Noname manuscript No.
(will be inserted by the editor)

Machine Translation between Hebrew and Arabic

Reshef Shilon · Nizar Habash · Alon Lavie · Shuly Wintner

Received: date / Accepted: date

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 on 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 the two systems we are currently developing, for translation in both directions.

Keywords Arabic · Hebrew · Transfer-based MT

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 society.

R. Shilon
Tel Aviv University, Tel Aviv, Israel. E-mail: reshefshilon@yahoo.com

N. Habash
Columbia University, New York. E-mail: habash@ccls.columbia.edu

A. Lavie
Carnegie Mellon University, Pittsburgh. E-mail: alavie@cs.cmu.edu

S. Wintner
University of Haifa, Haifa, Israel. E-mail: shuly@cs.haifa.ac.il

[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.

Machine translation between *very* close languages has of course been addressed in the past (Hajic, 1987; Hajic et al, 2000; Tantug et al, 2007). However, Hebrew and Arabic are not as close as, say, Czech and Slovak or Turkish and Turkmen, so more sophisticated approaches are called for.

The dominant paradigm in contemporary machine translation (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 Hebrew-to-Arabic and Arabic-to-Hebrew MT systems, using Stat-XFER (Lavie, 2008), a particularly suited framework 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 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 systems we are in the process of developing and reports some preliminary results. An early version of this work was published as Shilon et al (2010).

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.

2.1 Orthography

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 and differ only in diacritics. 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, are often realized as diacritics only: e.g., the Arabic orthographic word *wld* ‘boy’ can stand for */waladu/* (nom. def.), */waladū/* (nom.

[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.

[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). Phonetic forms are given between slashes.

indef.), and */waladī/* (gen. indef.), among others. Also, the distinction between the indicative, subjunctive and jussive imperfective forms of most Arabic verbs is lost in some scenarios when the words are non-diacriticized.

In both languages, some prepositions (e.g., *b* 'in, with', *l* 'to'), 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. For example, Hebrew *b+h+kth* 'in+the+classroom' is written *bkth*; and Arabic *l+Al+qlm* 'for the pen' is written *llqlm*. Arabic attaches pronominal direct objects as post-verbal clitics, a construction that, while grammatical, is rarely used in contemporary Hebrew. Hebrew uses the definite direct object marker *at* instead.

(1) *raiti* *awtm*
raiti at +hm
see.1sg.past def.acc they.acc
'I saw them' (Hebrew)

(2) *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 has a very broad meaning, and from templates that add vowels (and, possibly, also consonants) to the root, yielding a lexeme. Hebrew and Arabic have many shared roots. 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* 'addresee' from the same root, which does not exist in Arabic, whereas Arabic has *ktAb* 'book', which does not exist in Hebrew.

2.3 Inflectional morphology

Nouns and adjectives inflect for number, gender and definiteness. In addition, both languages share the difference between the *formal gender* of nouns, which is the gender according to the surface form (expressed in suffixes), and the *functional gender*, which is the gender that is used in agreement. Both languages have many nouns with different formal and functional genders. 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. Another important difference between the two languages is that Arabic encodes case on nouns, whereas Hebrew does not.

Nominals typically come in three varieties (called *states*): absolute, definite and *construct* state, which is used in genitive constructions (see Section 2.4). Feminine nouns in the construct state behave differently: In Hebrew such forms trigger 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'. Construct state

inflection in Arabic and Hebrew is similar in other cases, e.g., the masculine plural suffix *im* (Hebrew) and *yn* (Arabic) is shortened to *i/y*.

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).

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: 1. the perfective form is used for the past tense in Arabic and Hebrew; 2. 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; 3. the imperative; and 4. 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 can 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, 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 (the subject if it is definite, or an expletive pronoun if the subject is indefinite). In addition, the subject of the clause should be in accusative case. Hebrew has no parallel construction. On the other hand, when a Hebrew sentence begins with an adverbial, 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. 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 of 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) *Al+qlm* *Al+jmyl*
pen-m.sg.def pretty.m.sg.def
'The pretty pen' (Arabic)

(5) *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.

Subject–verb In both languages the verb and the subject NP agree on person, number and gender. However, in Arabic VSO sentences the verb is always in singular:

(6) *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 adjectival phrase. 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:

(7) *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 (8).

In Hebrew, but not in Arabic, such relations can also be expressed in a different construction, using the possessive preposition *šl* 'of'. 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 a *šl* PP (9).

(8) *iwm* *h+hwldt*
day.m.sg.indef the+birth.f.sg
'The birthday' (Hebrew)

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

Pro-drop In both languages, a subject pronoun can be omitted if the verb is in past, 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. In the following sentence, the relativizer and the encliticized pronoun agree with the antecedent irrational plural noun, and therefore are feminine singular:

(10) *Al+ÂqlAm* *Alty* *Ăštrý+hA* *Al+wld*
pen-m.pl.def REL.f.sg 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 Example 10. Relative clauses that modify indefinite nouns have no relativizer, as in Example 11. The Hebrew relative clause alwyas starts with a relativizer which carries no agreement features.

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

rÂyt *wldA* *qrÂ* *ktAbA*
see.1st.sg.past boy.sg.m.indef 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+*, which can be used in this function only if the embedded verb is in the present. A similar phenomenon in Arabic uses 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 these issues 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) *nkx* *b+* *h+pgišh*
attend.3sg.m.past in+ meeting.def
'he attended the meeting' (Hebrew)

(13) *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 distances V–O dependencies, since the subject may intervene between the verb and its preposition.

(14) *Âṣrb* *rŷys* *Al+Hkwmħ* *ywm* *Al+ÂrbṣA'* *fy* *jlsħ*
express.3sg.m.past leader government.def day Wednesday in meeting
Al+Hkwmħ *Al+Âsbwṣyħ* *ṣn* *Âml* *+h* ...
government.def weekly.def upon hope he.poss
'The prime-minister expressed Wednesday during the government weekly meeting his hope ...'

This example demonstrates the potential 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 statistical, n-gram-based language model to correctly select the preposition.

Another lexical challenge stems from the fact that existing Arabic lexical resources (Buckwalter, 2004; Habash, 2004) 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 morphological analysis, transfer and generation. The complex morphology induces inherent data sparsity problems, magnifying the limitation imposed by the dearth of available parallel corpora (Habash and Sadat, 2006).

Translating from Hebrew to Arabic, 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 Habash (2004) which requires proper specification of the morpho-syntactic features in order to generate the correct inflected form. Clitics are generated separately and are then attached as a post-process (El Kholy and Habash, 2010).

In the Arabic-to-Hebrew direction we use Habash (2004) as a morphological analyzer and morphological disambiguator. This helps us reduce the amount of hypotheses in the lattice. For generation we use the reverse direction of Itai and Wintner (2008) as a generator, which inflects better for gender than its Arabic counterpart. Due to the morphological disambiguator in Arabic and the generator in Hebrew, translation in this direction currently performs better.

3.3 Syntactic challenges

Several possible correspondences between Hebrew and Arabic word order may exist. 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* (see 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 (when the relative clause is combined with the relativizer) 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.

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

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

(15) *Al+wld* *Alðy* *rÂyt* *+h* *fy* *Al+mTAr* *Al+kbyr*
boy.sg.m.def REL.sg.m see.1.sg.past he.acc in airport.m.def big.m.def
Twyl
tall.m.indef
'The boy I saw at the big airport is tall' (Arabic)

In the case of subject–verb agreement on number, when the Arabic form of the verb is generated, it is unknown whether the verb will be placed before or after the subject. This poses a challenge for generating the correct form of the verb.

A more complex issue is the plural form 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 source.

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 enhances 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 determined only at relatively late stages, resulting in huge hypothesis spaces earlier.

For every verb the Arabic morphological 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 not), 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.

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

4.1 Using English as pivot

The dominant Hebrew-to-Arabic MT system, Google,[6] 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 (Hebrew on the left, Arabic on the right of the arrows):

(16)

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

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

klb+km	⟹	*Alklb*
dog.sg+poss.2.pl.m	⟹	dog.sg.def

'your dog' ⟹ 'the dog'

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 language model.

(17)

mwrim	/	*mwrwt*	⟹	*mʕlmyn*
teachers.m	/	teachers.f	⟹	teachers.m

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

(18)

Tblh	⟹	*TAwlħ*	*bnq*	⟹	*sAHl*
table (data)	⟹	table (furniture)	bank (financial)	⟹	bank (shore)

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

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

(19)

mwrwt	*ipwt*	*aklw*	⟹
teacher.pl.f.indef	pretty.pl.f.indef	eat.3.pl.past	

Aklt	*Almʕlmyn*	*jmylħ*
eat.3.sg.f.past	teacher.pl.m.acc/gen.def	pretty.sg.f.indef

'pretty teachers ate' ⟹ 'teachers ate pretty'

[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.

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 nominative case). 5. Verb conjugation error: 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.

Consider the example of Figure 1. This is an augmented synchronous context-free rule that maps correspondences between the source language (SL) and the target language (TL). This rule maps a SL noun phrase to a TL noun phrase, hence `NP::NP`. Furthermore, the rule specifies that the SL noun phrase is built up from `NP2 PREP PRO`, whereas the TL noun phrase has a different structure, namely `NP2 PRO`. But this is not all: each of the non-terminals on both sides of the rules is associated with a feature structure that encodes more detailed information, and constraints can be imposed on the feature structures that prevent the rule from firing. Specifically, SL feature structures are pointed to by indexed X-s, whereas TL feature structures are referred to as Y-s. Thus, the specification `(X1::Y1)` says that the first element in the body of the SL rule (`NP2`) corresponds to the first daughter of the TL rule (again, `NP2`). Furthermore, the specification `((Y1 poss) = +)` means that the rule can only fire of the value of the `poss` feature of the TL `NP2` is `+`. Other constraints in this rule verify that features of the TL `PRO` (referred to as `Y2`) correspond to values of the SL `PRO`, namely `X3`.

Rules such as the one exemplified above inform the *transfer engine* of Stat-XFER, which 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. The output of the engine is a lattice of alternative translation segments, arising from syntactic ambiguity, lexical ambiguity and multiple translation equivalents of lexical items. The other component of the system is a monotonic *decoder*, used to create complete translation hypotheses from the lattice. The task of the decoder is to select a linear sequence of adjoining but non-overlapping translation units that maximizes the overall score of the TL string given the SL string, using a beam-search that controls the underlying parsing and transfer process.

```
{NP_POSS,1}       # rulename
;;SL: H SPR $LKM  # source example
;;TL: ktAb +km    # target example

# morpheme POS mapping
NP::NP [NP2 PREP PRO] -> [NP2 PRO]
(
# morpheme alignment
(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)
)
```

Fig. 1 Example of a transfer rule

Scores are based on a log-linear combination of several features, including a TL language model, rule probabilities, a measure of fragmentation and the source-to-target relative sentence length.

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, 2004b), 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 (Monson et al, 2008).

Specifically, for our Hebrew-to-Arabic system we use a Hebrew morphological analyzer (Itai and Wintner, 2008), a medium-sized dictionary, an Arabic morphological generator (Habash, 2004), and a tokenized version of the Arabic GigaWord corpus as a language model. We manually constructed a grammar, currently consisting of over 40 rules, 21 of which are NP rules. Some rules manipulate bound 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. For our Arabic-to-Hebrew system, we use the same components in the reverse direction, adding an Arabic morphological disambiguator (Habash, 2004), and using a tokenized version of Hebrew as a language model. We detail both systems below.

5 Transfer-based Hebrew-Arabic machine translation

We created two Stat-XFER MT systems, translating from Hebrew to Arabic and from Arabic to Hebrew, whose transfer rules successfully implement solutions for many of the problematic issues raised in Section 3, focusing on gapping morphological differences and enforcing agreement. We correctly generate and decode both Arabic and Hebrew verbs with encliticized object pronouns, NP-internal structure, agreement between subject and adjectival-predicate, and subject–verb agreement (on number, gender and person). We also correctly translate structures that do not exist in the target language, such as the Hebrew definite accusative marker *at*, the genitive *šl* and double genitive constructions, and the Arabic future

markers *swf* and *s+*. We implemented rules to enforce agreement on rationality and gender between nouns and adjectives, and to relate verbs to their subcategorized prepositions; but we still lack the large-scale lexical resources needed to fully solve some of these problems.

As an example, refer back to the transfer rule of Figure 1. It 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 now discuss solutions we implemented for some of the challenges listed in Section 3.

Subject–predicate 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. This rule is depicted in Figure 2.

```
{S_NP_ADJ,1}     # rule name
;;SL: H ILD GDWL # source example
;;TL: Alwld kbyr # target example
# morpheme POS mapping
S::S [NP ADJP] -> [NP ADJP]
(X1::Y1)     # morpheme alignment
(X2::Y2)
# Hebrew side agreement
((X1 def) = +)
((X2 def) = -)
((X1 num) = (X2 num))
((X1 gen) = (X2 gen))
# Arabic side agreement
((Y1 rational) = +)
((Y1 def) = +)
((Y2 def) = -)
((Y1 num) = (Y2 num))
((Y1 gen) = (Y2 gen))
((Y1 case) = nominative)
((Y2 case) = nominative)
```

Fig. 2 Subject–predicate agreement

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. 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 language model 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. See Figure 3

Subject–verb number agreement Recall that the Arabic verb is in singular if it precedes the subject. Therefore, 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. This decision is taken when we handle the sentence level, where we know whether the subject NP is pronominal or not, and can deduce the word order. See Figure 4

```
{S_NP_ADJ_IRRAT,1}
;;SL: H$WLXNWT GDWLIM
;;TL: AlTAwlAt kbyrp
S::S [NP ADJP] -> [NP ADJP]
(X1::Y1)
(X2::Y2)
# Hebrew side agreement
((X1 def) = +)
((X2 def) = -)
((X1 num) = (X2 num))
((X1 gen) = (X2 gen))
# Arabic side agreement
((Y1 rational) = -)
((Y1 def) = +)
((Y2 def) = -)
((Y1 num) = plural)
((Y2 num) = singular)
((Y2 gen) = feminine)
((Y1 case) = nominative)
((Y2 case) = nominative)
```

Fig. 3 Irrational plural noun agreement

```
{S_VB_NP_swap, 1}
;; SL: HLIDIM AKLW
;; TL: Akl AlAwlAd
S::S [NP VB] -> [VB NP] # POS mapping
(X1::Y2)                # POS alignment
(X2::Y1)
# Hebrew side agreement
((X1 num) = (X2 num))
((X1 gen) = (X2 gen))
((X1 per) = (X2 per))
# Arabic side agreement
((Y1 num) = singular)
((Y1 per) = (Y2 per))
```

Fig. 4 Subject–verb number agreement

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 (Figure 5) 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.

```
{VERB_NEG_lm,0}
;;SL: LA AKL
;;TL: lm yAkl
VB::VB [NEG V] -> ["lm" V]
(X2::Y2)
((X2 tense) = past)
((Y2 aspect) = imperfect)
((Y2 mood) = jussive)
((Y2 per) = (X2 per))
((Y2 gen) = (X2 gen))
((Y2 voice) = (X2 voice))
```

Fig. 5 Verb negation

6 Preliminary results and evaluation

The two MT systems are now fully implemented, although their coverage is still limited. To evaluate the performance of the systems, we created two test sets, one for each direction. All sentences in our development and test sets are extracted from newspaper texts; the Hebrew

reference corpus was manually translated by two translators to Arabic, whereas for the Arabic reference corpus we obtained three translations. In the Hebrew-to-Arabic test set, 84% of the Hebrew side morphemes had at least one entry in our bilngual lexicon, and 87% of the Arabic side morphemes in the Arabic-to-Hebrew test set had at least one such entry.

As the systems are still under development, and several components are not yet functioning at full scale, we constrain the evaluation to smaller, simpler sentences for which we have good lexical coverage of the source language sentence. We selected all sentences of length 10 (words) or less, with at most one totally unknown morpheme in our lexicon. This resulted in a set of 39 sentences in the Hebrew-to-Arabic system, 28 sentences in the Arabic-to-Hebrew system. Out-of-vocabulary morphemes in the input sentences were manually completed for the smaller evaluation sets. As a baseline, we use the same systems with no grammar rules. Figure 6 depicts actual translations produced with the systems on some of our development set sentences.

(20) 1. *Âς ln* *mSdr* *rsmy* *sdAny* *Ân* *TyAryn* *rwsyyn*
inform.past.3ms source.sg official.sg Sudani.sg that pilot.du Russian.du
AxtTfA
kidnap.past.3.du
'Official Sudani sources informed that two Russian pilots were kidnapped' (Arabic input)

2. *mqwr* *ršmi* *swdni* *hwdi'* *š* *šni* *Tiisim* *rwsim*
source.sg official.sg Sudani.sg inform.past.3ms that two pilot.m.pl Russian.pl
xTpw
kidnap.pl
'Official Sudani sources informed that two Russian pilots kidnapped' (Hebrew with gramamr)

3. *hwdi'w* *mqwrwt* *ršmiim* *swdni* *š* *Tiis* *rwsih* *xTwp*
inform.past.3mp source.pl official.pl Sudani.sg that pilot.m.sg Russia kidnap.sg.passive
'Informed official Sudani sources that a pilot Russia kidnapped' (Hebrew without grammar)

(21) 1. *Âkd* *AlHryry* *Ân* *ςlAqt+h* *mς* *swryA* *ttTwr* *ĂijAbA*
emphesize.past.3ms AlHariri that relation+his with Syria evolve.fut.3fs posistively
'AlHariri confirmed that his relations with Syria are evolving positively' (Arabic input)

2. *hxriri* *ašr* *š+* *qšr* *+w* *'m* *swrih* *hštnh* *xiwbi*
the-Hariri confirm.past.3ms that relation his with Syria change.past.3ms positive.m.sg
'The Hariri confirmed that his relations with Syria changed positive' (Hebrew with gramamr)

3. *ašr* *h-xriri* *š+* *qšr* *awtw* *'m* *swrih* *hštnh*
confirm.past.3ms the-Hariri that relation him with Syria change.past.3ms
xiwbi
positive.m.sg
'Confirmed the Hariri that relation him with Syria changed positive' (Hebrew without grammar)

(22) 1. *xbri+h* *šl* *h+nšiah* *hm* *mšpTnim* *m'wlim*
friend.m.pl+her of the+president.f.sg they lawyer.m.pl excellent.m.pl
'The president's friends are excellent lawyers' (Hebrew input)

2. *ÂςDA'* *Alrŷys* *hm* *mHAmwn* *mmtAzħ*
member.pl the+president.m.sg they lawyer.m.pl.nom excellent.m.sg.nom/gen
'The members of the president are excellent lawyer' (Arabic with gramamr)

3. *Sdyq+y* *Ân* *Ăly* *Al+rŷys+h* *mHAmyAã* *mmtAzAã*
friend+my that to the+president+his lawyer.sg.acc excellent.sg.acc
incoherent (Arabic without gramamr)

Fig. 6 Translation examples: Arabic to Hebrew (20,21) and Hebrew to Arabic (22)

Consider (20): the translation reflects correct transfer of number and enforcement of N-Adj agreement in both NPs. In addition, the dual form in Arabic, which does not produc-

tively exist in Hebrew, is properly translated into the plural form in the noun, adjective and verb, and the explicit *'šni'* 'two' is generated in the correct gender. However, the passive form of the verbs is not properly generated. In the baseline system (20.3), agreement is violated in both NPs, the dual number is not properly handled, and the Arabic adjective *rwsyyn* ('Russian') is assigned the wrong POS.

In (21), the grammar-based system (2) correctly generates the possessive prounoun (as opposed to (3)), while in both systems the proper name *AlHryry* is not properly translated. In (22), the grammar-based system (2) correctly handles the Hebrew double genitive construction, translating it to the Arabic genitive construction, and correctly treats the nominal predicate construction. There are still errors in N-Adj agreement on number and gender, and in the translation of the subject noun *nšiah* (wrong gender). These issues arise from lattice exploision. The baseline translation (3), on the other hand, is totally incoherent.

We also report automatic evaluation results on this simplified test set. Table 6 lists BLEU (Papineni et al, 2002) and METEOR (Lavie et al, 2004a) scores for both systems.

	With rules		Without rules	
	BLEU	METEOR	BLEU	METEOR
Hebrew to Arabic	0.107	0.301	0.143	0.310
Arabic to Hebrew	0.275	0.467	0.231	0.417

Table 1 Evaluation results

Evidently, the Arabic to Hebrew system performs much better than the Hebrew to Arabic one. The grammar yields a significant improvement in the Arabic-to-Hebrew system, but it actually damages the Hebrew-to-Arabic system. The main reason for the deterioration in quality of translation using the grammar is lattice explosion, due to the great number of hypotheses. This is caused by two major factors: (1) Lacking a high-quality morphological disambiguator for Hebrew; and (2) the number of possibilities returned by the Arabic generator. When using a smaller bilingual lexicon with fewer translation options, the output is far better. We are currently working on ways to solve this issue, by incorporating a morphological disambiguator for Hebrew, and minimizing the number of results returned by the Arabic generator by merging results with identical surface forms and different feature structures.

To better understand these results, we performed a deep analysis of five sentences in each direction, focusing on the various potential sources of errors during the translation process. Table 6 lists the number of errors that can be attributed to each component: lexicon, grammar, decoder (when the correct hypothesis is present in the lattice but not selected), morphological analyzer, generator and disambiguation module.

	Lexicon	Grammar	Decoder	Analyzer	Generator	Disambiguation
H2A	14	11	4	3	1	
A2H	5	11	3		1	2

Table 2 Number of errors by type

We now take a closer look at one of the sentences in the smaller test set. The Arabic input is *nHn dAŷmAã nqwl lhm AðhbwA wqçwA AlĂtfAq* ('we always tell them: go sign the agreement') and the Hebrew references are *anxnw tmid awmrim lhm lkw xtmw 'l hhskm*, and twice *anw tmid awmrim lhm: lkw, xtmw 'l hhskm*. Our Arabic-to-Hebrew system produces the following output:

(23)	*anxnw*	*tmidi*	*amr*	*lhm*	*kli*	*hhskm*	*ngn*
	we	constant	tell.past.3rd.sg.masc	to them	tools	the agreement	play

Several errors occur in the translation of this sentence:

- The Arabic lexical entry dAŷmAã is not matched. The reason is that our lexicon is specified for case diacritics, whereas the analyzer's output does not include them.
- The pair *wq∼'* ⇔ *xtm* 'sign' is missing from the dictionary.
- The Arabic disambiguation module wrongly chooses the verbal template (*Âðohab-a* instead of *ðahab-a*), and predicts the wrong aspect (perfective instead of imperative)
- A rule for Subj-ADV-V is missing. As a result, subject-verb agreement is not enforced.
- The grammar lacks a rule for translating Arabic imperfective to Hebrew present tense.
- The grammar lacks a rule that inserts the preposition *'l* 'on'; there is no matching preposition in the Arabic input.

From the detailed error analysis and its numerical summary a clearer picture of the development status appears, where the grammar and lexicon are the crucial factors responsible for most of the errors. While augmenting and tuning the rules in the grammar is relatively easy, augmenting the bilingual lexicon is a hard task that currently remains open.

7 Outlook

To our knowledge, this is 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 discussed the shortcomings of a English-pivot-based approach to Hebrew-Arabic MT. Finally, we presented preliminary evaluation results on small evaluation sets of short, simple sentences.

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 Hebrew morphological disambiguation, and implementing more transfer rules for both systems.

Acknowledgements We are grateful to Gennadi Lembersky for his help. This research was supported by THE ISRAEL SCIENCE FOUNDATION (grant No. 137/06).

References

Brown PF, Cocke J, Della Pietra SA, Della Pietra VJ, Jelinek F, Lafferty JD, Mercer RL, Roossin PS (1990) A statistical approach to machine translation. Computational Linguistics 16(2):79–85

Buckwalter T (2004) Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, Philadelphia

El Kholy A, Habash N (2010) Techniques for Arabic morphological detokenization and orthographic denormalization. In: Proceedings of LREC-2010

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

Habash N, Sadat F (2006) Arabic preprocessing schemes for statistical machine translation. In: Moore RC, Bilmes JA, Chu-Carroll J, Sanderson M (eds) HLT-NAACL, The Association for Computational Linguistics

Habash N, Soudi A, Buckwalter T (2007) On Arabic transliteration. In: Soudi A, Neumann G, van den Bosch A (eds) Arabic Computational Morphology, Text, Speech and Language Technology, vol 38, Springer, chap 2, pp 15–22, URL http://dx.doi.org/10.1007/978-1-4020-6046-5_2

Hajic J (1987) Ruslan: An MT system between closely related languages. In: Proceedings of the 3rd Conference of The European Chapter of the Association for Computational Linguistics, pp 113–117

Hajic J, Hric J, Kubon V (2000) Machine translation of very close languages. In: Proceedings of the Sixth Conference on Applied Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, pp 7–12, DOI 10.3115/974147.974149, URL http://www.aclweb.org/anthology/A00-1002

Hanneman G, Ambati V, Clark JH, Parlikar A, Lavie A (2009) An improved statistical transfer system for French–English machine translation. In: StatMT '09: Proceedings of the Fourth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Morristown, NJ, USA, pp 140–144

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

Kumar S, Och FJ, Macherey W (2007) Improving word alignment with bridge languages. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Association for Computational Linguistics, Prague, Czech Republic, pp 42–50, URL http://www.aclweb.org/anthology/D/D07/D07-1005

Lavie A (2008) Stat-XFER: A general search-based syntax-driven framework for machine translation. In: Gelbukh AF (ed) CICLing, Springer, Lecture Notes in Computer Science, vol 4919, pp 362–375

Lavie A, Vogel S, Levin L, Peterson E, Probst K, Llitjós AF, Reynolds R, Carbonell J, Cohen R (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, DOI http://doi.acm.org/10.1145/974740.974747

Lavie A, Sagae K, Jayaraman S (2004a) The significance of recall in automatic metrics for mt evaluation. In: Frederking RE, Taylor K (eds) AMTA, Springer, Lecture Notes in Computer Science, vol 3265, pp 134–143

Lavie A, Wintner S, Eytani Y, Peterson E, Probst K (2004b) 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

Monson C, Font Llitjós A, Ambati V, Levin L, Lavie A, Alvarez A, Aranovich R, Carbonell J, Frederking R, Peterson E, Probst K (2008) Linguistic structure and bilingual informants help induce machine translation of lesser-resourced languages. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), URL http://www.lrec-conf.org/proceedings/lrec2008/

Munteanu DS, Marcu D (2005) Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics 31(4):477–504, DOI http://dx.doi.org/10.1162/089120105775299168

Muraki K (1987) PIVOT: Two-phase machine translation system. In: MT Summit Manuscripts and Program, pp 81–83

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp 311–318, DOI http://dx.doi.org/10.3115/1073083.1073135

Shilon R, Habash N, Lavie A, Wintner S (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

Tantug AC, Adali E, Oflazer K (2007) Machine translation between turkic languages. In: Proceedings of ACL 2007, Companion Volume, The Association for Computer Linguistics

Varga D, Halácsy P, Kornai A, Nagy V, Németh L, Trón V (2005) Parallel corpora for medium density languages. In: Proceedings of RANLP'2005, pp 590–596

Wu H, Wang H (2007) Pivot language approach for phrase-based statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Association for Computational Linguistics, Prague, Czech Republic, pp 856–863, URL http://www.aclweb.org/anthology/P07-1108