M.; Salerno, V.M. Adaptation to New Microphones Using Artificial Neural Networks With Trainable Activation Functions. *IEEE Trans. Neural. Netw. Learn. Syst.* **2017**, *28*, 1959–1965. [CrossRef] [PubMed]
7. Zhang, A.; Wang, H.; Li, S.; Cui, Y.; Liu, Z.; Yang, G.; Hu, J. Transfer Learning with Deep Recurrent Neural Networks for Remaining Useful Life Estimation. *Appl. Sci.* **2018**, *8*, 2416. [CrossRef]
8. Anastasopoulos, A.; Chiang, D. A case study on using speech-to-translation alignments for language documentation. In Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages, Honolulu, HI, USA, 6–7 March 2017.
9. Adda, G.; Stüker, S.; Adda-Decker, M.; Ambouroue, O.; Besacier, L.; Blachon, D.; Bonneau-Maynard, H.; Godard, P.; Hamlaoui, F.; Idiatov, D. Breaking the unwritten language barrier: The BULB project. In Proceedings of the SLTU (Spoken Language Technologies for Under-Resourced Languages), Yogyakarta, Indonesia, 9–12 May 2016; pp. 8–14.

10. Post, M.; Kumar, G.; Lopez, A.; Karakos, D.; Callison-Burch, C.; Khudanpur, S. Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus. In Proceedings of the IWSLT, Heidelberg, Germany, 5–6 December 2013.
11. Hasegawa-Johnson, M.A.; Jyothi, P.; McCloy, D.; Mirbagheri, M.; Liberto, G.M.D.; Das, A.; Ekin, B.; Liu, C.; Manohar, V.; Tang, H. ASR for Under-Resourced Languages from Probabilistic Transcription. *IEEE/ACM Trans. Audio Speech Lang.* **2017**, *25*, 50–63. [CrossRef]
12. Sadat, F.; Habash, N. Combination of Arabic preprocessing schemes for statistical machine translation. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 20 July 2006; pp. 1–8.
13. Habash, N. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Columbus, OH, USA, 16–17 July 2008; pp. 57–60.
14. Li, S.; Zhao, Z.; Hu, R.; Li, W.; Liu, T.; Du, X. Analogical reasoning on Chinese morphological and semantic relations. *arXiv*, 2018; arXiv:1805.06504.
15. Koehn, P.; Knowles, R. Six challenges for neural machine translation. *arXiv*, **2017**; arXiv:1706.03872.
16. Almahairi, A.; Cho, K.; Habash, N.; Courville, A. First result on Arabic neural machine translation. *arXiv*, 2016, arXiv:1606.02680.
17. Belinkov, Y.; Glass, J. Large-Scale Machine Translation between Arabic and Hebrew: Available Corpora and Initial Results. In Proceedings of the Workshop on Semitic Machine Translation, Austin, TX, USA, 1 November 2016; pp. 7–12.
18. Habash, N.; Hu, J. Improving Arabic-Chinese statistical machine translation using English as pivot language. In Proceedings of the Fourth Workshop on Statistical Machine Translation, Athens, Greece, 30–31 March 2009; pp. 173–181.
19. Shilon, R.; Habash, N.; Lavie, A.; Wintner, S. Machine Translation between Hebrew and Arabic. *Mach. Transl.* **2012**, *26*, 177–195. [CrossRef]
20. Ghurab, M.; Zhuang, Y.; Wu, J.; Abdullah, M.Y. Arabic-Chinese and Chinese-Arabic Phrase-Based Statistical Machine Translation Systems. *Inf. Technol. J.* **2010**, *9*, 666–672. [CrossRef]
21. Junczys-Dowmunt, M.; Dwojak, T.; Hoang, H. Is neural machine translation ready for deployment? A case study on 30 translation directions. In Proceedings of the IWSLT, Seattle, WA, USA, 8–9 December 2016.
22. Zalmout, N.; Habash, N. Optimizing Tokenization Choice for Machine Translation across Multiple Target Languages. *Prague Bull. Math. Linguist.* **2017**, *108*, 257–269. [CrossRef]
23. Hassan, H.; Sima'an, K.; Way, A. Syntactically Lexicalized Phrase-Based SMT. *IEEE Trans. Audio Speech Lang.* **2008**, *16*, 1260–1273. [CrossRef]
24. Avramidis, E.; Koehn, P. Enriching morphologically poor languages for statistical machine translation. In Proceedings of the ACL/HLT, Columbus, OH, USA, 15–20 June 2008; pp. 763–770.
25. Bhattacharyya, P. Role of Morphology Injection in Statistical Machine Translation. *arXiv*, **2017**, arXiv:1709.05487.
26. Kuntarič, S.; Krek, S.; Šikonja, M.R. Comparing Standard and Factored Models in Statistical Machine Translation from English to Slovene Using the Moses System. *Slov. 2.0 Empir. Appl. Interdiscip. Res.* **2018**, *5*, 1–26.
27. Durrani, N.; Haddow, B.; Heafield, K.; Koehn, P. Edinburgh's machine translation systems for European language pairs. In Proceedings of the Eighth Workshop on Statistical Machine Translation, Sofia, Bulgaria, 8–9 August 2013; pp. 114–121.
28. Badr, I.; Zbib, R.; Glass, J. Segmentation for English-to-Arabic statistical machine translation. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Columbus, OH, USA, 16–17 July 2008; pp. 153–156.
29. Rajeh, H.A.; Li, Z.; Ayedh, A.M. A Novel Approach by Injecting CCG Supertags into an Arabic–English Factored Translation Machine. *Arab. J. Sci. Eng.* **2016**, *41*, 3071–3080. [CrossRef]
30. Li, S.; Wong, D.F.; Chao, L.S. Korean-Chinese statistical translation model. In Proceedings of the 2012 International Conference on Machine Learning and Cybernetics (ICMLC), Xi'an, China, 15–17 July 2012; pp. 767–772.
31. Sennrich, R.; Haddow, B. Linguistic input features improve neural machine translation. In Proceedings of the First Conference on Machine Translation, Berlin, Germany, 11–12 August 2016.

32. García-Martínez, M.; Barrault, L.; Bougares, F. Factored neural machine translation. *arXiv*, **2016**, arXiv:1609.04621.
33. Du, J.; Way, A. Pre-Reordering for Neural Machine Translation: Helpful or Harmful? *Prague Bull. Math. Linguist.* **2017**, *108*, 171–182. [CrossRef]
34. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Prague, Czech Republic, 25–27 June 2007; pp. 177–180.
35. Pasha, A.; Al-Badrashiny, M.; Diab, M.T.; El Kholy, A.; Eskander, R.; Habash, N.; Pooleery, M.; Rambow, O.; Roth, R. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In Proceedings of the LREC, Reykjavik, Iceland, 26–31 May 2014; pp. 1094–1101.
36. Sajjad, H.; Guzmán, F.; Nakov, P.; Abdelali, A.; Murray, K.; Al Obaidli, F.; Vogel, S. QCRI at IWSLT 2013: Experiments in Arabic-English and English-Arabic spoken language translation. In Proceedings of the 10th International Workshop on Spoken Language Technology (IWSLT), Heidelberg, Germany, 5–6 December 2013.
37. Habash, N.; Soudi, A.; Buckwalter, T. On Arabic Transliteration. In *Arabic Computational Morphology*; Springer: Berlin, Germany, 2007; pp. 15–22.
38. Aldarmaki, H.; Diab, M. Robust part-of-speech tagging of Arabic text. In Proceedings of the Second Workshop on Arabic Natural Language Processing, Beijing, China, 30 July 2015; pp. 173–182.
39. Birch, A.; Osborne, M.; Koehn, P. CCG supertags in factored statistical machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 23 June 2007; pp. 9–16.
40. Eisele, A.; Chen, Y. MultiUN: A Multilingual Corpus from United Nation Documents. In Proceedings of the LREC, Valletta, Malta, 17–23 May 2010.
41. Chang, P.-C.; Galley, M.; Manning, C.D. Optimizing Chinese word segmentation for machine translation performance. In Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, OH, USA, 19 June 2008; pp. 224–232.
42. Tseng, H.; Chang, P.; Andrew, G.; Jurafsky, D.; Manning, C. A conditional random field word segmenter for sighan bakeoff 2005. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea, 14–15 October 2005.
43. Och, F.J.; Ney, H. A Systematic Comparison of Various Statistical Alignment Models. *Comput. Linguist.* **2003**, *29*, 19–51. [CrossRef]
44. Heafield, K.; Pouzyrevsky, I.; Clark, J.H.; Koehn, P. Scalable modified Kneser-Ney language model estimation. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria, 4–9 August 2013; pp. 690–696.
45. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA, 7–12 July 2002; pp. 311–318.
46. Koehn, P. Statistical significance tests for machine translation evaluation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, Spain, 25–26 July 2004; pp. 388–395.