Challenges in Building an Arabic-English GHMT System with SMT Components

Nizar Habash[†], **Bonnie Dorr**[‡], **Christof Monz**[§]
[†]Center for Computational Learning Systems, Columbia University
habash@cs.columbia.edu
[‡]Department of Computer Science, University of Maryland
bonnie@umiacs.umd.edu
[§]Department of Computer Science, Queen Mary, University of London
christof@dcs.qmul.ac.uk

Abstract

The research context of this paper is developing hybrid machine translation (MT) systems that exploit the advantages of linguistic rule-based and statistical MT systems. Arabic, as a morphologically rich language, is especially challenging even without addressing the hybridization question. In this paper, we describe the challenges in building an Arabic-English generation-heavy machine translation (GHMT) system and boosting it with statistical machine translation (SMT) components. We present an extensive evaluation of multiple system variants and report positive results on the advantages of hybridization.

1 Introduction

The research context of this work is developing hybrid machine translation (MT) systems that exploit the advantages of linguistic rule-based and statistical MT systems. Arabic, as an example of a morphologically rich language, is especially challenging even without addressing the hybridization question. In this paper, we describe the challenges in building an Arabic-English generation-heavy machine translation (GHMT) system (Habash, 2003a) and extending it with statistical machine translation (SMT) components.

A major challenge for working with Arabic is the proliferation of inconsistent morphological representations in different resources and tools for Arabic natural language processing (NLP) (Habash, 2006). This inconsistency is heightened when one attempts to combine techniques used in linguistically-aware MT approaches with those of surface-based statistical MT approaches, where the level of representation of the phrase (beyond the word) is different for each of these two approaches. We describe how we address this issue in our system and present an extensive evaluation addressing its various strengths and weaknesses. We show positive improvements when extending our basic GHMT system with SMT components.

The remainder of this paper is organized as follows: the next section (Section 2) discusses previous work on hybridization in MT. This is followed by a discussion of Arabic-specific challenges for MT implementations in Section 3. Section 4 describes the Arabic components of our basic GHMT system. Section 5 describes the extensions we made to integrate SMT components into the GHMT system. Section 6 presents three evaluations of multiple MT system variants.

2 Previous Work

We discuss research related to our approach in the areas of generation-heavy MT and MT hybridization.

2.1 Generation-Heavy MT

GHMT is an asymmetrical hybrid approach that addresses the issue of MT resource poverty in source-poor/target-rich language pairs by exploiting symbolic and statistical target-language resources (Habash and Dorr, 2002; Habash, 2003a; Habash,

Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 56-65,
Cambridge, August 2006.

2003b). Expected source-language resources include a syntactic parser and a simple one-to-many translation dictionary. No transfer rules or complex interlingual representations are used. Rich target-language symbolic resources such as word lexical semantics, categorial variations and subcategorization frames are used to overgenerate multiple structural variations from a target-language-glossed syntactic dependency representation of source-language sentences. This symbolic overgeneration accounts for possible translation divergences, cases where the underlying concept or "gist" of a sentence is distributed differently in two languages such as *to put butter on* and *to butter* (Dorr, 1993). The overgeneration is constrained by multiple statistical target-language models including surface n-grams and structural n-grams. The source-target asymmetry of systems developed in this approach makes them more easily retargetable to new source languages (provided a source-language parser and translation dictionary). In this paper, we describe these two specific extensions for Arabic in detail (Section 4).

2.2 MT Hybridization

Research into MT hybrids has increased over the last few years as research in the two main competing paradigms—rule-based MT and corpus-based (statistical) MT—is approaching a plateau in performance. In the case of statistical approaches this has recently led to approaches that rely not just on surface forms but also on symbolic knowledge such as morphological information and syntactic structure. In the next two subsections, we review this body of research. Our own research however, differs in that we are approaching the hybridization question from the opposite direction, i.e., how to incorporate SMT components into rule-based systems (Senellart, 2006). Nonetheless, the research on SMT-based hybrids has influenced many of our decisions and directions.

2.2.1 Morphology-Based Approaches

The anecdotal intuition in the field is that reduction of morphological sparsity often improves translation quality. This reduction can be achieved by increasing training data or via morphologically-driven preprocessing (Goldwater and McClosky, 2005). Recent investigations of the effect of morphology on SMT quality focused on morphologically rich languages such as German (Nießen and Ney, 2004); Spanish, Catalan, and Serbian (Popović and Ney, 2004); and Czech (Goldwater and McClosky, 2005). These studies examined the effects of various kinds of tokenization, lemmatization and part-of-speech (POS) tagging and showed a positive effect on SMT quality.

Lee (2004) investigated the use of automatic alignment of POS-tagged English and affix-stem segmented Arabic to determine appropriate tokenizations of Arabic. Her results showed that morphological preprocessing helps, but only for smaller corpora. Habash and Sadat (2006) reached similar conclusions on a much larger set of experiments including various preprocessing schemes and techniques. They showed that genre variation interacts with preprocessing decisions.

Within our approach, working with Arabic morphology is especially challenging. We discuss this issue in more detail in Section 3.

2.2.2 Syntax-Based Approaches

More recently a number of statistical MT approaches included syntactic information as part of the preprocessing phase, the decoding phase or the n-best rescoring phase.

Collins et al. (2005) incorporated syntactic information as part of preprocessing the parallel corpus. A series of transformations on the source parse trees were applied to make the order of the source-language words and phrases closer to that of the target language. The same reordering was done for a new source sentence before decoding. They showed a modest statistically significant improvement over basic phrase-based MT.

Quirk et al. (2005) used sub-graphs of dependency trees to deal with word-order differences between the source and the target language. During training, dependency graphs on the source side were projected onto the target side by using the alignment links between words in the two languages. The use of syntactic information is the main difference between their approach and phrase-based statistical MT approaches. During decoding, the different subgraphs were combined in order to generate the most likely dependency tree. This approach has been shown to provide significant improvements over a

phrase-based SMT system.

Och et al. (2004) experimented with a wide range of syntactic features to rescore the n-best lists generated by their statistical MT system. Although some features—e.g., POS tags and parse-tree to string mappings—led to slight improvements over the baseline, larger improvements were obtained by using simpler, non-syntactic features, such as IBM Model 1 alignments.

Similar to Collins et al. (2005) and Quirk et al. (2005), our approach uses source-language syntactic (specifically dependency) representations to capture generalizations about the source-language text. Unlike both of them, we do not use or learn specific mappings between the syntactic structure of source and target languages. Instead, our approach maps the source language to a syntactically language-independent representation which forms the basis for target-language generation.

3 Arabic Challenges

Arabic is a morphologically complex language with a large set of morphological features. These features are realized using both concatenative (affixes and stems) and templatic morphology (root and patterns) with a variety of morphological and phonological adjustments that appear in word orthography and interact with orthographic variations. As a result, there are many different possible representations of Arabic morphological tokens that have been used in different resources for Arabic NLP (Habash, 2006).

For statistical MT, in principle, it does not matter what level of morphological representation is used so long as the input is on the same level as that of the training data. However, in practice, there are certain concerns with issues such as sparsity, ambiguity, and training data size. Symbolic MT approaches tend to capture more abstract generalizations about the languages they translate between compared to statistical MT. This comes at a cost of being more complex than statistical MT, involving more human effort, and depending on already existing resources for morphological analysis and parsing.

This dependence on existing resources highlights the problem of variation in morphological representations for Arabic. In a typical situation, the input/output text of an MT system is in simple white-space tokenization. But, a statistical parser (such as (Collins, 1997) or (Bikel, 2002)) trained out-of-the-box on the Penn Arabic Treebank (Maamouri et al., 2004) assumes the same kind of tokenization it uses (4-way normalized segments into conjunction, particle, word and pronominal clitic). This means a separate tokenizer is needed to convert input text to this representation (Habash and Rambow, 2005; Diab et al., 2004).

An additional issue with a treebank-trained statistical parser is that its input/output is in normalized segmentation that does not contain morphological information such as features or lexemes that are important for translation. Arabic-English dictionaries use lexemes and proper translation of features, such as number and tense, requires access to these features in both source and target languages. As a result, additional conversion is needed to relate the normalized segmentation to the lexeme and feature level. Of course, in principle, the treebank and parser could be modified to be at the desired level of representation (i.e. lexeme and features). But this may be a labor-intensive task for researchers interested in MT.

4 Extending GHMT to Arabic

As described earlier, our English-targeted GHMT system can be used with a new source language given that a dependency parse and a word-based translation lexicon are provided. In the following sub-sections, we describe these two components in our Arabic-English GHMT system. The reusable English generation component in GHMT is called EXERGE (Expansive Rich Generation for English). It is discussed in detail in (Habash, 2003a).

4.1 Analysis Issues

This sub-section describes the necessary steps for processing an Arabic input sentence.

4.1.1 Tokenization and POS Tagging

For tokenization, we use the Penn Arabic Treebank (PATB) tokenization scheme, which is most compatible with statistical parsers trained on the PATB (Maamouri et al., 2004). For the POS tagset, we use the collapsed tagset for PATB (24 tags). We use the Morphological Analysis and Disambiguation

(MADA) tool for Arabic preprocessing (Habash and Rambow, 2005) together with TOKAN, a general tokenizer for Arabic (Habash, 2006). MADA uses the ALMORGEANA (Arabic Lexeme-based Morphological Analysis and Generation) system, which is an alternative engine to Buckwalter's AraMorph that uses the same lexical files.

4.1.2 Chunking

We employ a rule-based segment chunker to address two issues. First, the Arabic sentence length, which averages over 35 words with PATB tokenization (in the news genre), slows down the parser and increases its chances of producing null parses. Second, the use of punctuation and numbers in by-lines in news requires template handling in analysis and generation, which needs to be updated depending on the genre. Instead, we choose to preserve source-language order for such cases by chunking them out and treating them as special chunk separators that are translated independently. The rules currently implemented use the following chunk separators. POS information is used in this process.

- Arabic conjunction proclitic `w/CC`[1] *and*
- Numbers (`CD`) and punctuation (`PUNC`)
- The subordinating conjunction `An/IN` *that*

On average, sentences had 3.3 chunk separators.

4.1.3 Parsing

For parsing, we use the Bikel parser (Bikel, 2002) trained on the PATB (Part 1). The default output of the parser is an unlabeled constituency representation. The tokens in the parser are surface words in the PATB tokenization scheme.

4.1.4 Postparsing

The specifications of EXERGE require an input dependency tree labeled with minimal syntactic relations (subj, obj, obj2, and mod). Moreover, the nodes must have lexemes and features from a pre-specified set of feature names and values (Habash, 2003a). The output of the parsing step undergoes operations such as relation labeling and node-structure modification. Some of these operations are similar to the Spanish post-parsing processing for Matador (Spanish-English GHMT) (Habash, 2003b).

Constituency to Dependency We convert constituencies to dependencies using modified head-percolation rules from Bikel parser applied with the Const2Dep tool[2] (Habash and Rambow, 2004).

Lexeme Selection MADA is only a morphological disambiguation tool that makes no sense-disambiguation choices. Therefore, multiple lexemes are still available as ambiguous options at the tree nodes. In some cases, the parser overrides the POS tag that was chosen initially by MADA. As a result, we need to re-visit discarded morphological analyses again. We re-apply the ALMORGEANA system on the tokenized words and then filter analyses using the following criteria. In case no analysis matches, all analyses are passed on to the next filter.

- Analyses with PATB tokenizable clitics are ignored because the word is already tokenized.
- Analyses that match the word's POS are selected. Others are ignored. The POS matching is fuzzy since the tagset used by ALMORGEANA (15 tags) is more coarse-grained than the PATB tagset (24 tags). Also, since there are common cases of mismatch in Arabic, certain seemingly mismatched cases are allowed, e.g., noun, adjective and proper noun.
- We use a statistical unigram lexeme and feature model. The model was trained on PATB (part 1 and part 2) and 1 million words from Arabic Gigaword (Graff, 2003) disambiguated using MADA. The lexemes are chosen based on their unigram counts. Ties are broken with feature unigrams.

Dependency Tree Restructuring The following operations are applied to the dependency tree:

- Idafa Handling: The Idafa construction is a syntactic construction indicating the relationship of possession between two nouns, i.e., Noun1 *of* Noun2. Nouns in this construction

[1] All Arabic transliterations in this paper are provided in the Buckwalter transliteration scheme (Buckwalter, 2002).

[2] The Const2Dep tool was provided by Rebecca Hwa.

are modified to include an intervening node that has no surface value but is glossed to *of/'s/*empty**.

- The untokenized prefix `Al+` *the* is turned into a separate node that is a dependent on the word to which it is attached.

- Feature mapping: We map Arabic-specific features to language-independent features used in EXERGE. For example, the untokenized prefix `s+` *will* is mapped to the feature `TENSE:FUT` and the Arabic perfective aspect verb feature is turned into `TENSE:PAST`.

Relation Labeling An Arabic subject may be: (a) pro-dropped (verb conjugated), (b) pre-verbal (full conjugation with verb), or (c) post-verbal (3person and gender agreement only). Third-person masculine/feminine singular verbs are often ambiguous as to whether they are case (a), where the adjacent noun is an object, or (c), where the adjacent noun is a subject. A verb can have zero, one or two objects. Pronominal objects are always cliticized to the verb, which means they can appear between the verb and the nominal subject. For passive verbs, the subject position is reserved for a *PRO* and the feature is passed along. In principle, Arabic's rich case system can account for the different configurations and also allow many variations in order, but since most cases are diacritical (and thus optionally written), that information is not always available. Arabic prose (non-poetry) writers generally avoid such syntactic acrobatics.

We use heuristics based on Arabic syntax to determine the relation of the verb to its nominal (common and proper), pronominal and relativizing children.

4.1.5 Subtree Phrase Construction

Each node in the dependency tree is annotated with the full projection of the subtree it heads. This subtree phrase is later used to interface with the statistical MT component.

4.2 Lexical Translation Issues

One of the main challenges in resource usage is the transformation of the lexicon of the Buckwalter Arabic morphological analyzer (BAMA) (Buckwalter, 2002) into a form that is readily usable by our GHMT system. The original Buckwalter lexicon contains English glosses for Arabic stem entries used in morphological analysis. Since the glosses are attached to stems, they are sometime inflected for number or voice. Table 1 illustrates the entries associated with three lemmas: `kuwfiy~_1`, `kAtib_1` and `katab-u_1`. Each entry consists of four tab-separated columns. The first two columns contain the undiacritized and diacritized stems, respectively. The third column specifies a morphological category which controls what affixes can be attached to the stem. Column four contains one or more English glosses. An optional fifth column marks the POS of the entry. The processing of this resource includes the following operations:

- POS determination. We determine the POS of the entry using the POS specified in the fifth column when present; otherwise, we use the form of the morphological category. For example, `PV` and `IV` indicate POS *verb*, whereas `N` and `Nap` indicate POS *noun*.

- Gloss slash expansion. The forward slash is used in the English gloss to specify alternatives, e.g., "of/from Kufa" for "of Kufa" or "from Kufa." We detect such cases and expand them appropriately.

- Parenthetical removal. Gloss parenthetical comments, such as "(Iraq)" in the entry for `kuwfiy~_1`, are removed from the gloss.

- Gloss depluralization. A plural gloss is discarded if the singular form of the gloss is used for a different stem of the same lemma. For example, the gloss "writers" for a plural stem of the lemma `kAtib_1` in Table 1 is removed since the singular form "writer" appears under a different stem of the same lemma.

- Gloss depassivization. Passive verb forms in English glosses are depassivized to ensure a lexemic translation. However, we made the decision to include both passive and active forms in the current version because of the high degree of ambiguity between these two forms in Arabic.

The following are the entries in our final lexicon which correspond to those in Table 1:

Table 1: Entries in the BAMA stem lexicon

```
;; kuwfiy~_1
kwfy   kuwfiy~   Nall        of/from Kufa (Iraq);Kufic        <pos>kuwfiy~/ADJ</pos>
;; kAtib_1
kAtb   kAtib     N/ap        writer;author
kAtb   kAtib     N/ap        clerk
ktAb   kut~Ab    N           authors;writers
ktb    katab     Nap         authors;writers
;; katab-u_1
ktb    katab     PV          write
ktb    kotub     IV          write
ktb    kutib     PV_Pass     be written;be fated;be destined
ktb    kotab     IV_Pass_yu  be written;be fated;be destined
```

```
kuwfiy~_1 AJ Kufic/from_Kufa/of_Kufa
kAtib_1   N  author/clerk/writer
katab-u_1 V  be_destined/be_fated/
             be_written/destine/fate/write
```

5 Integration of SMT Components into GHMT

The main challenge for integrating SMT components into GHMT is that the conception of the phrase (anything beyond a single word) is radically different. *Phrase-based* SMT systems take a phrase to be a sequence of words with no hidden underlying structure (Koehn, 2004). On the other hand, for systems that use parsers, like GHMT, a phrase has a linguistic structure that defines it and its behavior in a bigger context. Both kinds come with problems.

Statistical phrases are created from alignments, which may not be clean. This results in *jagged* edges to many phrases. For example, the phrase `. on the other hand , the` (containing seven words starting with a period and ending with "the") overlaps multiple linguistic phrase boundaries. Another related phenomenon is that of *statistical hallucination*, e.g., the translation of `AlswdAn w` (literally, *Sudan and*) into *enterprises and banks*.

Linguistic phrases come with a different set of problems. Since parsing technology for Arabic is still behind English,[3] many linguistic phrases are misparsed creating *symbolic hallucinations* that affect the rest of the system. A common example is the incorrect attachment of a prepositional phrase that modifies a complete sentence to one of its noun phrases.

We investigate two variants of a basic approach to using statistical phrases in the GHMT system. The phrase-based SMT system we use is Pharaoh (Koehn, 2004). We limit the statistical translation-table phrases used to those that correspond to completely projectable subtrees in the linguistic dependency representation of the input sentence. More complex solutions that use statistical phrases covering parts of a linguistic phrase are left for future work.

In the first variant, (GHMT + Phrase Table, henceforth GHMTPHT), we use the phrase table produced by Pharaoh as a multi-word surface dictionary. In the generation process, when a subtree is matched to an entry in this dictionary, an additional path in the generation lattice is created using the phrase-table entry in addition to the basic GHMT generation.

In the second variant, (GHMT + Pharaoh, henceforth GHMTPHARAOH), we use Pharaoh to translate the subtree projections for all the subtrees in the input sentence. These translations are added as alternatives to the basic GHMT system. Results comparing these two variants and a few others are described in Section 6.

The basic idea here is to exploit GHMT's focus on phrase structure generation (global level) together with a phrase-based SMT system's robustness (local phrases). One particular case in Arabic that we investigate later is the position of the subject relative to the verb. When we have a correct parse, moving the subject, which follows the verb in Arabic over 35% of the time, to a preverbal position is easy for GHMT (given a correct parse) but can be hard for a phrase-based SMT system, especially with sub-

[3]The parser we used in this paper is among the best available, yet its performance for Arabic is in the lower 70s percent (labeled constituency PARSEVAL F-1 score).

ject noun phrases exceeding the system's distortion limit.

6 Evaluation

We use the standard NIST MTEval datasets for the years 2003, 2004 and 2005 (henceforth MT03, MT04 and MT05, respectively).[4] The 2002 MTEval test set was used for Minimum Error Training (Och, 2003).

All of the training data used here are available from the Linguistic Data Consortium (LDC). We use an Arabic-English parallel corpus of about 5 million words to train the translation model.[5] For Arabic preprocessing, the Arabic Treebank scheme is used (Habash and Sadat, 2006). All systems use the same surface trigram language model, trained on approximately 340 million words of English newswire text from the English Gigaword corpus.[6]

English preprocessing simply included downcasing, separating punctuation from words and splitting off "'s". Trigram language models are implemented using the SRILM toolkit (Stolcke, 2002). Both BLEU (Papineni et al., 2002; Callison-Burch et al., 2006) and NIST (Doddington, 2002) metric scores are reported. All scores are computed against four references with n-grams of maximum length four. As a post-processing step, the translations of all systems are true-cased, and all results reported below refer to the case-sensitive BLEU and NIST scores.

We conducted three sets of evaluations that explore different aspects of the data sets and the system variants: a full system evaluation, a genre-specific evaluation, and a qualitative evaluation of specific linguistic phenomena.

6.1 Full Evaluation

Six system variants are compared:

- GIST is a simple gisting system that produces a sausage lattice from the English glosses in the output of the Buckwalter Arabic morphological analyzer (BAMA). Arabic word order is preserved and English realization is limited to the variants provided in BAMA.
- GHMT is the system described in Section 4. The lexical translation is limited to the Buckwalter lexicon.
- GHMTPHT is a variant of GHMT that uses a statistical phrase table as support multi-word surface dictionary (see Section 5).
- GHMTPHARAOH is the second variant discussed in Section 5. It uses Pharaoh to generate subtree phrases.
- PHARAOHBW is the Pharaoh phrase-based SMT system trained on the basic training set in addition to the entries in the Buckwalter lexicon.
- PHARAOH is the Pharaoh phrase-based SMT system trained only on the basic training set.

The results of the full systems are presented in Table 2. The lowest performing system is GIST, as expected. GHMT, using only the Buckwalter lexicon and no other training data, more than doubles the GIST score. This indicates that the system is making more correct lexical choices and word order realization beyond simple gisting.

GHMTPHT and GHMTPHARAOH provide substantial improvements over GHMT. In GHMTPHT, only 54.6% of subtrees find a match in the phrase table; as opposed to GHMTPHARAOH which guarantees a statistical translation for all subtrees. This accounts for the large difference between the two scores. This is a positive result for improving a non-statistical MT system with SMT components. However, the scores are still lower than the fully statistical system. We discuss the differences further in Section 6.3.

The primarily statistical systems PHARAOH and PHARAOHBW outperform all other systems. PHARAOH does better than PHARAOHBW for MT03 and MT05 but not for MT04. For all three data sets, the differences are not statistically significant.

As the amount of dependence on training data increases, we see a bigger divide between the different data sets. MT03 and MT05 behave similarly but

[4] http://www.nist.gov/speech/tests/mt/

[5] The parallel text includes Arabic News, eTIRR, English translation of Arabic Treebank, and Ummah.

[6] Distributed by the Linguistic Data Consortium: http://www.ldc.upenn.edu

Table 2: True-cased results of various systems on NIST MTEval test sets

Test Set	Metric	GIST	GHMT	GHMTPHT	GHMTPHARAOH	PHARAOHBW	PHARAOH
MT03	BLEU	0.0811	0.1479	0.2362	0.3379	0.4128	0.4162
	NIST	5.1846	6.0528	7.3213	8.2569	9.9205	9.9300
MT04	BLEU	0.0651	0.1402	0.2110	0.2777	0.3546	0.3522
	NIST	4.3904	6.0935	7.0981	7.5834	9.2038	9.1291
MT05	BLEU	0.0607	0.145	0.2313	0.3239	0.3935	0.3960
	NIST	4.7259	6.2636	7.4836	8.3687	9.6980	9.6615

Table 3: Genre-specific true-cased results of various systems on NIST MT04 test set

Genre	Metric	GIST	GHMT	GHMTPHT	GHMTPHARAOH	PHARAOHBW	PHARAOH
News	BLEU	0.0817	0.1617	0.2582	0.3434	0.4266	0.4244
	NIST	4.8989	6.358	7.6143	8.3132	9.7206	9.6796
Speech	BLEU	0.0429	0.1276	0.1821	0.2447	0.3088	0.3043
	NIST	3.2993	5.3923	6.2022	6.6354	7.8796	7.7164
Editorial	BLEU	0.0575	0.1144	0.1542	0.1914	0.2704	0.2703
	NIST	3.7633	4.9751	5.4724	5.4608	7.2344	7.1812

MT04 lags behind. One of the reason behind this behavior is that MT04 is a mixed genre data set. In the next section, we examine the differences in the genres in more detail.

6.2 Genre Evaluation

The MTEval 2004 data set is special in that it has a mix of genre (200 documents: 100 news, 50 speeches and 50 editorials). The training data we used is all Arabic news. We wanted to investigate the difference in behavior among variants with different types of symbolic and statistical resources. Table 3 presents the scores for genre-specific subsets of the MT04 test set.

The difference in scores across the different systems is consistent with the full evaluation in Table 2. The difference across the genre is very clear, with the news subset performing at a similar score level to that of the MT03 and MT05 test sets in Table 2.

Upon examination of the documents in MT04, we see several variations across the genres that explain the differences. In particular, speeches and editorials have a much higher rate of first and second person pronouns and verbs, include interrogative sentences, and use more flowery and fiery language than news. Out-of-vocabulary (OOV) rates in the the different subsets as measured against the basic training set data is as follows: news (2.02%), speeches (2.01%) and editorials (2.34%). The differences are very small. This confirms that it is style/use difference that is the biggest contributor to the difference in scores.

The fact that we see similar differences in GIST and GHMT as in PHARAOH contradicts our hypothesis that GHMT is more genre-independent than SMT approaches. We believe this is a result of the Arabic linguistic resources we use being biased towards news-genre. For example, the Arabic treebank used for training the parser is only in the news genre. The Buckwalter lexicon potentially also has some internal bias toward news genre because it was developed in tandem with the Arabic treebank.

6.3 Qualitative Evaluation

Automatic evaluation systems are often criticized for not capturing linguistic subtleties. This is clearly apparent in the field's moving back toward using human evaluation metrics such as HTER (Snover et al., 2006). We conducted a small human evaluation of verb and subject realization in eight random documents from MT04. The documents contained 47 sentences and reflect the distribution of genre in the MT04 test set. We compare three systems GHMT, GHMTPHARAOH and PHARAOH.

The evaluation was conducted using one bilingual Arabic-English speaker (native Arabic, almost native English). The task is to determine for every verb that appears in the Arabic input whether it is

Table 4: Verb and subject realization in eight documents from MT04

		GHMT		GHMTPHARAOH		PHARAOH	
Genre	**Verb Count**	**Verbs Seen**	**Realized Subject**	**Verbs Seen**	**Realized Subject**	**Verbs Seen**	**Realized Subject**
News	46	44 (95.7%)	29 (65.9%)	42 (91.3%)	31 (73.8%)	40 (87.0%)	29 (72.5%)
Speech	48	41 (85.4%)	21 (51.2%)	42 (87.5%)	24 (57.1%)	29 (60.4%)	12 (41.4%)
Editorial	29	20 (69.0%)	8 (40.0%)	17 (58.6%)	9 (52.9%)	19 (65.5%)	10 (52.6%)
All	123	105 (85.4%)	58 (55.2%)	101 (82.1%)	64 (63.4%)	88 (71.5%)	51 (58.0%)

realized or not in the English translation. If realized, we then determine whether its subject is mapped to the appropriate position in English. Since translation divergences can cause an Arabic verb to appear as a noun in English, a nominalized translation is accepted as a valid realization. The subject of a non-verbal translation is considered correctly assigned if the meaning of the relationship of the original subject-verb pair is preserved. Correct realization of the verb object is not considered here, and neither are non-verbal Arabic translations to verb forms in English.

The results are presented in Table 4 for each genre and also collectively. For each of the three systems studied, two columns are presented. The first presents the count of verbs and their percentage of all Arabic verbs (from the column Verb Count). The second column presents the number of correctly realized subjects and their percentage relative to the *seen verbs*.

Both the percentage of verbs seen and realized subjects show a drop as we go from news genre to speeches and editorials. This is consistent with the automatic evaluation scores. The percentage of verbs seen is much higher in GHMT compared to PHARAOH. This is consistent with previous findings comparing GHMT to SMT systems (Habash, 2003b). The relative percentage of realized subjects is lower mostly due to chunking and parsing errors on the Arabic input. A positive result is the performance of the GHMTPHARAOH system which although slightly below GHMT in terms of verbs seen, has a higher percentage of realized subjects. In fact, it is the highest among the three systems. We believe this is a result of statistical phrase robustness which is independent of the parse correctness. So, for example, even if the verb and its subject are misparsed as a compounding of two nouns (POS tag error and parse error), the SMT translation of their projected subtree produces the right verb-subject pair in the correct relative order. Clearly, further research is needed to investigate similar phenomena so the respective strengths of both approaches can be further exploited.

7 Conclusions and Future Work

We presented the challenges and details of an implementation of an Arabic-English GHMT system extended with SMT components. We described an extensive evaluation of multiple system variants and reported positive results on the advantages of hybridization. Manual evaluation of verb-subject realization showed that our symbolic approach (GHMT) and our hybrid approach (GHMTPHARAOH) outperform a purely statistical approach, with the hybrid approach yielding the best performance.

In the future, we plan to extend the use of statistical phrases in the GHMT system to parts of the linguistic tree. We also plan to further investigate how statistical phrases can be used in making symbolic components more robust for MT purposes. The evaluation we did uncovered a wealth of research problems that will serve as the basis of future research. We believe many of the insights of this work are applicable to other languages, particularly those with rich morphologies.

Acknowledgments

This work has been supported, in part, under Army Research Lab Cooperative Agreement DAAD190320020 and the GALE program of the Defense Advanced Research Projects Agency, Contracts No. HR0011-06-2-0001 and HR0011-06-C-0023. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of ARL or DARPA.

References

Daniel M. Bikel. 2002. Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine. In *Proceedings of the International Conference on Human Language Technology Research (HLT)*, San Diego, CA.

Tim Buckwalter. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania. Catalog No.: LDC2002L49.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL'06)*, Trento, Italy.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause Restructuring for Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, MI.

Michael Collins. 1997. Three Generative, Lexicalised Models for Statistical Parsing. In *Proceedings of the ACL (jointly with the EACL)*, Madrid, Spain.

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. 2004. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In *Proceedings of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04)*, Boston, MA.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. In *Proceeding of the ARPA Workshop on Human Language Technology*.

Bonnie J. Dorr. 1993. *Machine Translation: A View from the Lexicon*. The MIT Press, Cambridge, MA.

Sharon Goldwater and David McClosky. 2005. Improving Statistical MT through Morphological Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver, Canada.

David Graff. 2003. Arabic Gigaword. Linguistic Data Consortium, University of Pennsylvania. Catalog No.: LDC2003T12.

Nizar Habash and Bonnie J. Dorr. 2002. Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation. In *Proceedings of the Association for Machine Translation in the Americas, AMTA-2002*, Tiburon, CA.

Nizar Habash and Owen Rambow. 2004. Extracting a Tree Adjoining Grammar from the Penn Arabic Treebank. In *Proceedings of Traitement Automatique du Langage Naturel (TALN-04)*. Fez, Morocco.

Nizar Habash and Owen Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of (ACL'05)*.

Nizar Habash and Fatiha Sadat. 2006. Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL06)*, New York, NY.

Nizar Habash. 2003a. *Generation Heavy Hybrid Machine Translation*. Ph.D. thesis, University of Maryland College Park.

Nizar Habash. 2003b. Matador: A Large Scale Spanish-English GHMT System. In *Proceedings of Machine Translation Summit (MT Summit IX)*, New Orleans, LA.

Nizar Habash. 2006. Arabic Morphological Representations for Machine Translation. In *Arabic Computational Morphology: Knowledge-based and Empirical Methods. Editors A. van den Bosch and A. Soudi.*

Philipp Koehn. 2004. Pharaoh: a Beam Search Decoder for Phrase-based Statistical Machine Translation Models. In *Proceedings of the Association for Machine Translation in the Americas (AMTA-2004)*, Washington DC.

Young-Suk Lee. 2004. Morphological Analysis for Statistical Machine Translation. In *Proceedings of (HLT-NAACL04)*.

Mohamed Maamouri, Ann Bies, and Tim Buckwalter. 2004. The Penn Arabic Treebank: Building a Large-scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Franz Josef Och, Daniel Gildea, Anoop Sarkar, Kenji Yamada, Sanjeev Khudanpur, Alex Fraser, Shankar Kumar, David Smith, Libin Shen, Viren Jain, Katherine Eng, Zhen Jin, and Dragomir Radev. 2004. A Smorgasbord of Features for Statistical Machine Translation. In *Proceedings of (HLT-NAACL04)*.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of (ACL'03)*, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of (ACL'02)*, Philadelphia, PA.

Maja Popović and Hermann Ney. 2004. Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proceedings of (ACL'05)*.

Sonja Nießen and Hermann Ney. 2004. Statistical Machine Translation with Scarce Resources using Morpho-syntactic Information. *Computational Linguistics*, 30(2).

Jean Senellart. 2006. Boosting Linguistic Rule-based MT System with Corpus-based Approaches. In *Presentation. GALE PI Meeting, Boston, MA.*

Matt Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the Association for Machine Translation in the Americas (AMTA 2006)*, Cambridge, MA.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Denver, CO.