Morphological Constraints for Phrase Pivot Statistical Machine Translation

Ahmed El Kholy ame2127@columbia.edu
Center for Computational Learning Systems, Columbia University

Nizar Habash nizar.habash@nyu.edu
Computer Science, New York University Abu Dhabi

Abstract

The lack of parallel data for many language pairs is an important challenge to statistical machine translation (SMT). One common solution is to pivot through a third language for which there exist parallel corpora with the source and target languages. Although pivoting is a robust technique, it introduces some low quality translations especially when a poor morphology language is used as the pivot between rich morphology languages. In this paper, we examine the use of synchronous morphology constraint features to improve the quality of phrase pivot SMT. We compare hand-crafted constraints to those learned from limited parallel data between source and target languages. The learned morphology constraints are based on projected alignments between the source and target phrases in the pivot phrase table. We show positive results on Hebrew-Arabic SMT (pivoting on English). We get 1.5 BLEU points over a phrase pivot baseline and 0.8 BLEU points over a system combination baseline with a direct model built from parallel data.

1 Introduction

One of the main challenges in statistical machine translation (SMT) is the scarcity of parallel data for many language pairs especially when the source and target languages are morphologically rich. A common SMT solution to the lack of parallel data is to pivot the translation through a third language (called pivot or bridge language) for which there exist abundant parallel corpora with the source and target languages. The literature covers many pivoting techniques. One of the best performing techniques, phrase pivoting (Utiyama and Isahara, 2007), builds an induced new phrase table between the source and target. One of the main issues of this technique is that the size of the newly created pivot phrase table is very large. Moreover, many of the produced phrase pairs are of low quality which affects the translation choices during decoding and the overall translation quality.

In this paper, we focus on improving phrase pivoting. We introduce morphology constraint scores which are added to the log linear space of features in order to determine the quality of the pivot phrase pairs. We compare two methods of generating the morphology constraints. One method is based on hand-crafted rules relying on the authors knowledge of the source and target languages; while in the other method, the morphology constraints are induced from

available parallel data between the source and target languages which we also use to build a direct translation model. We then combine both the pivot and direct models to achieve better coverage and overall translation quality. We show positive results on Hebrew-Arabic SMT. We get 1.5 BLEU points over a phrase-pivot baseline and 0.8 BLEU points over a system combination baseline with a direct model built from given parallel data.

Next, we briefly discuss some related work. In Section 3, we review the best performing pivoting strategy and how we use it. In Section 4, we discuss the linguistic differences among Hebrew, Arabic, and the pivot language, English. This is followed by our approach to using morphology constraints in Section 5. We finally present our experimental results in Section 6 and a case study in Section 7.

2 Related Work

Many researchers have investigated the use of pivoting (or bridging) approaches to solve the data scarcity issue (Utiyama and Isahara, 2007; Wu and Wang, 2009; Khalilov et al., 2008; Bertoldi et al., 2008; Habash and Hu, 2009). The main idea is to introduce a pivot language, for which there exist large source-pivot and pivot-target bilingual corpora. Pivoting has been explored for closely related languages (Hajič et al., 2000) as well as unrelated languages (Koehn et al., 2009; Habash and Hu, 2009). Many different pivot strategies have been presented in the literature. The following three are the most common. The first strategy is the sentence translation technique in which we first translate the source sentence to the pivot language, and then translate the pivot language sentence to the target language (Khalilov et al., 2008). The second strategy is based on phrase pivoting (Utiyama and Isahara, 2007; Cohn and Lapata, 2007; Wu and Wang, 2009). In phrase pivoting, a new source-target phrase table (translation model) is induced from source-pivot and pivot-target phrase tables. Lexical weights and translation probabilities are computed from the two translation models. The third strategy is to create a synthetic source-target corpus by translating the pivot side of source-pivot corpus to the target language using an existing pivot-target model (Bertoldi et al., 2008). In this paper, we use the phrase pivoting approach, which has been shown to be the best with comparable settings (Utiyama and Isahara, 2007).

There has been recent efforts in improving phrase pivoting. One effort focused on improving alignment symmetrization targeting pivot phrase systems (El Kholy and Habash, 2014). In another recent effort, Multi-Synchronous Context-free Grammar (MSCFG) is leveraged to triangulate source-pivot and pivot-target synchronous Context-free Grammar (SCFG) rule tables into a source-target-pivot MSCFG rule table that helps in remembering the pivot during decoding. Also, pivot LMs are used to assess the naturalness of the derivation (Miura et al., 2015).

In our own previous work, we demonstrated quality improvement using connectivity strength features between the source and target phrase pairs in the pivot phrase table (El Kholy et al., 2013). These features provide quality scores based on the number of alignment links between words in the source phrase to words of the target phrase. In this work, we extend on the connectivity scores with morphological constraints through which we provide quality scores based on the morphological compatibility between the connected/aligned source and target words.

Since both Hebrew and Arabic are morphologically rich, we should mention that there has been a lot of work on translation to and from morphologically rich languages (Yeniterzi and Oflazer, 2010; Elming and Habash, 2009; El Kholy and Habash, 2010; Habash and Sadat, 2006; Kathol and Zheng, 2008). Most of these efforts are focused on syntactic and morphological processing to improve the quality of translation.

Until recently, there has not been much parallel Hebrew-English and Hebrew-Arabic data (Tsvetkov and Wintner, 2010), and consequently little work on Hebrew-English and Hebrew-

Translation Model	Training Corpora Size	Phrase Table	
		# Phrase Pairs	Size
Hebrew-English	≈1M words	3,002,887	327MB
English-Arabic	≈60M words	111,702,225	14GB
Pivot_Hebrew-Arabic	N/A	$>$ 30 Billion	≈2.5TB

Table 1: Translation Models Phrase Table comparison in terms of number of lines and sizes.

Arabic SMT. Lavie et al. (2004) built a transfer-based translation system for Hebrew-English and so did Shilon et al. (2012) for translation between Hebrew and Arabic. Our previous work discussed above (El Kholy et al., 2013) was demonstrated on Hebrew-Arabic with English pivoting.

3 Phrase Pivoting

In this section, we review the phrase pivoting strategy in detail as we describe how we built our baseline for Arabic-Hebrew via pivoting on English. We also discuss how we overcome the large expansion of source-to-target phrase pairs in the process of creating a pivot phrase table. In phrase pivoting (which is sometimes called triangulation or phrase table multiplication), we train a Hebrew-Arabic and an English-Arabic translation models, such as those used in the sentence pivoting technique. Based on these two models, we induce a new Hebrew-Arabic translation model. Since our models are based on a Moses phrase-based SMT system (Koehn et al., 2007), we use the standard set of phrase-based translation probability distributions.[1] We follow Utiyama and Isahara (2007) in computing the pivot phrase pair probabilities. The following are the set of equations used to compute the lexical probabilities (p_w) and the phrase translation probabilities (ϕ):

$$\phi(h|a) = \sum_e \phi(h|e)\phi(e|a)$$
$$\phi(a|h) = \sum_e \phi(a|e)\phi(e|h)$$
$$p_w(h|a) = \sum_e p_w(h|e)p_w(e|a)$$
$$p_w(a|h) = \sum_e p_w(a|e)p_w(e|h)$$

Above, h is the Hebrew source phrase; e is the English pivot phrase that is common in both Hebrew-English translation model and English-Arabic translation model; and a is the Arabic target phrase. We also build a Hebrew-Arabic reordering table using the same technique but we compute the reordering probabilities in a similar manner to Henriquez et al. (2010).

Filtering for Phrase Pivoting As discussed earlier, the induced Hebrew-Arabic phrase and reordering tables are very large. Table 1 shows the amount of parallel corpora used to train the Hebrew-English and the English-Arabic and the equivalent phrase table sizes compared to the induced Hebrew-Arabic phrase table.[2] We follow the work of El Kholy et al. (2013) and filter the phrase pairs used in pivoting based on log-linear scores. The main idea of the filtering process is to select the top [n] English candidate phrases for each Hebrew phrase from the Hebrew-English phrase table and similarly select the top [n] Arabic target phrases for each English phrase from the English-Arabic phrase table and then perform the pivoting process described earlier to create a pivoted Hebrew-Arabic phrase table. To select the top candidates, we

[1] Four different phrase translation scores are computed in Moses' phrase tables: two lexical weighting scores and two phrase translation probabilities.

[2] The size of the induced phrase table size is computed but not created.

first rank all the candidates based on the log linear scores computed from the phrase translation probabilities and lexical weights multiplied by the optimized decoding weights then we pick the top [n] pairs. In our experiments, we pick the top 1000 pairs for pivoting.

4 Linguistic Comparison

In this section we present the challenges of preprocessing Arabic, Hebrew, and English, and how we address them. Both Arabic and Hebrew are morphologically complex languages. One aspect of Arabic's complexity is its various attachable clitics and numerous morphological features (Habash, 2010). Clitics include conjunction proclitics, e.g., +و *w*+[3] 'and', prepositional proclitics, e.g., +ل *l*+ 'to/for', the definite article +ال *Al*+ 'the', and the class of pronominal enclitics, e.g., +هم *hm*+ 'their/them'. All of these clitics are separate words in English. Beyond the clitics, Arabic words inflect for person, gender, number, aspect, mood, voice, state and case. Additionally, Arabic orthography uses optional diacritics for short vowels and consonant doubling. This, together with Arabic's morphological richness, leads to a high degree of ambiguity: about 12 analyses per word, typically corresponding to two lemmas on average (Habash, 2010). We follow El Kholy and Habash (2010) and use the PATB tokenization scheme (Maamouri et al., 2004) in our experiments. The PATB scheme separates all clitics except for the determiner clitic *Al*+(DET). We use MADA v3.1 (Habash and Rambow, 2005; Habash et al., 2009) to tokenize the Arabic text. We only evaluate on detokenized and orthographically correct (enriched) output following the work of El Kholy and Habash (2010).

Similar to Arabic, Hebrew poses computational processing challenges typical of Semitic languages (Itai and Wintner, 2008; Shilon et al., 2012). Hebrew orthography also uses optional diacritics and its morphology inflects for gender, number, person, state, tense and definiteness. Furthermore, Similar to Arabic, Hebrew has a set of attachable clitics, e.g., conjunctions (such as +ו *w*+[4] 'and'), prepositions (such as +ב *b*+ 'in'), the definite article (+ה *h*+ 'the'), or pronouns (such as +הם *hm*+ 'their'). These issues contribute to a high degree of ambiguity that is a challenge to translation from Hebrew to English or to any other language. We follow Singh and Habash (2012)'s best preprocessing setup which utilized a Hebrew tagger (Adler, 2007) and produced a tokenization scheme that separated all clitics.

English, our pivot language, is quite different from both Arabic and Hebrew. English is poor in morphology and barely inflects for number and tense, and for person in a limited context. English preprocessing simply includes down-casing, separating punctuation and splitting off "'s".

5 Approach

One of the main challenges in phrase pivoting is the very large size of the induced phrase table. It becomes even more challenging if either the source or target language is morphologically rich. The number of translation candidates (fanout) increases due to ambiguity and richness which in return increases the number of combinations between source and target phrases. Since the only criteria of matching between the source and target phrase is through a pivot phrase, many of the induced phrase pairs are of low quality. These phrase pairs unnecessarily increase the search space and hurt the overall quality of translation. A basic solution to the combinatorial expansion is to filter the phrase pairs used in pivoting based on log-linear scores as discussed in Section 3, however, this doesn't solve the low quality problem.

[3] Arabic transliteration throughout the paper is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007).

[4] The following Hebrew 1-to-1 transliteration is used (in Hebrew alphabetical order): *abgdhwzxTiklmns'pcqršt*. All examples are undiacritized and final forms are not distinguished from non-final forms.

Similar to factored translation models (Koehn and Hoang, 2007) where linguistic (morphology) features are augmented to the translation model to improve the translation quality, our approach to address the quality problem is based on constructing a list of synchronous morphology constraints between the source and target languages. These constraints are used to generate scores to determine the quality of pivot phrase pairs. However, unlike factored models, we do not use the morphology in generation and the morphology information comes completely from external resources. In addition, since we work in the pivoting space, we only apply the morphology constraints to the connected words between the source and target languages through the pivot language. This guarantees a fundamental level of semantic equivalence before applying the morphology constraints especially if there is distortion between source and target phrases.

We build on our approach in El Kholy et al. (2013) where we introduced connectivity strength features between the source and target phrase pairs in the pivot phrase table. These features provide quality scores based on the number of alignment links between words in the source phrase and words in the target phrase. The alignment links are generated by projecting the alignments of the source-pivot phrase pairs and the pivot-target phrase pairs used in pivoting. We use the same concept but instead of using the lexical mapping between source and target words, we compute quality scores based on the morphological compatibility between the connected source and target words.

To choose which morphological features to work with, we performed an automatic error analysis on the output of the phrase-pivot baseline system. We did the analysis using AMEANA (El Kholy and Habash, 2011), an open-source error analysis tool for natural language processing tasks targeting morphologically rich languages. We found that the most problematic morphological features in the Arabic output are gender (GEN), number (NUM) and determiner (DET). We focus on those features in addition to (POS) in our experiments.

Next, we present our approach to generating the morphology constraint features using hand-crafted rules and compare this approach with inducing these constraints from Hebrew-Arabic parallel data.

5.1 Rule-based Morphology Constraints

Our rule-based morphology constraint features are basically a list of hand-crafted mappings of the different morphological features between Hebrew and Arabic. Since both languages are morphologically rich as explained in Section 4, it is straightforward to produce these mappings for GEN, NUM and DET. Note, however, that we also account for ambiguous cases; e.g., feminine gender in Arabic can map to words with ambiguous gender in Hebrew. We additionally use different POS tag sets for Arabic (47 tags) and Hebrew (25 tags) and in many cases one Hebrew tag can map to more than one Arabic tag; for example, three Arabic noun tags *abbrev, noun* and *noun_prop* map to two Hebrew tags *feminine, masculine* noun.[5] Table 2 shows a sample of the morphological mappings between Arabic and Hebrew.

After building the morphological features mappings, we use them to judge the quality of a given phrase pair in the phrase pivot model. We add two scores W_s and W_t to the log linear space. Given a *source-target* phrase pair $\bar{s}, \bar{t}$ and a word projected alignment a between the source word positions $i = 1, ..., n$ and the target word positions $j = 1, ..., m$, W_s and W_t are defined in equations 1 and 2. F is the set of morphological features (we focus on GEN, NUM, DET and POS). M_f is the hand-crafted rules mapping between Arabic and Hebrew feature values of feature $f \in F$. In case of ambiguity for a given feature; for example, a word's gender being masculine or feminine, we use the maximum likelihood value of this feature given the word. $MLE_f(i)$ is the maximum likelihood feature value of feature f for the source word at

[5]Please refer to (Habash et al., 2009) for a complete set of Arabic POS tag set and (Adler, 2007) for Hebrew POS tag set.

Features	Arabic	Hebrew
GEN	Feminine	Feminine / Both
	Masculine	Masculine / Both
NUM	Singluar	Singluar / Singluar-Plural
	Dual	Dual / Dual-Plural
	Plural	Plural / Dual-Plural / Singular-Plural
DET	No Determiner	No Determiner
	Determiner	Determiner

Table 2: Rule-based mapping between Arabic and Hebrew morphological features. Each feature value in Arabic can map to more than one feature value in Hebrew.

position i, and $MLE_f(j)$ is the maximum likelihood feature value of feature f for the target word at position j. The maximum likelihood feature values for Hebrew were computed from the Hebrew side of the training data. As for Arabic, the maximum likelihood feature values were computed from the Arabic side of the training data in addition to Arabic Gigaword corpus, which was used in creating the language model (more details in Section 6.1).

$$W_s = \frac{1}{|F|} \sum_{\forall f \in F} \sum_{\forall (i,j) \in a} \frac{1}{n} [(MLE_f(i), MLE_f(j)) \in M_f] \quad (1)$$

$$W_t = \frac{1}{|F|} \sum_{\forall f \in F} \sum_{\forall (i,j) \in a} \frac{1}{m} [(MLE_f(i), MLE_f(j)) \in M_f] \quad (2)$$

5.2 Induced Morphology Constraints

In this section, we explain our approach in generating morphology constraint features from a given parallel data between source and target languages. Unlike the rule-based approach we build a translation model between the source and target morphological features and we use the morphology translation probabilities as metric to judge a given phrase pair in the pivot phrase table. For the automatically induced constraints, we jointly model mapping between conjunctions of features attached to aligned words rather than tallying each feature match independently. Writing good manual rules for such feature conjunction mappings would be more difficult. Table 3 shows some examples of mapping (GEN), number (NUM) and determiner (DET) in Hebrew to their equivalent in Arabic and their respective bi-directional scores.

Hebrew (H)	Arabic (A)	$P_{FC}(A\|H)$	$P_{FC}(H\|A)$
[Fem+Dual+Det]	[Fem+Dual]	0.0006	0.0833
[Fem+Dual+Det]	[Fem+Dual+Det]	0.0148	0.3333
[Fem+Dual+Det]	[Fem+Singular] [Fem+Dual]	0.0052	0.0833
[Fem+Dual+Det]	[Masc+Dual+Det]	0.0047	0.5000

Table 3: Examples of induced morphology constraints for (GEN), number (NUM) and determiner (DET) and their respective scores.

As in rule-based approach, we add two scores W_s and W_t to the log linear space which are defined in equations 3 and 4. P_{FC} is the conditional morphology probability of a given feature

combination (FC) value. Similar to rule-based morphology constraints, we resort to the maximum likelihood value of a feature combination when the values are ambiguous. $MLE_{FC}(i)$ is the maximum likelihood feature combination (FC) value for the source word at position i while $MLE_{FC}(j)$ is the maximum likelihood feature combination (FC) value for the target word at position j.

$$W_s = \frac{1}{n} \sum_{\forall(i,j)\in a} P_{FC}(MLE_{FC}(i)|MLE_{FC}(j)) \tag{3}$$

$$W_t = \frac{1}{m} \sum_{\forall(i,j)\in a} P_{FC}(MLE_{FC}(j)|MLE_{FC}(i)) \tag{4}$$

5.3 Model Combinations

Since we use parallel data to induce the morphology constraints, it would make sense to measure the effect of combining (a) the pivot model with added morphology constraints, and (b) the direct model trained on the parallel data used to induce the morphology constraints. We perform the combination using Moses' phrase table combination techniques. Translation options are collected from one table, and additional options are collected from the other tables. If the same translation option (in terms of identical input phrase and output phrase) is found in multiple tables, separate translation options are created for each occurrence, but with different scores (Koehn and Schroeder, 2007). We show results over a learning curve in Section 6.5.

6 Experiments

In this section, we present a set of experiments comparing the use of rule-based versus induced morphology constraint features in phrase-pivot SMT as well as model combination to improve Hebrew-Arabic pivot translation quality.

6.1 Experimental Setup

In our pivoting experiments, we build two SMT models; one model to translate from Hebrew to English, and another model to translate from English to Arabic. The English-Arabic parallel corpus is about ($\approx$ 60M words) and is available from LDC[6] and GALE[7] constrained data. The Hebrew-English corpus is about ($\approx$ 1M words) and is available from sentence-aligned corpus produced by Tsvetkov and Wintner (2010). For the direct Hebrew-Arabic SMT model, we use a TED parallel corpus of about ($\approx$ 2M words) (Cettolo et al., 2012).

Word alignment is done using GIZA++ (Och and Ney, 2003). For Arabic language modeling, we use 200M words from the Arabic Gigaword Corpus (Graff, 2007) together with the Arabic side of our training data. We use 5-grams for all language models (LMs) implemented using the SRILM toolkit (Stolcke, 2002).

All experiments are conducted using the Moses phrase-based SMT system (Koehn et al., 2007). We use MERT (Och, 2003) for decoding weight optimization. Weights are optimized using a tuning set of 517 sentences developed by Shilon et al. (2010).

We use a maximum phrase length of size 8 across all models. We report results on a Hebrew-Arabic development set (Dev) of 500 sentence with a single reference and an evaluation set (Test) of 300 sentences with three references developed by Shilon et al. (2010). We evaluate using BLEU-4 (Papineni et al., 2002).

[6]LDC Catalog IDs: LDC2005E83, LDC2006E24, LDC2006E34, LDC2006E85, LDC2006E92, LDC2006G05, LDC2007E06, LDC2007E101, LDC2007E103, LDC2007E46, LDC2007E86, LDC2008E40, LDC2008E56, LDC2008G05, LDC2009E16, LDC2009G01.

[7]Global Autonomous Language Exploitation, or GALE, was a DARPA-funded research project.

6.2 Baselines

We compare the performance of adding the connectivity strength features (+*Conn*) to the phrase pivoting SMT model (*Phrase_Pivot*) and building a direct SMT model using all parallel He-Ar corpus available. The results are presented in Table 4. Consistently with our previous effort (El Kholy et al., 2013), the performance of the phrase-pivot model improves with the connectivity strength features. While the direct system is better than the phrase pivot model in general, the combination of both models leads to a high performance gain of 1.7/4.4 BLEU points in Dev/Test over the best performers of both the direct and phrase-pivot models.

Model	**Dev**	**Test**
Direct	9.7	20.4
Phrase_Pivot	8.3	19.8
Phrase_Pivot+Conn	9.1	20.1
Direct+Phrase_Pivot+Conn	**11.4**	**24.5**

Table 4: Comparing phrase pivoting SMT with connectivity strength features, direct SMT and the model combination. The results show that the best performer is the model combination in Dev and Test sets.

6.3 Rule-based Morphology Constraints

Model	**Dev**		**Test**	
	Single	**Combined**	**Single**	**Combined**
Direct	**9.7**	n/a	20.4	n/a
Phrase_Pivot+Conn	9.1	11.4	20.1	24.5
Phrase_Pivot+Conn+Morph_Rules	**9.6**	12.2*	20.9*	24.6
Phrase_Pivot+Conn+Morph_Auto	**9.6**	**12.4***	**21.6***	**25.3***

Table 5: Morphology constraints results. The "Single" columns show the results of a single model of either the direct model or the phrase pivoting models with additional morphological constraints features. The "Combined" show the results of system combination between the direct model and the different phrase pivoting models. In the first row, the "Combined" results are not applicable for the direct model. (*) marks a statistically significant result against both the direct and phrase-pivot baseline.

In this experiment, we show the performance of adding hand-crafted morphology constraints (+*Morph_Rules*) to determine the quality of a given phrase pair in the phrase-pivot translation model. The third row in Table 5 shows that although the rules are based on a one-to-one mapping between the different morphological features, the translation quality is improved over the baseline phrase-pivot model by 0.5/0.8 BLEU points in Dev/Test sets.

As expected, the system combination of the pivot model with the direct model improves the overall performance but the gain we get from the morphology constraints only appears in the Dev set with 0.8 BLEU points, and not much in the Test set.

6.4 Induced Morphology Constraints

In this experiment, we measure the effect of using induced morphology constraints (+*Morph_Auto*) on MT quality. The last row in Table 5 shows that the induced morphology constraints improve the results over the baseline phrase-pivot model by 0.5/1.5 BLEU points in

Parallel Data Size	Model	Dev		Test	
		Single	**Combined**	**Single**	**Combined**
125K	Direct	2.7	n/a	8.4	n/a
	Phrase_Pivot+Conn	9.1	10.4	20.1	20.9
	Phrase_Pivot+Conn+Morph_Auto	9.2	10.6	20.6	21.3
500K	Direct	5.9	n/a	15.1	n/a
	Phrase_Pivot+Conn	9.1	10.7	20.1	22.5
	Phrase_Pivot+Conn+Morph_Auto	9.7	11.2	20.8	22.8
2M	Direct	**9.7**	n/a	20.4	n/a
	Phrase_Pivot+Conn	9.1	11.4	20.1	24.5
	Phrase_Pivot+Conn+Morph_Auto	**9.6**	**12.4**	**21.6**	**25.3**

Table 6: Learning curve results of 100% (2M words), 25% (500K words) and 6.25% (125K words) of the parallel Hebrew-Arabic corpus.

Dev/Test sets and over the Rule-based morphology constraints by 0.7 BLEU points in the Test set.

Similar to the Rule-based constraints, the performance did not improve compared to the *direct model* in the Dev set; but, again, the Test set showed a great improvement of 1.5 and 1.2 BLEU points over the pivot and direct models, respectively. Also the system combination of the pivot model with the direct model improves the overall performance. The model using induced morphological features is the best performer with an increase in the performance gain by 1.0/0.8 BLEU points in Dev/Test sets. This shows that the benefit we get from the induced morphology constraints were not diluted when we do the model combination given the fact that the constraints were induced from the parallel data to start with.

It is important to note here that the induced morphology constraints outperformed the rule-based constraints across all settings. This shows that the complex morphology constraints extracted from the parallel data provide knowledge that can not be covered by simple linguistic rules. However, the simple rule-based approach comes in handy when there is no data between the source and target languages.

6.5 Learning Curve

In this experiment, we examine the effect of using less data in inducing morphology constraints rules and the overall performance when we combine systems. Table 6 shows the results on a learning curve of 100% (2M words), 25% (500K words) and 6.25% (125K words) of the parallel Hebrew-Arabic corpus.

As expected, The system combination between the direct translation models and the phrase-pivot translation model leads to an improvement in the translation quality across the learning curve even when there is small amount of parallel corpora. Despite the weak performance (2.7 BLEU) of the direct system built on 6.25% of the parallel Hebrew-Arabic corpus, the system combination leads to 1.4 BLEU points gain.

An interesting observation from the results is that we always get a performance gain from the induced morphology constrains across all settings. This shows that the system combination helps in adding more lexical translation choices while the constraints help in a different dimension, which is selecting the best phrase pairs from the pivot system.

Hebrew Source	המתווכים והסוחרים מסרבים לדבר בפומבי על המחירים. *the+middlemen and+the+traders refuse[m.p.] to+speak publicly about the+prices*
Arabic Reference	يرفض الوسطاء والتجار الحديث علنا عن الاسعار *refuse[m.s.] the+middlemen and+the+traders the+speaking publicly about the+prices*
Phrase_Pivot+Conn	وسطاء והסוחרים يرفض التحدث علنا عن الاسعار *middlemen* והסוחרים *refuse[m.s.] the+speaking publicly about the+prices*
Direct	המתווכים والتجار رفضوا الحديث على الملأ على الاسعار המתווכים *and+the+traders refused[m.p.] the+speaking upon the+public about the+prices*
Phrase_Pivot+Conn+ Morph_Auto	الوسطاء והסוחרים يرفضون التحدث علنا عن الاسعار *the+middlemen* והסוחרים *refuse[m.p.] the+speaking publicly about the+prices*
Direct+Phrase_Pivot+ Conn+Morph_Auto	وسطاء والتجار رفضوا التحدث علنا عن الاسعار *middlemen and+the+traders refused[m.p.] the+speaking publicly about the+prices*

Table 7: Translation examples.

7 Case Study

In this section we consider an example from our Dev set that captures many of the patterns and themes in the evaluation. Table 7 shows a Hebrew source sentence and its Arabic reference. This is followed by the output from the pivot system, the direct system, the Phrase_Pivot+Conn+Morph_Auto system and the combined system.

Two particular aspects should be noted. First is the complementary lexical coverage of the direct and pivot systems. This is seen in how one of each covers half of the phrase *middlemen and traders*. The combined system captures both. Secondly, the gender, number and tense of the main verb prove challenging in many ways (and this is an issue for a majority of the sentences in the Dev set). The Hebrew verb in the present tense is masculine and plural; and naturally follows the subject. The Arabic reference verb appears at the beginning of the sentence, in which location it only agrees with the subject in gender (while number is singular). Arabic Verbs in SVO order agree in gender and number. All the MT systems we compare leave the verb after the subject. The direct, Phrase_Pivot+Conn+Morph_Auto, and combination systems get the number and gender correctly; however, the direct and combined system make the verb tense past. The Phrase_Pivot+Conn+Morph_Auto example highlights the value of morphology constraints; but the example points out that they sometimes are hard to evaluate automatically, since there are morphosyntactically allowable forms that do not match the translation references.

8 Conclusion and Future Work

In this paper, we presented the use of synchronous morphology constraint features based on hand-crafted rules compared to rules induced from parallel data to improve the quality of phrase-pivot based SMT. We show that the two approaches lead to an improvement in the translation quality. The induced morphology constraints approach is a better performer, however, it relies on the fact there is a parallel corpus between source and target languages. We show positive results on Hebrew-Arabic SMT. We get 1.5 BLEU points over phrase-pivot baseline and 0.8 BLEU points over system combination baseline with direct model built from given parallel data.

In the future, we plan to work on reranking experiments as a post-translation step based on morphosyntactic information between source and target languages. We also plan to work on word reordering between morphologically rich language to maintain the relationship between the word order and the morphosyntactic agreement in the context of phrase pivoting.

Acknowledgments

The work presented in this paper was possible thanks to a generous Google Research Award. We would like to thank Reshef Shilon and Shuly Winter for helpful discussions and support with processing Hebrew. We also thank the anonymous reviewers for their insightful comments.

References

Adler, M. M. (2007). *Hebrew morphological disambiguation: An unsupervised stochastic word-based approach*. PhD thesis, Ben-Gurion University of the Negev.

Bertoldi, N., Barbaiani, M., Federico, M., and Cattoni, R. (2008). Phrase-based statistical machine translation with pivot languages. *Proceeding of IWSLT*, pages 143–149.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Cohn, T. and Lapata, M. (2007). Machine translation by triangulation: Making effective use of multi-parallel corpora. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 728.

El Kholy, A. and Habash, N. (2010). Orthographic and Morphological Processing for English-Arabic Statistical Machine Translation. In *Proceedings of Traitement Automatique du Langage Naturel (TALN-10)*. Montréal, Canada.

El Kholy, A. and Habash, N. (2010). Techniques for Arabic Morphological Detokenization and Orthographic Denormalization. In *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

El Kholy, A. and Habash, N. (2011). Automatic Error Analysis for Morphologically Rich Languages. In *MT Summit XIII*.

El Kholy, A. and Habash, N. (2014). Alignment symmetrization optimization targeting phrase pivot statistical machine translation. In *Proceedings of the 17th annual conference of the European Association for Machine Translation, EAMT 2014*.

El Kholy, A., Habash, N., Leusch, G., Matusov, E., and Sawaf, H. (2013). Language independent connectivity strength features for phrase pivot statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Sofia, Bulgaria. Association for Computational Linguistics.

Elming, J. and Habash, N. (2009). Syntactic Reordering for English-Arabic Phrase-Based Machine Translation. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, pages 69–77, Athens, Greece.

Graff, D. (2007). Arabic Gigaword 3, LDC Catalog No.: LDC2003T40. Linguistic Data Consortium, University of Pennsylvania.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Habash, N. and Hu, J. (2009). Improving Arabic-Chinese Statistical Machine Translation using English as Pivot Language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181, Athens, Greece.

Habash, N. and Rambow, O. (2005). Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan.

Habash, N., Rambow, O., and Roth, R. (2009). MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In Choukri, K. and Maegaard, B., editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*. The MEDAR Consortium.

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

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In van den Bosch, A. and Soudi, A., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Hajič, J., Hric, J., and Kubon, V. (2000). Machine Translation of Very Close Languages. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP'2000)*, pages 7–12, Seattle.

Henriquez, C., Banchs, R. E., and Mariño, J. B. (2010). Learning reordering models for statistical machine translation with a pivot language.

Itai, A. and Wintner, S. (2008). Language resources for Hebrew. *Language Resources and Evaluation*, 42(1):75–98.

Kathol, A. and Zheng, J. (2008). Strategies for building a Farsi-English smt system from limited resources. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH'2008)*, pages 2731–2734, Brisbane, Australia.

Khalilov, M., Costa-jussá, M. R., Fonollosa, J. A. R., Banchs, R. E., Chen, B., Zhang, M., Aw, A., Li, H., Mariño, J. B., Hernández, A., and Q., C. A. H. (2008). The talp & i2r smt systems for iwslt 2008. In *International Workshop on Spoken Language Translation. IWSLT 2008, pg. 116–123.*

Koehn, P., Birch, A., and Steinberger, R. (2009). 462 machine translation systems for europe. *Proceedings of MT Summit XII*, pages 65–72.

Koehn, P. and Hoang, H. (2007). Factored translation models. In *EMNLP-CoNLL*, pages 868–876.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227. Association for Computational Linguistics.

Lavie, A., Probst, K., Peterson, E., Vogel, S., Levin, L., Font-Llitjos, A., and Carbonell, J. (2004). A trainable transfer-based machine translation approach for languages with limited resources. In *Proceedings of European Association for Machine Translation Workshop on Broadening horizons of machine translation and its applications*, Malta.

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (2004). The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Miura, A., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2015). Improving pivot translation by remembering the pivot. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 573–577, Beijing, China. Association for Computational Linguistics.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–52.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Shilon, R., Habash, N., Lavie, A., and Wintner, S. (2010). Machine translation between hebrew and arabic: Needs, challenges and preliminary solutions. In *Proceedings of AMTA*.

Shilon, R., Habash, N., Lavie, A., and Wintner, S. (2012). Machine translation between Hebrew and Arabic. *Machine Translation*, 26:177–195.

Singh, N. and Habash, N. (2012). Hebrew morphological preprocessing for statistical machine translation. In *16th annual conference of the European Association for Machine Translation (EAMT)*, Trento, Italy.

Stolcke, A. (2002). SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO.

Tsvetkov, Y. and Wintner, S. (2010). Automatic acquisition of parallel corpora from websites with dynamic content. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 3389–3392.

Utiyama, M. and Isahara, H. (2007). A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York. Association for Computational Linguistics.

Wu, H. and Wang, H. (2009). Revisiting pivot language approach for machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 154–162, Suntec, Singapore. Association for Computational Linguistics.

Yeniterzi, R. and Oflazer, K. (2010). Syntax-to-morphology mapping in factored phrase-based statistical machine translation from english to turkish. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 454–464, Uppsala, Sweden. Association for Computational Linguistics.