Machine Translation between Hebrew and Arabic: Needs, Challenges and Preliminary Solutions

Reshef Shilon
Dept. of Linguistics
Tel Aviv U.
Tel Aviv, Israel

Nizar Habash
CCLS
Columbia U.
New York, NY

Alon Lavie
LTI
Carnegie Mellon U.
Pittsburgh, PA

Shuly Wintner
Dept. of Computer Science
U. of Haifa
Haifa, Israel

Abstract

Hebrew and Arabic are related but mutually incomprehensible languages with complex morphology and scarce parallel corpora. Machine translation between the two languages is therefore interesting and challenging. We discuss similarities and differences between Hebrew and Arabic, the benefits and challenges that they induce, respectively, and their implications for machine translation. We highlight the shortcomings of using English as a pivot language and advocate a direct, transfer-based and linguistically-informed (but still statistical, and hence scalable) approach. We report preliminary results of such a system that we are currently developing.

1 Introduction

Modern Hebrew and Modern Standard Arabic, both Semitic languages, share many orthographic, lexical, morphological, syntactic and semantic similarities, but they are still not mutually comprehensible.[1] Most native Hebrew speakers in Israel do not speak Arabic, and the vast majority of Arabs (outside Israel) do not speak Hebrew. Machine translation (MT) between these two language has the potential to bridge over political and cultural differences and bring the disputing peoples in the Middle East somewhat closer together by better understanding each other's societies.

The dominant paradigm in contemporary MT (Brown et al., 1990) relies on large-scale parallel corpora from which correspondences between the two languages can be extracted. However, such abundant parallel corpora currently exist only for few language pairs; and low- and medium-density languages (Varga et al., 2005) require alternative approaches. Specifically, no parallel corpora exist for Hebrew–Arabic.[2]

As an alternative to the pure statistical approach, we are currently developing a Hebrew-to-Arabic MT system, using the Stat-XFER framework (Lavie, 2008), which is particularly suited for low-resource language pairs. We discuss in Section 2 some linguistic properties of the two languages. Section 3 describes the implications on MT of the similarities and, in particular, differences between the two languages. In Section 4 we discuss possible solutions to these challenges, advocating in Section 5 a linguistically-aware, transfer-based approach. Section 6 describes the system we are in the process of developing and reports some preliminary results.

2 Linguistic properties

Hebrew and Arabic are both closely-related (West) Semitic languages, implying that they share many linguistic properties and structures, even though they are not mutually comprehensible. We briefly discuss some of the similarities and differences below.

[1] In certain respects, Arabic Dialects have morpho-syntactic features closer to Hebrew than Modern Standard Arabic, e.g., the absence of nominal case and verbal mood, the behavior of the feminine ending in genitive constructions, the gender-number invariance of the relativizer, and the dominance of SVO order over VSO order. We do not discuss Arabic dialects here.

[2] Several web sites have *comparable* contents, e.g., Wikipedia or the Israeli daily YNet (`http://www.ynet.co.il`); A small set of translated political essays is available from Gush Shalom (`http://www.gush-shalom.org/`) and Zavit Akheret (`http://zavita.co.il/`); the Bible is not available in Modern Hebrew.

2.1 Orthography

Letters and diacritics While Hebrew and Arabic use different writing systems, they share many orthographic similarities. Their orthographies consist of a system of letters, denoting consonants and long vowels, and diacritics, which denote short vowels. In both languages, the diacritics are typically omitted in contemporary texts, which leads to high morphological ambiguity, and makes text analysis a harder task.[3]

Translating *to* non-diacriticized Arabic (or Hebrew) has its advantages, since many variant words share the same non-diacriticized form. For example, distinction in gender in second person pronouns is lost in some scenarios in both languages: the Hebrew forms */katavta/* 'you (2.sg.m) wrote' and */katavt/* 'you (2.sg.f) wrote' collapse into the non-diacriticized form *ktbt*; and the Arabic forms */baytuka/* 'your (2.sg.m) house' and */baytuki/* 'your (2.sg.f) house' collapse into the non-diacriticized form *bytk*. Moreover, Arabic case and mood features, absent in Hebrew, often realize as diacritics only: e.g., the Arabic orthographic word *wld* 'boy' can stand for *waladu* (nom. def.), *waladũ /waladun/* (nom. indef.), and *waladĩ /waladin/* (gen. indef.), among others.

Clitics In both languages, some prepositions (e.g., *b* 'in, with', *l* 'to, for'), conjunctions (e.g., *w* 'and') and the definite article are attached as proclitics to the following word. Attachment of more than one particle can trigger orthographic modifications:

(1) (a) *bkth*

b+	h+	kth
in	the	classroom

'in the classroom' (Hebrew)

(b) *llqlm*

l+	Al+	qlm
for	the	pen

'for the pen' (Arabic)

Arabic attaches pronominal direct objects as postverbal clitics, a construction that, while grammatical, is rarely used in contemporary Hebrew. Hebrew uses the definite direct object marker *at* instead.

(2) (a) *raiti* *awtm*

raiti	at	+hm
see.1sg.past	def.acc	they.acc

'I saw them' (Hebrew)

(b) *rÂythm*

rÂyt	+hm
see.1sg.past	they.acc

'I saw them' (Arabic)

2.2 Word formation

As in other Semitic languages, most nouns and verbs are built from a lexical *root*, a morpheme consisting of consonants only which generally denotes a vague semantic meaning, and from templates that add vowels (and, possibly, also consonants) to the root, yielding a lexeme. Many roots are shared between Hebrew and Arabic. For example, the root *k.t.b* 'write' has the same basic meaning in both languages, but it is used in different templates and yields different lexemes. The past tense, 1st person plural form of the verb 'write' is *ktbnw* in Hebrew, *ktbnA* in Arabic; the noun 'letter (message)' is derived from the same root, and is *mktb* in Hebrew, *mktwb* in Arabic. However, Hebrew also has *mkwtb* 'addressee' from the same root, which does not exist in Arabic, whereas Arabic has *ktAb* 'book', which does not exist in Hebrew.

Knowing the meaning of the root may lead to better selection of a translation, even though there are often semantic differences, as well as many cases of completely different roots.

2.3 Inflectional morphology

Inflectional morphology in both languages is rich and productive.

Nominal morphology Nouns and adjectives inflect for number, gender and definiteness. However, Arabic nominals have three values for the number feature (singular, plural and *dual*), whereas the dual form only exists in Hebrew in a few frozen cases. Furthermore, Arabic has an irregular way for producing the plural form of nouns (the 'broken plural'), whereas in Hebrew plural forms are regularly related to their singular counterparts. One important

[3] To facilitate readability we use a transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexicographic order, are *abgdhwzxTiklmns'pcqršt*. For Arabic we use the transliteration scheme of Habash et al. (2007): (in alphabetical order) *AbtθjHxdðrzssšSDTĎςγfqklmnhwy* and the additional symbols: ' ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ , ŷ ئ , ħ ة, ý ى, a ـَ, u ـُ, i ـِ, ∼ ـّ, ã ـً, ũ ـٌ, ĩ ـٍ. Phonetic forms are given between slashes.

difference between the two languages is that Arabic encodes case on nouns, whereas Hebrew does not.

Another difference is the form of feminine nouns in the genitive construction (Section 2.4). In Hebrew this construction triggers a change of the feminine ending *-h* to *-t*. In Arabic the feminine ending is always *ħ*, combining the duality of *h* and *t*, which changes to *t* only before a possessive pronominal enclitic. For example, in Hebrew the feminine noun *xtwlh* 'cat' changes in this construction into *xtwlt rxwb* 'street cat'; but in Arabic, *qTħ* 'cat' changes in *qTtnA* 'our cat' but not in *qTħ šArς* 'street cat'.

Many similar pronouns are common to both languages, and pronouns inflect for the same features (number, gender, person and case). This makes translation of pronouns easier. Both nouns and prepositions can combine with cliticized pronominal suffixes that encode number, gender and person (of the possessor or the object of the preposition), e.g., *lnw* 'to us (Hebrew)', *lnA* 'to us' (Arabic).

Verbal morphology Verbs inflect for number, gender, person and tense, and the two languages share a complex and similar verb structure and inflection system. The two languages share the same four verbal forms: **a.** the perfective form is used for the past tense in Arabic and Hebrew; **b.** the imperfective is used for the future tense in Hebrew but is used for a variety of tenses in Arabic (past, present and future) in coordination with various moods and particles; **c.** the imperative; and **d.** the active participle used for present tense in Hebrew and to a lesser extent as a deverbal in Arabic. The ambiguity of the Arabic imperfective form is a challenge for translation since it could correspond to multiple Hebrew forms: the negated forms of the Hebrew *ktb/kwtb/iktwb* 'he wrote/writes/will-write' translate to Arabic *lm/lA/ln yktb* all using the same verb with different moods and particles combining tense and negation (in the case of *lm* and *ln*).

Passivization is implemented differently in the two languages. Hebrew predominantly employs a morphological mechanism whereby an active verbal pattern has a passive counterpart. This is highly productive for two patterns (*pi‘el–pu‘al* and *hif‘il–huf‘al*), less so for the third (*pa‘al–nif‘al*). Arabic utilizes a different mechanism of vowel change, which is productive for almost all verbal patterns.

In both Hebrew and Arabic, the second person singular masculine and third person singular feminine forms are homonymous across the verbal paradigm in the imperfective/future tense. For example, *tktwb* 'you.sg.m/she will write' (Hebrew), *tktb* 'you.sg.m write/she writes' (Arabic). This is a clear case of morphological ambiguity that does *not* have to be resolved in translation.

2.4 Syntax

Word order The dominant word order is SVO in Hebrew and VSO in Arabic (although other orders are possible), but there are some syntactic constraints on this default order. In Arabic, an embedded clause after the subordinating conjunction *An* must start with a noun (such as the subject or an expletive pronoun). In addition, the subject of the clause should be in accusative case. Hebrew has no parallel construction. On the other hand, when a sentence begins with an adverbial in Hebrew, the default order is VSO.

Agreement Both Arabic and Hebrew have a complex agreement system, involving features such as person, number, gender, and definiteness. In both languages agreement constraints hold between the following POS pairs:

N-Adj When an adjective modifies a noun, they should agree on number, gender and definiteness. Noun-phrase (NP) internal word order is identical.

(3) *h+ild/Al+wld* *h+gbwh/Al+Twyl*
the+boy.sg.m the+tall.sg.m
'The tall boy' (Hebrew/Arabic)

A peculiarity in Arabic is that the agreement features of plural, irrational (non-human) nouns are always singular feminine, regardless of the gender of the singular noun, and ignoring the semantic plurality of the noun. Every reference to that noun in the sentence must agree with these features:

(4) (a) *Al+qlm* *Al+jmyl*
pen-m.sg.def pretty.m.sg.def
'The pretty pen' (Arabic)

(b) *Al+ÂqlAm* *Al+jmylħ*
pen-m.pl.def pretty.f.sg.def
'The pretty pens' (Arabic)

Quant-N Subtle agreement constraints hold between quantifiers (e.g., numerals) and the nouns they modify. These constraints differ across the two languages.

Subj-V In both languages the verb and the subject NP agree on person, number and gender. However, in Arabic VSO sentences, the verb agrees on person and gender with the subject, but always appears in singular form:

(5) *ktb* *Al+ÂwlAd*
write-past.sg.m boy-pl.m.def
'The boys wrote' (Arabic)

Verbless predicates Both languages have a common construction of verbless sentences, where the predicate is either a PP, another NP or an adjective. In both latter cases, the subject and the predicate must agree on number and gender, but the subject must be definite and the predicate indefinite.

(6) *Al+wld* *Twyl*
boy.m.sg.def tall.m.sg.indef
'the boy is tall (Arabic)

Genitive constructions In both languages a noun–noun construction (called *smikhut* in Hebrew, *idafa* in Arabic) is used to express genitive relations. The head of the structure is the first noun, which determines the number and gender agreement features. The definiteness of this structure is marked on the second noun only.

(7) *sfr/ktAb* *h+ild/Al+wld*
book.indef the+boy.def
'The boy's book' (Hebrew/Arabic)

In Hebrew, but not in Arabic, such relations can also be expressed in a different construction, using the possessive preposition *šl* 'of'.

(8) *h+sfr* *šl* *h+ild*
the+book.def of the+boy.def
'The boy's book' (Hebrew)

Hebrew exhibits yet another construction of double genitives, which does not exist in Arabic. In this construction, the antecedent noun is followed both by a cliticized possessive pronoun and by the genitive marker *šl* with the possessor.

(9) *sfr+w* *šl* *h+ild*
book+his of the+boy.def
'The boy's book' (Hebrew)

Pro-drop In both languages, a subject pronoun can be omitted if the verb is in perfective/past, imperfective/future or imperative forms. The agreement features of the subject can be deduced from the morphological form of the verb. This may facilitate translation in some cases: target pronouns do not have to be explicitly generated when they are missing in the source language.

Relative clauses In Arabic the relativizer carries gender and number features, and has to agree with the antecedent noun modified by the relative clause.

(10) *Al+ÂqlAm* *Al∼ty*
pen-m.pl.def REL.f.sg
Ăštrý+hA *Al+wld*
buy-past.3.m.sg+she-acc. boy-m.sg.def
'The pens which the boy bought' (Arabic)

Such relative clauses modify only definite nouns, as in (10). Relative clauses that modify indefinite nouns have no relativizer. The Hebrew relative clause starts with a relativizer which carries no agreement features.

(11)(a) *raiti* *ild* *š+*
see.1st.sg.past boy.sg.m.indef REL
qra *sfr*
read.3rd.sg.past book.sg.indef
'I saw a boy who read a book' (Hebrew)

(b) *rÂyt* *wldA*
see.1st.sg.past boy.sg.m.indef
qrÂ *ktAbA*
read.3rd.sg.past book.sg.indef
'I saw a boy [who] read a book' (Arabic)

Hebrew also has a construction in which the relativizer is the definite article *h+*. This construction can be used for relative clauses only if the embedded verb is in the present. A similar use appears in Arabic using the definite article with the active participle deverbal form.

3 Challenges

The similar characteristics of Arabic and Hebrew can indeed be beneficial for MT, but the differences listed above pose some intricate challenges. We list some of those below and suggest possible solutions to the issues presented in the following section.

3.1 Lexical challenges

As in other language pairs, Hebrew and Arabic verbs have different subcategorization frames for corresponding verbs. Some Hebrew verbs require a specific preposition before the indirect object while in Arabic the object is direct, and vice versa.

(12)(a)

nkx	*b+*	*h+pgišh*
attend.3sg.m.past	in+	meeting.def

‘he attended the meeting’ (Hebrew)

(b)

HDr	*Al+jlsħ*
attend.3sg.m.past	meeting.def

‘he attended the meeting’ (Arabic)

This phenomenon is of course not special to Hebrew-Arabic. However, combined with differences in word order between the two languages, its effect is enhanced. While the language model (LM) may correctly choose the preposition in the Arabic output sentence based on the local context, this is less likely in sentences with long-distance V–P dependencies, since the subject may intervene between the verb and its preposition.

(13)

Âςrb	*rŷys*	*Al+Hkwmħ*
express.3sg.m.past	leader	government.def

ywm	*Al+ÂrbςA’*	*fy*	*jlsħ*	*Al+Hkwmħ*
day	Wednesday	in	meeting	government.def

Al+Âsbwςyħ	*ςn*	*Âml*	*+h*	...
weekly.def	upon	hope	he.poss	

‘The prime minister expressed on Wednesday his hope ...’

This example demonstrates the possible distance between the verb *Âςrb* ‘express’ and its required preposition *ςn*, which are separated by the subject NP and other temporal and locative adjuncts. This distance hampers the ability of a LM to correctly select the preposition.

Another lexical challenge stems from the fact that existing Arabic lexical resources do not encode information on gender and rationality of nouns, which is crucial for enforcing N-Adj agreement. The implication is that in order to generate Arabic, one must overgenerate both masculine and feminine forms, delegating the choice to the language model, which chooses poorly in long-distance dependencies.

3.2 Morphological challenges

Translating between two morphologically rich languages poses challenges in analysis, transfer and generation. The complex morphology induces an inherent data sparsity problem, and the limitation imposed by the dearth of available parallel corpora is magnified (Habash and Sadat, 2006).

We use a morphological analyzer (Itai and Wintner, 2008) for the Hebrew source, with no morphological disambiguation module.[4] This causes many wrong analyses to be processed and dramatically increases the size of the hypothesis lattice.

For generation, we use an Arabic morphological generator (Habash, 2004) which requires proper specification of the morpho-syntactic features in order to generate the correct inflected form. Clitics are generated separately and then attached as a post-process (El Kholy and Habash, 2010).

3.3 Syntactic challenges

Arabic word order is relatively free, as in Hebrew. This means that there are many possible correspondences between Hebrew and Arabic word orders. Since the dominant word order in Arabic is VSO, the verb and its object are not necessarily consecutive. As a result, the variability of possible sentence structures has to be accounted for on the sentence level, rather than on levels such as VP.

Generating the correct word order in an embedded clause that starts with *An* (Section 2.4) is a complex issue. It requires generation of several different structures at the embedded sentence level, forcing the subtle order constraints according to the embedded sentence structure, and afterwards validating that this was indeed inside an embedded clause.

A major challenge stems from constructions and word formations in Hebrew that do not exist in Arabic. For example, the Hebrew double genetive construction does not directly correspond to an Arabic construction (see Section 2.4). Here, the Hebrew cliticized possessive pronoun must be omitted, and the corresponding Arabic *idafa* structure has to be generated with the proper case assignment.

As we have shown in section 2.4, Arabic poses many syntactic challenges in correctly forcing agreement. For example, N-Adj agreement in verbless sentences whose predicate is an adjectival phrase requires identification of the heads of the subject and the (potentially distant) indefinite adjectival

[4] Such a module is under development. Experiments with available POS taggers resulted in poorer performance.

predicate, and forcing agreement between them:

(14)	*Al+wld*	*Alðy*	*rÂyt*	*+h*
	boy.sg.m.def	REL.sg.m	see.1.sg.past	he.acc
	fy	*Al+mTAr*	*Al+kbyr*	*Twyl*
	in	airport.m.def	big.m.def	tall.m.indef

‘The boy I saw at the big airport is tall’ (Arabic)

In the case of V-S number agreement, when the Arabic form of the verb is generated, the information of whether to place the verb before or after the subject is still unknown (see 3.4). This poses a challenge for correctly generating the output.

A more complex issue is the plural of irrational nouns in Arabic. As demonstrated in 10, any reference to such a noun must use singular feminine agreement features. This requires information about the irrationality of the plural noun, particles that need to agree with it, and enforcement of long distance agreement.

Another challenge is to generate the correct aspectual form of the Arabic imperfective verb in an embedded clause. Since Hebrew does not have an aspectual system, the correct Arabic form must be generated using information that does not originate from the Hebrew verb.

3.4 Computational challenges

Every MT system handles the problem of potential lattice explosion. This is even stronger in translating from and to morphologically rich languages, such as ours. The lack of a morphological disambiguator during analysis magnifies this effect. This issue is especially true in the case of our system, which processes both the source and the target languages bottom-up simultaneously, in order to prune target hypotheses during parsing. Some syntactic choices are hence determined only at relatively late stages, resulting in huge hypothesis spaces in earlier stages.

For every verb the Arabic generator returns 109 possible forms (excluding possible clitics). This is the number of possible results out of the cartesian product of several many-valued morpho-syntactic features: person, gender, number, aspect (perfective, imperfective and imperative), voice (passive or active), and mood (indicative, subjunctive or jussive). For every noun, 72 forms are returned (excluding possible clitics), as a result of the various values of the features gender, number, case, possessiveness and definiteness.

4 Possible approaches

As the standard paradigm of statistical MT is not applicable to Hebrew-to-Arabic MT, due to the dearth of available parallel corpora, two alternatives present themselves. One is translating using a third language (most naturally, English) as a pivot (Muraki, 1987; Wu and Wang, 2007); the other is relying on linguistically-motivated transfer rules, augmented by deep linguistic processing of both the source and the target languages.[5] We consider both approaches below.

4.1 Using English as pivot

The dominant Hebrew-to-Arabic MT system is Google’s.[6] Google has been known to use ‘bridge’ languages in translation (Kumar et al., 2007). We provide evidence that Google’s Hebrew-to-Arabic MT uses English as a pivot, and demonstrate the shortcomings of this approach.[7]

As a first test, we use the number- and gender-ambiguity of second-person pronouns in English (*you*). Since Hebrew and Arabic use separate forms for these pronouns, direct translation is not expected to be ambiguous; however, Google produces the following wrong translations in such cases:

(15)	*atm*	/	*atn*	⟹	*Ant*
	you.pl.m	/	you.pl.f	⟹	you.sg.m/f

amrti	*lkm*	⟹
say.1sg.past	to+you.2.pl.m-dat.	⟹
qlt	*lk*	
say.1sg.past	to+you.2.sg.m/f-gen.	

The second test uses the fact that plural nouns in English are unspecified for gender, whereas in Hebrew and Arabic they are. Here, gender is lost in translation of plurality, and the decoder chose the most common option according to the LM.

(16)	*mwrim*	/	*mwrwt*	⟹	*mςlmyn*
	teachers.m	/	teachers.f	⟹	teachers.m

[5] A third approach is to use comparable corpora (Munteanu and Marcu, 2005); but with no parallel data whatsoever, this is unlikely to succeed.

[6] `http://www.google.com/language_tools`, accessed May 5th, 2010.

[7] Another Hebrew-to-Arabic MT system, `http://www.microsofttranslator.com/`, also uses English as a pivot language, and shows similar characteristics.

In the third test, we use words which are lexically ambiguous in English but not in Hebrew or Arabic.

(17)(a) *Tblh* ⟹ *TAwlħ*
table (data) ⟹ table (furniture)

(b) *bnq* ⟹ *sAHl*
bank (financial) ⟹ bank (shore)

(c) *idni* ⟹ *ktyb*
manual (by-hand) ⟹ manual (booklet)

Finally, we used proper names and morphologically complex words in Hebrew. On hard-to-translate items, Google resorts to transliteration; here, *English* rather than Arabic transliteration was output.

(18) *k+b+mdint* *h+ihwdim* ⟹
as+in+state jews.def ⟹
Achabmdynat *Al+yhwd*
(English transliteration) jews.def

The implication of using a morphologically-poor language as a pivot in translating between two morphologically-rich languages is that much information is lost in the process, and the output tends to be either wrong or ungrammatical. Example 19 summarizes the problems.

(19) *mwrwt* *ipwt* *aklw*
teacher.pl.f.indef pretty.pl.f.indef eat.3.pl.past
⟹ *Aklt* *Almςlmyn*
⟹ eat.3.sg.f.past teacher.pl.m.acc/gen.def
jmylħ
pretty.sg.f.indef
'pretty teachers ate' ⟹ 'teachers ate pretty'

The following issues can be observed: (1) Gender mismatch (feminine *mwrwt* vs. masculine *Almςlmyn*). The reason is that English nouns are unspecified for gender. (2) Number mismatch (plural *ipwt* and singular *jmylħ*). This results in the wrong translation and a disfluency in the target sentence. The reason is that English adjectives are unspecified for number. (3) Definiteness mismatch (Hebrew is indefinite while in Arabic the noun is definite and the adjective is not). (4) Case mismatch: Hebrew is unspecified, Arabic is accusative/genitive (as opposed to the correct case nominative). (5) Verb conjugation error: the gender of the verb that precedes the plural subject *Almςlmyn* is in feminine singular form, although the subject is *rational* plural masculine.

4.2 Transfer-based translation

As an alternative to using English as a pivot language, we advocate a knowledge-based approach. A linguistically-aware transfer approach has several advantages in our case. Source-language morphological analysis provides a tokenization and analysis of the input sentence into morphemes with their morpho-syntactic features. Then, transfer rules and a transfer lexicon map source words and (linguistic) phrases into the target language, bridging over syntactic differences across the languages. Finally, a target-language morphological generator creates inflected morphemes from the yield of the target tree fragments; a subsequent detokenization step then recreates the correct orthographic forms.

We use the Stat-XFER framework (Lavie, 2008), which uses 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 fully-implemented 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 correspondences 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 or Mapudungun-to-Spanish.

Specifically, we use a Hebrew morphological analyzer (Itai and Wintner, 2008), a mediumsized dictionary, an Arabic morphological generator (Habash, 2004), and a tokenized version of the Arabic Gigaword (Graff et al., 2006) corpus as a language model. We manually constructed a grammar, currently consisting of 42 rules. Some rules manipulate morphemes. After decoding (which uses the language model) we detokenize the output sentence in its morpheme representation (El Kholy and Habash, 2010) to produce the final translation. We detail the system below.

5 Solutions

We have successfully implemented many of the problematic issues raised above, focusing on gapping morphological differences and enforcing agreement. We correctly generate and decode Arabic verbs with encliticized object pronouns, NP-internal structure, agreement between subject and adjectival-predicate, Subj–V agreement (on number, gender and person), and the translation of the Hebrew double genitive construction. We implemented rules for the enforcement of N-Adj agreement on rationality and gender, and for V–Prep long distance collocation; but we still do not have the large-scale lexical resources to fully solve these problems.

As an example of a transfer rule, Figure 1 depicts the rule that maps Hebrew phrases such as *hsfr šlkm* 'your (2.pl.m) book' to Arabic phrases like *ktAb +km* 'your (2.pl.m) book'. This is an instance of a Hebrew genitive construction using *šl* 'of' with a cliticized pronoun, mapped into an Arabic construction which uses an enclitic pronoun on the noun.

We discuss below solutions we implemented for some of the challenges listed in Section 3.

N-Adj agreement In local contexts, this is relatively easy, since a simple rule can use unification constraints to force agreement on all features. When the subject and the adjectival predicate are distant, the agreement features of the head of the subject must be propagated up the NP, and agreement is checked at the sentence level.

Irrational plural noun agreement The naïve solution is to lexically determine the rationality of each noun, and let two different rules generate the verb in the correct form according to the subject's rationality (given that the subject is plural). However, information on rationality is not currently available.

```
{NP_POSS,1} # rulename

;;SL: H SPR $L +KM  # source example
;;TL: ktAb +km    # target example

# morpheme POS mapping
NP::NP [NP2 PREP PRO] -> [NP2 PRO]
(
# morpheme aligning
(X1::Y1)
(X3::Y2)

# lexical constraint on SL
((X2 lex) = $L)

# syntactic constraint on TL
((Y1 poss) = +)

# syntactic constraints on SL-TL
((Y1 def) = (*NOT* +))
((Y2 per) = (X3 per))
((Y2 num) = (X3 num))
((Y2 gen) = (X3 gen))

# propagation of features
(X0 = X1)
(Y0 = Y1)
)
```

Figure 1: Example of a transfer rule

Another solution is to generate both the feminine singular form and the plural form with the original gender of the singular form, and let the LM decide. This may solve the problem in local contexts, but as we show in 10, the phenomenon extends to long-distance dependencies.

Our preferred solution is to combine the two approaches. Two hypotheses are generated, one for the rational form and one for the irrational form. Using the rules, we account for complex NPs with relative clauses, and force agreement among all relevant references to the antecedent noun. By propagating the agreement features up to higher levels of the tree, we guarantee that the predicate agrees with the subject NP, whether it is a *regular* rational plural or an *irregular* irrational plural.

V-Subj number agreement Both the singular and the plural forms of the verb may have to be generated. In Hebrew analysis, we can only determine whether the subject is a pronoun or a full NP at a level that contains both subject and verb. Orthogonally, in Arabic generation, we have to decide whether to use the singular form of the Arabic verb and place it before the NP subject, or use the number-agreeing form after the NP subject.

Generating correct aspect Hebrew verbs in the future tense may be translated into the indicative imperfective and subjunctive imperfective forms in Arabic. As the choice is determined by the preceding word, transfer rules are perfectly placed to address the issue. If the preceding word is a preposition denoting intention, we choose the subjunctive form; otherwise, we choose the indicative form. This also reduces the lattice size.

Negated Hebrew verbs in the past tense also have two possible translations: the negated perfective form *mA ktbt* 'I didn't write', and the jussive form with the negative preposition *lm Aktb* 'I didn't write'. We generate both structures and let the LM choose according to local context.

As for other usages of the imperfective jussive tense, these are rare cases that involve specific prepositions. Therefore these constructions are dealt with explicitly using designated transfer rules.

6 Preliminary results

While we still do not have robust evaluation results, we provide in Figure 2 a few example translations of simple phrases to demonstrate the capabilities of the system. We compare our results with Google's.

Example (20 b) demonstrates correct N-Adj agreement for rational and irrational plural nouns and correct treatment of NP conjunction structure. In example (20 c), Google fails on generating the correct constituent structure, lexical translation of 'policewomen' and enforcing agreement, and generates an incoherent result.

Example (21 b) demonstrates correct translation of the preposition, differing word order, V-Subj number agreement in Arabic, and conversion of a possessive construction using *šl* from Hebrew to *Idafa* in Arabic. In example (21 c), Google fails on translating the Hebrew verb correctly, enforcing case, and the correct choice of preposition (*HDr* requires a direct object).

Example (22 b) demonstrates correct translation of the double genitive and verbless predicate constructions. Google's translation is incoherent.

7 Outlook

To our knowledge, we have presented the first computationally oriented discussion of Arabic and Hebrew targeting MT between the two languages. We highlighted the similarities and differences between the two languages and their consequences on the process of MT. We also presented some results comparing to an English-pivot-based approach to Hebrew-Arabic MT.

This is still work in progress and our results are indeed preliminary. However, we demonstrate that our system is capable of producing non-trivial translations, mapping complex morphological and syntactic structures across the two languages in a way that an English-mediated translation fails to achieve. Furthermore, unlike traditional rule-based systems, our approach is fully scalable, and relies on a large target-language model to favor more fluent translations. We are currently incorporating a larger-scale Hebrew-Arabic dictionary and some limited parallel data, overcoming several technical issues involving Arabic morphological generation and implementing more transfer rules.

Acknowledgements

We are grateful to Gennadi Lembersky for his continuous help. Reshef Shilon and Shuly Wintner were partly supported by The Israel Science Foundation (grant No. 137/06). Alon Lavie's work was supported in part by NSF grants IIS-0534217 and IIS-0915327. All of Nizar Habash's contributions were funded by Columbia University's Center for Computational Learning Systems (CCLS).

References

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Ahmed El Kholy and Nizar Habash. 2010. Techniques for Arabic morphological detokenization and orthographic denormalization. In *Proc. of LREC-2010*.

David Graff, Ke Chen, Junbo Kong and Kazuaki Maeda. 2006. Arabic Gigaword. Linguistic Data Consortium.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proc. of HLT-NAACL*.

Nizar Habash, Abdelhadi Soudi, and Timothy Buckwalter. 2007. On Arabic transliteration. In Abdelhadi Soudi,

(20) (a) *mkwniwt* *ipwt* *wšwtrwt* *ipwt*
car.pl.f.indef pretty.pl.f.indef and+policewomen.f.indef pretty.pl.def.indef
'pretty cars and pretty policewomen'

(b) *syArAt* *jmylħ* *wšrTyAt* *jmylAt*
car.pl.f.indef pretty.sg.f and+policewomen.f.indef pretty.pl.f.indef
'pretty cars and pretty policewomen'

(c) *syArAt* *AlšrTħ* *lTyf* *lTyf*
car.pl.f.def police.sg.f.def pretty.sg.m.indef pretty.sg.m.indef
'The police's cars pretty pretty'

(21) (a) *hncigim* *šlkm* *nkxw* *bišibh*
representative.pl.m.def you.pl.m.poss attend.past.3.pl in+meeting.sg.f.def
'your representatives attended the meeting'

(b) *HDr* *mmvlwkm* *Aljlsħ*
attend.past.sg.m representative.pl.m.nom+you.pl.m.poss meeting.def
'your representatives attended the meeting'

(c) *wmmvlwkm* *AlHADryn* *fy* *AlAjtmAς*
and+representative.pl.m.nom+you.pl.m.poss attend.participle.pl.m.def.acc/gen in meeting.def
'And your representatives that attended the meeting'

(22) (a) *mkwnith* *šl* *hmnhlt* *gdwlh*
car.sg.f.def+she.poss of principal.sg.f.def big.sg.f.indef
'The principal's car is big'

(b) *syArħ* *Almdyrħ* *kbyrħ*
car.sg.f.indef principal.sg.f.def big.sg.f.indef
'The principal's car is big'

(c) *Alf* *AlrAysy* *llsyArAt*
thousand main.sg.m.indef to+the+car.pl.f.def
'The cars' thousand main'

Figure 2: Example Hebrew to Arabic translations: (a) Hebrew; (b) our Arabic output; (c) Google's output

Guenter Neumann, and Antal van den Bosch, editors, *Arabic Computational Morphology*. Springer.

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

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

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

Shankar Kumar, Franz J. Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *Proc. of EMNLP-CoNLL*.

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 *Proc. of TMI-2004: Theoretical and Methodological Issues in Machine Translation*.

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.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Kazunori Muraki. 1987. PIVOT: Two-phase machine translation system. In *MT Summit Manuscripts and Program*, pages 81–83.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proc. of RANLP*.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proc. of ACL*.