Incorporating Linguistic Knowledge in Statistical Machine Translation: Translating Prepositions

Reshef Shilon
Dept. of Linguistics
Tel Aviv University
Israel

Hanna Fadida
Dept. of Computer Science
Technion
Israel

Shuly Wintner
Dept. of Computer Science
University of Haifa
Israel

Abstract

Prepositions are hard to translate, because their meaning is often vague, and the choice of the correct preposition is often arbitrary. At the same time, making the correct choice is often critical to the coherence of the output text. In the context of statistical machine translation, this difficulty is enhanced due to the possible long distance between the preposition and the head it modifies, as opposed to the local nature of standard language models. In this work we use monolingual language resources to determine the set of prepositions that are most likely to occur with each verb. We use this information in a transfer-based Arabic-to-Hebrew statistical machine translation system. We show that incorporating linguistic knowledge on the distribution of prepositions significantly improves the translation quality.

1 Introduction

Prepositions are hard to translate. Prepositional phrases modify both nouns and verbs (and, in some languages, other parts of speech); we only focus on verbs in this work. When a prepositional phrase modifies a verb, it can function as a complement or as an adjunct of the verb. In the former case, the verb typically determines the preposition, and the choice is rather arbitrary (or idiomatic). In fact, the choice of preposition can vary among synonymous verbs even in the same language. Thus, English *think* takes either *of* or *about*, whereas *ponder* takes no preposition at all (we view direct objects as prepositional phrases with a null preposition in this work.) Hebrew *hkh* "hit" takes the accusative preposition *at*, whereas the synonymous *hrbic* "hit" takes *l* "to". Arabic *tfAdY* "watch out" takes a direct object or *mn* "from", whereas *A$fq* "be careful of" takes *En* "on" and *tHrz* "watch out" takes *mn* "from".[1]

In the latter case, where the prepositional phrase is an adjunct, the choice of preposition does convey some meaning, but this meaning is vague, and the choice is often determined by the noun phrase that follows the preposition (the *object* of the preposition). Thus, temporals such as *last week*, *on Tuesday*, or *in November*, locatives such as *on the beach*, *at the concert*, or *in the classroom*, and instrumentals such as *with a spoon*, are all translated to prepositional phrases with *the same* preposition, *b* "in", in Hebrew (*b+šbw‘ š‘br*, *b+ywm šliši*, *b+nwbmbr*, *b+ym*, *b+qwncrT*, *b+kth*, and *b+kp*, respectively).

Clearly, then, prepositions cannot be translated literally, and the head that they modify, as well as the object of the preposition, have to be taken into account when a preposition is chosen to be generated. Standard phrase-based statistical machine translation (MT) does not always succeed in addressing this challenge, since the coherence of the output text is determined to a large extent by an n-gram language model. While such language models can succeed to discriminate in favor of the correct preposition in local contexts, in long-distance dependencies they are likely to fail.

We propose a method for incorporating linguistic knowledge pertaining to the distribution of prepositions that are likely to occur with verbs in a transfer-based statistical machine translation system. First, we use monolingual language resources to rank the possible prepositions that various verbs subcategorize for. Then, we use this information in an Arabic-to-Hebrew MT system.

[1] To facilitate readability we use a transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexicographic order, are abgdhwzxTiklmnspcqršt. For Arabic we use the transliteration scheme of Buckwalter (2004).

Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data (Hybrid2012), EACL 2012, pages 106–114,
Avignon, France, April 23 2012.

The system is developed in the framework of Stat-XFER (Lavie, 2008), which facilitates the explicit expression of synchronous (extended) context-free transfer rules. We use this facility to implement rules that verify the correct selection of prepositions by the verbs that subcategorize them. We show that this results in significant improvement in the translation quality.

In the next section we briefly survey related work. Section 3 introduces the Stat-XFER framework in which our method is implemented. We present the problem of translating prepositions between Hebrew and Arabic in Section 4, and discuss possible solutions in Section 5. Our proposed method consists of two parts: acquisition of verb-preposition mappings from corpora (Section 6), and incorporation of this knowledge in an actual transfer-based MT system (Section 7). Section 8 provides an evaluation of the results. We conclude with suggestions for future research.

2 Related Work

An explicit solution to the challenges of translating prepositions was suggested by Trujillo (1995), who deals with the problem of translating spatial prepositions between Spanish and English in the context of a lexicalist transfer-based MT framework. Trujillo (1995) categorizes spatial prepositions according to a lexical-semantic hierarchy, and after parsing the source language sentence, uses the representation of prepositions in the transfer process, showing improvement in performance compared to other transfer-based systems. This requires resources much beyond those that are available for Arabic and Hebrew.

More recent works include Gustavii (2005), who uses transformation-based learning to infer rules that can correct the choice of preposition made by a rule-based MT system. Her reported results show high accuracy on the task of correctly generating a preposition, but the overall improvement in the quality of the translation is not reported. Li et al. (2005) focus on three English prepositions (*on*, *in* and *at*) and use WordNet to infer semantic properties of the immediate context of the preposition in order to correctly translate it to Chinese. Again, this requires language resources that are unavailable to us. WordNet (and a parser) are used also by Naskar and Bandyopadhyay (2006), who work on English-to-Bengali translation.

The closest work to ours is Agirre et al. (2009), who translate from Spanish to Basque in a rule-based framework. Like us, they focus on prepositional phrases that modify verbs, and include also the direct object (and the subject) in their approach. They propose three techniques for correctly translating prepositions, based on information that is automatically extracted from monolingual resources (including verb-preposition-head dependency triplets and verb subcategorization) as well as manually-crafted selection rules that rely on lexical, syntactic and semantic information. Our method is similar in principle, the main differences being: (i) we incorporate linguistic knowledge in a *statistical* decoder, facilitating scalability of the MT system, (ii) we use much more modest resources (in particular, we do not parse either of the two languages), and (iii) we report standard evaluation measures.

Much work has been done regarding the automatic acquisition of subcategorization frames in English (Brent, 1991; Manning, 1993; Briscoe and Carroll, 1997; Korhonen, 2002), Czech (Sarkar and Zeman, 2000), French (Chesley and Salmon-alt, 2006), and several other languages. The technique that we use here (Section 6) can now be considered standard.

3 Introduction to Stat-XFER

The method we propose is implemented in the framework of Stat-XFER (Lavie, 2008), a statistical machine translation engine that includes a declarative formalism for symbolic transfer grammars. A grammar consists of a collection of synchronous context-free rules, which can be augmented by unification-style feature constraints. These transfer rules specify how phrase structures in a source-language correspond and transfer to phrase structures in a target language, and the constraints under which these rules should apply. The framework also includes a transfer engine that applies the transfer grammar to a source-language input sentence at runtime, and produces collections of scored word- and phrase-level translations according to the grammar. Scores are based on a log-linear combination of several features, and a beam-search controls the underlying parsing and transfer process.

Crucially, Stat-XFER is a statistical MT framework, which uses statistical information to weigh word translations, phrase correspon-

dences and target-language hypotheses; in contrast to other paradigms, however, it can utilize both automatically-created and manually-crafted language resources, including dictionaries, morphological processors and transfer rules. Stat-XFER has been used as a platform for developing MT systems for Hindi-to-English (Lavie et al., 2003), Hebrew-to-English (Lavie et al., 2004), Chinese-to-English, French-to-English (Hanneman et al., 2009) and many other low-resource language pairs, such as Inupiaq-to-English and Mapudungun-to-Spanish.

In this work, we use the Arabic-to-Hebrew MT system developed by Shilon et al. (2010), which uses over 40 manually-crafted rules. Other resources include Arabic morphological analyzer and disambiguator (Habash, 2004), Hebrew morphological generator (Itai and Wintner, 2008) and a Hebrew language model compiled from available corpora (Itai and Wintner, 2008).

While our proposal is cast within the framework of Stat-XFER, it can be in principle adapted to other syntax-based approaches to MT; specifically, Williams and Koehn (2011) show how to employ unification-based constraints to the target-side of a string-to-tree model, integrating constrain evaluation into the decoding process.

4 Translating prepositions between Hebrew and Arabic

Modern Hebrew and Modern Standard Arabic, both closely-related Semitic languages, share many orthographic, lexical, morphological, syntactic and semantic similarities, but they are still not mutually comprehensible. Machine translation between these two languages can indeed benefit from the similarities, but it remains a challenging task. Our current work is situated in the framework of the only direct MT system between these two languages that we are aware of, namely Shilon et al. (2010).

Hebrew and Arabic share several similar prepositions, including the frequent *b* "in, at, with" and *l* "to". However, many prepositions exist in only one of the languages, such as Arabic *En* "on, about" or Hebrew *šl* "of". Hebrew uses a preposition, *at*, to introduce definite direct objects (which motivates our choice of viewing direct objects as special kind of prepositional phrases, which may sometimes be introduced by a null preposition). The differences in how the two languages use prepositions are significant and common, as the following examples demonstrate.

(1) *AErb* *Al+wzyr* *En* *Aml+h*
expressed.3ms the+minister on hope+his
'The minister expressed his hope' (Arabic)

h+šr *hbi'* *at* *tqwt+w*
the+minister expressed.3ms acc hope+his
'The minister expressed his hope' (Hebrew)

(2) *HDr* *Al+wzyr* *Al+jlsp*
attended.3ms the+minister the+meeting
'The minister attended the meeting' (Arabic)

h+šr *nkx* *b+* *h+išibh*
the+minister attended.3ms in the+meeting
'The minister attended the meeting' (Hebrew)

In (1), the Arabic preposition *En* "on, about" is translated into the Hebrew accusative marker *at*. In contrast, (2) demonstrates the opposite case where the Arabic direct object (no preposition) is translated into a Hebrew prepositional phrase introduced by *b* "in". Clearly, despite the lexical and semantic similarity between many Hebrew and Arabic prepositions, their licensing by semantically-equivalent verbs is different in both languages.

An important issue is the selection of prepositions to model. We focus on a small list of the most common prepositions in both languages. The list was constructed by counting prepositions in monolingual corpora from the news domain in the two languages (500K tokens in Arabic, 120K tokens in Hebrew). In total, the Arabic data includes 70K prepositions, which comprise 14% of the corpus tokens, whereas the Hebrew data includes 19K prepositions, or 16% of the tokens. Not surprisingly, the most frequent prepositions were those that are commonly used to introduce complements. The data are listed in Table 1.

Based on these data, we decided to focus on the set of top nine Arabic prepositions (*fy*, *l*, *b*, *mn*, *ElY*, *AlY*, *En*, *mE* and the direct object), and the top six Hebrew prepositions (*b*, *l*, *m*, *'l*, *'m*, and the direct object), comprising over 80% of all preposition occurrences in our corpora.[2] These are also the most common complement-preceding prepositions, and therefore pose the main challenge for the task of machine translation.

[2]The preposition *k* "as" is omitted since it is translated directly to itself in most cases.

	Arabic				Hebrew			
Rank	Preposition	Count	%	$\sum$%	Preposition	Count	%	$\sum$%
1	*fy* "in"	13128	18.7	18.7	*b* "in"	6030	31.6	31.6
2	dir-obj	12626	17.9	36.7	*l* "to"	3386	17.7	49.3
3	*l* "to"	9429	13.4	50.1	dir-obj	3250	17.0	66.3
4	*b* "in, with"	7253	10.3	60.4	*m* "from"	1330	6.9	73.3
5	*mn* "from"	6859	9.7	70.2	*'l* "on"	1066	5.5	78.9
6	*ElY* "on"	5304	7.5	77.8	*k* "as"	354	1.8	80.7
7	*AlY* "to"	4458	6.3	84.1	*'m* "with"	338	1.7	82.5
8	*En* "on, about"	1871	2.6	86.8	*bin* "between"	191	1.0	84.6
9	*mE* "with"	1380	1.9	88.8	*'d* "until"	159	0.8	85.4
10	*byn* "between"	1045	1.4	90.3	*lpni* "before"	115	0.6	86.0

Table 1: Counts of Arabic and Hebrew most frequent prepositions. The columns list, for each preposition, its count in the corpus, the percentage out of all prepositions, and the accumulated percentage including all the higher-ranking prepositions.

5 Possible solutions

In order to improve the accuracy of translating prepositions in a transfer-based system, several approaches can be taken. We discuss some of them in this section.

First, accurate and comprehensive statistics can be acquired from large monolingual corpora of the target language regarding the distribution of verbs with their subcategorized prepositions and the head of the noun phrase that is the object of the preposition. As a backoff model, one could use a bigram model of only the preposition and the head of the following noun phrase, e.g., (*on, Wednesday*). This may help in the case of temporal and locative adjuncts that are less related to the preceding verb. Once such data are acquired, they may be used in the process of scoring hypotheses, if a parser is incorporated in the process.

One major shortcoming of this approach is the difficulty of acquiring the necessary data, and in particular the effect of data sparsity on the accuracy of this approach. In addition, a high quality parser for the target language must be available, and it must be incorporated during the decoding step, which is a heavy burden on performance.

Alternatively, one could acquire lexical and semantic mappings between verbs, the type of their arguments, the selectional restrictions they impose, and the possible prepositions used to express such relations. This can be done using a mapping from surface forms to lexical ontologies, like WordNet (Fellbaum, 1998), and to a syntactic-semantic mapping like VerbNet (Schuler, 2005) which lists the relevant preceding preposition. Similar work has been done by Shi and Mihalcea (2005) for the purpose of semantic parsing. These lexical-semantic resources can help map between the verb and its possible arguments with their thematic roles, including selectional restrictions on them (expressed lexically, using a WordNet synset, like *human* or *concrete*).

The main shortcoming of this solution is that such explicit lexical and semantic resources exist mainly for English. In addition, even when translating into English, this information can only assist in limiting the number of possible prepositions but not in determining them. For example, one can talk *about* the event, *after* the event, or *at* the event. The information that can determine the correct preposition is in the source sentence.

Finally, a potential solution is to allow translation of source-language prepositions to a limited set of possible target-language prepositions, and then use both target-language constraints on possible verb-preposition matches and an n-gram language model to choose the most adequate solution. Despite the fact that this solution does not model the probability of the target preposition given its verb and the original sentence, it limits the number of possible translations by taking into account the target-language verb and the possible constraints on the prepositions it licenses. This method is also the most adequate for a scenario that employs a statistical decoder, such as the one used in Stat-XFER. This is the solution we advocate in this paper. We describe the acquisition of Hebrew verb–preposition statistics in the following section, and the incorporation of this knowledge in a machine translation system in Section 7.

6 Acquisition of verb–preposition data

To obtain statistics on the relations between verbs and prepositions in Hebrew we use the *TheMarker, Knesset* and *Arutz 7* corpora (Itai and Wintner, 2008), comprising 31M tokens. The corpora include 1.18M (potentially inflected) verb tokens, reflecting 4091 verb (lemma) types.

The entire corpus was morphologically analyzed and disambiguated (Itai and Wintner, 2008). We then collected all instances of prepositions that immediately follow a verb; this reflects the assumption that such prepositions are likely to be a part of the verb's subcategorization frame. A special treatment of the direct object case was required, because a Hebrew direct object is introduced by the accusative marker *at* when it is definite, but not otherwise. Since constituent order in Hebrew is relatively free, the noun phrase that immediately follows the verb can also be its subject. Therefore, we only consider such noun phrases if they do not agree with the verb in gender and number (and are therefore not subjects).

We then use maximum likelihood estimation to obtain the conditional probability of each preposition following a verb. The result is a database of verb-preposition pairs, with an estimate of their probabilities. Examples include *nkll* "be included", for which *b* "in" has 0.91 probability; *hstpq* "be satisfied" *b* "in" (0.99); *xikh* "wait" *l* "to" (0.73); *ht'lm* "ignore" *m* "from" (0.83); and *htbss* "base" *'l* "on" (0.93). Of course, some other verbs are less clear-cut.

From this database, we filter out verb-preposition pairs whose score is lower than a certain threshold. We are left with a total of 1402 verbs and 2325 verb-preposition pairs which we use for Arabic-to-Hebrew machine translation, as explained in the next section. Note that we currently ignore the probabilities of the prepositions associated with each verb; we only use the probabilities to limit the set of prepositions that are licensed by the verb. Ranking of these prepositions is deferred to the language model.

7 Incorporating linguistic knowledge

We implemented the last method suggested in Section 5 to improve the quality of the Arabic-to-Hebrew machine translation system of Shilon et al. (2010) as follows.

First, we modified the output of the Hebrew morphological generator to reflect also, for each verb, the list of prepositions licensed by the verb (Section 6). Stat-XFER uses the generator to generate inflected forms of lemmas obtained from a bilingual dictionary. Each such form is associated with a feature structure that describes some properties of the form (e.g., its gender, number and person). To the feature structures of verbs we add an additional feature, ALLOWED_PREPS, whose value is the list of prepositions licensed by the verb. For example, the feature structure of the Hebrew verb *sipr* "tell" is specified as:

```
(allowed_preps = (*OR* at l))
```

Thus, whenever the Hebrew generator returns an inflected form of the verb *sipr*, the feature ALLOWED_PREPS lists the possible prepositions *at* and *l* "to", that are licensed by this verb.

Then, we modified the transfer grammar to enforce constraints between the verb and its objects. This was done by adding a new non-terminal node to the grammar, OBJ, accounting for both direct and indirect objects. The idea is to encode the actual preposition (in fact, its surface form) as a feature of the OBJ node (Figure 1), and then, when a sentence is formed by combining a verb with its subject and object(s), to check the value of this

```
{OBJ_ACC_AT,0}
OBJ::OBJ [NP] -> ["AT" NP]
(X1::Y2)
((X1 def) = +)
((Y2 prep) = AT) #mark preposition
(X0 = X1)
(Y0 = Y2

{OBJ_PP,0}
OBJ::OBJ [PREP NP] -> [PREP NP]
(X1::Y1)
(X2::Y2)
((Y0 prep) = (Y1 lex)) #mark prep.
(X0 = X1)
(Y0 = Y1)

{OBJ_NP_PP_B, 0}
OBJ::OBJ [NP] -> ["B" NP]
(X1::Y2)
((Y0 prep) = B) #mark preposition
(X0 = X1)
(Y0 = Y2)
```

Figure 1: Propagating the surface form of the preposition as a feature of the OBJ node.

```
{S_VB_NP_OBJ_swap, 1}
S::S [VB NP OBJ] -> [NP VB OBJ]
(X1::Y2)
(X2::Y1)
(X3::Y3)
((X1 num) = singular) # Arabic agr.
((X1 per) = (X2 per))
((Y1 num) = (Y2 num)) # Hebrew agr.
((Y1 gen) = (Y2 gen))
((Y1 per) = (Y2 per))
((Y2 allowed_preps) = (Y3 prep))
```

Figure 2: Enforcing agreement between a verb VB and its object OBJ on the Hebrew side.

feature against the ALLOWED_PREPS feature of the verb (Figure 2).

Consider Figure 1. The first rule maps an Arabic direct object noun phrase to a Hebrew direct object, and marks the preposition *at* on the Hebrew OBJ node as the value of the feature PREP. The second rule maps an Arabic prepositional phrase to Hebrew prepositional phrase, marking the Hebrew OBJ (referred to here as Y1 lex) with the value of the feature PREP. The third rule maps an Arabic noun phrase to a Hebrew prepositional phrase introduced by the preposition *b* "in".

The rule in Figure 2 enforces sentence-level agreement between the feature ALLOWED_PREPS of the Hebrew verb (here, Y2 allowed_preps) and the actual preposition of the Hebrew object (here, Y3 prep).

To better illustrate the effect of these rules, consider the following examples, taken from the system's actual output (the top line is the Arabic input, the bottom is the Hebrew output). There can be four types of syntactic mappings between Arabic and Hebrew arguments: (NP, NP), (NP, PP), (PP, NP) and (PP, PP). Examples (3) and (4) demonstrate correct translation of the Arabic direct object into the Hebrew direct object (with and without the Hebrew definite accusative marker *at*, respectively). Example (5) demonstrates the correct translation of the Arabic direct object to a Hebrew PP with the preposition *l* "to". Example (6) demonstrates the correct translation of an Arabic PP introduced by *En* "on, about" to a Hebrew direct object, and Example (7) demonstrates the translation of Arabic PP introduced by *b* "in, with" into a Hebrew PP introduced by *'m* "with".

(3) *rAyt* *Al+wld*
see.past.1s the+boy

raiti *at* *h+ild*
see.past.1s acc.def the+boy
'I saw the boy'

(4) *rAyt* *wldA*
see.past.1s boy.acc.indef

raiti *ild*
see.past.1s boy
'I saw a boy'

(5) *Drb* *Al+Ab* *Al+wld*
hit.past.3ms the+father the+boy

h+ab *hrbic* *l+* *h+ild*
the+father hit.past.3ms to the+boy
'The father hit the boy'

(6) *AErb* *Al+wzyr* *En* *Aml+h*
express.past.3ms the+minister on hope+his

h+šr *hbi'* *at*
the+minister express.past.3ms acc.def.
tqwt+w
hope+his
'The minister expressed his hope'

(7) *AjtmE* *Al+wzyr* *b+* *Al+wld*
meet.past.3ms the+minister in the+boy

h+šr *npgš* *'m* *h+ild*
the+minister meet.past.3ms with the+boy
'The minister met the boy'

In (3), the input Arabic NP is definite and is marked by accusative case. A designated rule adds the string *at* before the corresponding Hebrew output, to mark the definite direct object. We create a node of type OBJ for both (direct) objects, with the feature PREP storing the lexical content of the preposition in the target language. Finally, in the sentence level rule, we validate that the Hebrew verb licenses a direct object, by unifying the feature PREP of OBJ with the feature ALLOWED_PREPS of the verb VB.

In (4), a similar process occurs, but this time no additional *at* token is added to the Hebrew output (since the direct object is indefinite). The same preposition, *at*, is marked as the PREP feature of OBJ (we use *at* to mark the direct object, whether the object is definite or not), and again, the feature PREP of OBJ is validated against the feature ALLOWED_PREPS of VB.

Example (5) is created using a rule that maps an Arabic direct object to a Hebrew prepositional phrase introduced by a different preposition, here *l* "to". Such rules exist for every Hebrew preposition from the set of common prepositions we focus on, since we have no prior knowledge of which preposition should be generated. We mark the lexical preposition *l* on the feature PREP of the Hebrew OBJ node, and again, this is validated in the sentence level against the prepositions allowed by the verb.

In example (6) we use rules that map an Arabic prepositional phrase to a Hebrew noun phrase. Here, the Arabic preposition is not translated at all, and the Hebrew definite accusative marker *at* is added, depending on the definiteness of the Hebrew noun phrase. The only difference in example (7) compared to previous examples is the translation of the Arabic preposition into a different Hebrew preposition. This is implemented in the bilingual lexicon, in a lexical entry that maps the Arabic preposition *b* "in, with" to the Hebrew preposition *'m* "with".

These rules help to expand the lexical variety of the prepositions on one hand (as in Example (7)), while at the same time disqualifying some hypotheses that employ prepositions that are not licensed by the relevant verb, using unification-style constraints. After this process, the lattice may still include several different hypotheses, from which the decoder statistically chooses the best one.

8 Evaluation

To evaluate the contribution of the proposed method, we created a test set of 300 sentences from newspaper texts, which were manually translated by three human translators. Of those, we selected short sentences (up to 10 words), for which the bilingual lexicon used by the system had full lexical coverage. This resulted in a set of 28 sentences (still with three reference translations each), which allowed us to focus on the actual contribution of the preposition-mapping solution rather than on other limitations of the MT system. Unfortunately, evaluation on the entire test set without accounting for full lexical coverage yields such low BLEU scores that the comparison between different configurations of the system is meaningless.

As a baseline system, we use exactly the same setup, but withhold any monolingual linguistic knowledge regarding verb-prepositions relations:

1. We omit the restrictions (stated in the grammar) on which prepositions Hebrew verbs license, such that each verb can be followed by each preposition.

2. We limit the lexical variance between prepositions in the lexicon, to only allow translation-pairs that occur in the bilingual dictionary. For example, we use the mapping of Arabic *ElY* "on" to Hebrew *'l* "on" (which occurs in the bilingual dictionary), but remove the mapping of Arabic *ElY* "on" to Hebrew *b* "in", which does not carry the same meaning.

Table 2 lists the BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2011) scores of both systems.

	BLEU	METEOR
Baseline	0.325	0.526
With prepositions	0.370	0.560

Table 2: Automatic evaluation scores.

The system that incorporates linguistic knowledge on prepositions significantly ($p < 0.05$) outperforms the baseline system. A detailed analysis of the obtained translations reveals that the baseline system generates prepositions that are not licensed by their head verb, and the language model fails to choose the hypothesis with the correct preposition, if such a hypothesis is generated at all.

As an example of the difference between the outputs of both systems, consider Figure 3. The Arabic input is given in (8). The output of the system that incorporates our treatment of prepositions is given in (9). Here, the Hebrew verb *hdgiš* "emphasize" is followed by the correct definite accusative marker *at*. The output of the baseline system is given in (10). Here, the Hebrew verb *aišr* "approve" is followed by the wrong preposition, *'l* "on", which is not licensed in this location. Consequently, the lexical selections for the following words of the translation differ and are not as fluent as in (9), and the output is only partially coherent.

(8) *Akd* *AlHryry* *ElY* *AltzAm+h* *b+* *Al+byAn* *Al+wzAry*
emphasize.past.3ms AlHaryry on obligation+his in the+announcement the+ministerial
l+ *Hkwmp* *Al+whdp* *Al+wTnyp*
to government the+unity the+national
'Alharyry emphasized his obligation in the ministerial announcement to the national government'

(9) *alxriri* *hdgiš* *at* *xwbt+w* *b+* *h+hwd'h*
Alharyry emphasize.past.3ms def.acc obligation+his in the+announcement
h+mmšltit *l+* *mmšlt* *h+axdwt* *h+lawmit*
the+governmental to government the+unity the+national
'Alharyry emphasized his obligation in the governmental announcement to the national government'

(10) *alxriri* *aišr* *'l* *zkiwn* *šl+w* *b+* *h+hwd'h* *h+mmšltit*
Alharyry confirm.past.3ms on permit of+his in the+announcement the+governmental
l+ *mmšlt* *h+axdwt* *h+lawmit*
to government the+unity the+national
'Alharyry confirmed on his permit in the governmental announcement to the national government'

Figure 3: Example translation output, with and without handling of prepositions.

9 Conclusion

Having emphasized the challenge of (machine) translation of prepositions, specifically between Hebrew and Arabic, we discussed several solutions and proposed a preferred method. We extract linguistic information regarding the correspondences between Hebrew verbs and their licensed prepositions, and use this knowledge for improving the quality of Arabic-to-Hebrew machine translation in the context of the Stat-XFER framework. We presented encouraging evaluation results showing that the use of linguistic knowledge regarding prepositions indeed significantly improves the quality of the translation.

This work can be extended along various dimensions. First, we only focused on verb arguments that are prepositional phrases here. However, our Hebrew verb-subcategorization data include also information on other types of complements, such as subordinate clauses (introduced by the complementizer *š* "that") and infinitival verb phrases. We intend to extend our transfer grammar in a way that will benefit from this information in the future. Second, we currently do not use the weights associated with specific prepositions in our subcategorization database; we are looking into ways to incorporate this statistical information in the decoding phase of the translation.

Furthermore, our database contains also statistics on the distribution of nouns following each preposition (which are likely to function as the heads of the object of the preposition); such information can also improve the accuracy of translation, and can be incorporated into the system. Another direction is to acquire and incorporate similar information on deverbal nouns, which license the same prepositions as the verbs they are derived from. For example, *xtimh 'l hskm* "signing.noun an agreement", where the Hebrew preposition *'l* "on" must be used, as in the corresponding verbal from *xtm 'l hskm* "signed.verb an agreement". We will address such extensions in future research.

Acknowledgements

We are grateful to Alon Itai, Alon Lavie, and Gennadi Lembersky for their help. This research was supported by THE ISRAEL SCIENCE FOUNDATION (grant No. 137/06).

References

Eneko Agirre, Aitziber Atutxa, Gorka Labaka, Mikel Lersundi, Aingeru Mayor, and Kepa Sarasola. 2009. Use of rich linguistic information to translate prepositions and grammar cases to Basque. In *Proceedings of the XIII Conference of the European*

Association for Machine Translation, EAMT-2009, pages 58–65, May.

Michael R. Brent. 1991. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 209–214.

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363.

Tim Buckwalter. 2004. *Buckwalter Arabic Morphological Analyzer Version 2.0*. Linguistic Data Consortium, Philadelphia.

Paula Chesley and Susanne Salmon-alt. 2006. Automatic extraction of subcategorization frames for French. In *Proceedings of the Language Resources and Evaluation Conference, LREC-2006*.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.

Ebba Gustavii. 2005. Target language preposition selection – an experiment with transformation-based learning and aligned bilingual data. In *Proceedings of EAMT-2005*, May.

Nizar Habash. 2004. Large scale lexeme based arabic morphological generation. In *Proceedings of Traitement Automatique du Langage Naturel (TALN-04)*, Fez, Morocco.

Greg Hanneman, Vamshi Ambati, Jonathan H. Clark, Alok Parlikar, and Alon Lavie. 2009. An improved statistical transfer system for French–English machine translation. In *StatMT '09: Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 140–144.

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

Anna Korhonen. 2002. *Subcategorisation acquisition*. Ph.D. thesis, Computer Laboratory, University of Cambridge. Techical Report UCAM-CL-TR-530.

Alon Lavie, Stephan Vogel, Lori Levin, Erik Peterson, Katharina Probst, Ariadna Font Llitjós, Rachel Reynolds, Jaime Carbonell, and Richard Cohen. 2003. Experiments with a Hindi-to-English transfer-based MT system under a miserly data scenario. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(2):143–163.

Alon Lavie, Shuly Wintner, Yaniv Eytani, Erik Peterson, and Katharina Probst. 2004. Rapid prototyping of a transfer-based Hebrew-to-English machine translation system. In *Proceedings of TMI-2004: The 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October.

Alon Lavie. 2008. Stat-XFER: A general search-based syntax-driven framework for machine translation. In Alexander F. Gelbukh, editor, *CICLing*, volume 4919 of *Lecture Notes in Computer Science*, pages 362–375. Springer.

Hui Li, Nathalie Japkowicz, and Caroline Barrière. 2005. English to Chinese translation of prepositions. In Balázs Kégl and Guy Lapalme, editors, *Advances in Artificial Intelligence, 18th Conference of the Canadian Society for Computational Studies of Intelligence*, volume 3501 of *Lecture Notes in Computer Science*, pages 412–416. Springer, May.

Christopher D. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 235–242.

Sudip Kumar Naskar and Sivaji Bandyopadhyay. 2006. Handling of prepositions in English to Bengali machine translation. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, pages 89–94.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Anoop Sarkar and Daniel Zeman. 2000. Automatic extraction of subcategorization frames for Czech. In *Proceedings of the 18th conference on Computational linguistics*, pages 691–697.

Karin Kipper Schuler. 2005. *Verbnet: a broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In Alexander F. Gelbukh, editor, *CICLing*, volume 3406 of *Lecture Notes in Computer Science*, pages 100–111. Springer.

Reshef Shilon, Nizar Habash, Alon Lavie, and Shuly Wintner. 2010. Machine translation between Hebrew and Arabic: Needs, challenges and preliminary solutions. In *Proceedings of AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas*, November.

Indalecio Arturo Trujillo. 1995. *Lexicalist machine translation of spatial prepositions*. Ph.D. thesis, University of Cambridge, April.

Philip Williams and Philipp Koehn. 2011. Agreement constraints for statistical machine translation into German. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 217–226, Edinburgh, Scotland, July.