Mach Translat
DOI 10.1007/s10590-009-9056-7

Symbolic-to-statistical hybridization: extending generation-heavy machine translation

Nizar Habash · Bonnie Dorr · Christof Monz

Received: 28 January 2009 / Accepted: 5 October 2009

Abstract The last few years have witnessed an increasing interest in hybridizing surface-based statistical approaches and rule-based symbolic approaches to machine translation (MT). Much of that work is focused on extending statistical MT systems with symbolic knowledge and components. In the brand of hybridization discussed here, we go in the opposite direction: adding statistical bilingual components to a symbolic system. Our base system is Generation-heavy machine translation (GHMT), a primarily symbolic asymmetrical approach that addresses the issue of *Interlingual* MT resource poverty in source-poor/target-rich language pairs by exploiting symbolic and statistical target-language resources. GHMT's statistical components are limited to target-language models, which arguably makes it a simple form of a *hybrid system*. We extend the hybrid nature of GHMT by adding statistical bilingual components. We also describe the details of retargeting it to Arabic–English MT. The morphological richness of Arabic brings several challenges to the hybridization task. We conduct an extensive evaluation of multiple system variants. Our evaluation shows that this new variant of GHMT—a primarily symbolic system extended with monolingual and bilingual statistical components—has a higher degree of grammaticality than a phrase-based statistical MT system, where grammaticality is measured in terms of correct verb-argument realization and long-distance dependency translation.

N. Habash (✉)
Center for Computational Learning Systems, Columbia University, New York, NY, USA
e-mail: habash@ccls.columbia.edu

B. Dorr
Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA
e-mail: bonnie@umiacs.umd.edu

C. Monz
Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
e-mail: c.monz@uva.nl

Published online: 11 November 2009

Keywords Hybrid machine translation · Generation-heavy machine translation · Statistical machine translation · Arabic–English machine translation

1 Introduction

The various approaches to machine translation (MT) can be grouped into two camps: the symbolic/rule-based and the statistical/corpus-based approaches.

Symbolic MT approaches are further classified according to the MT pyramid (Vauquois 1968) into word-based, transfer and interlingual approaches. These three classes vary in the linguistic depth of the representations they use. Word-based approaches (also called 'gisting') translate the words without using any deeper representation other than perhaps word morphology. Transfer approaches build syntactic analyses of the source language (SL), which are then *transferred* to syntactic representations of the target language (TL) using a transfer lexicon (dictionary) which relates syntactic structures and words in SL to TL. The TL syntactic representation is then realized (generated). Interlingual approaches use a deeper representation, an interlingua (IL), that abstracts away out of the SL and its syntax into a semantic language-independent representation. Interlingual approaches require components for analysis and generation that can translate into and from IL, respectively, for the SL and TL. To accomplish this, two SL–IL and TL–IL lexicons (dictionaries) with symmetrical coverage are needed. Interlingual approaches typically use syntactic parsing and generation components comparable to those used in transfer systems. The deeper the symbolic approach, the more expensive the resources it requires. Resource richness and resource symmetry (for SL and TL) are some of the main challenges for symbolic MT approaches. *Resource poverty* often refers to the lack of monolingual and even multilingual resources involving the "poor" language.

In contrast to these approaches, statistical/corpus-based approaches have a much simpler symmetry requirement: the presence of parallel corpora for SL and TL. For these approaches, the degree of resource poverty is defined primarily in terms of the size of monolingual and bilingual corpora for the language pair in question. These surface-based statistical MT approaches translate between languages purely on the basis of word-to-word or phrase-to-phrase translation tables that do not involve any deeper syntactic/semantic analysis. In the last decade and a half, the success of surface-based statistical MT approaches has changed the face of the field, which was previously dominated by linguistic rule-based symbolic approaches.

In their simplest and purest form, statistical MT approaches are surface-based and linguistic-rule-free and, correspondingly, symbolic MT systems are statistics-free. The last few years have witnessed an increased interest in hybridizing the two approaches to create systems that exploit the advantages of both linguistic rules and statistical techniques in MT systems. Much of that work is focused on extending statistical MT systems with symbolic knowledge and components. In the brand of hybridization discussed here, we go in the opposite direction: adding statistical knowledge and components to a symbolic system. This approach to hybridization is of particular interest to researchers and companies who have developed robust symbolic systems over many years and want to extend their systems with statistical components.

As a starting point, we build on a primarily symbolic Generation-heavy machine translation (GHMT) system (Habash and Dorr 2002; Habash 2003a; Habash et al. 2003). This is an asymmetrical approach that addresses the issue of *Interlingual* MT resource poverty in source-poor/target-rich language pairs by exploiting symbolic and statistical TL resources. The source-target asymmetry of systems developed in this approach makes them more easily retargetable to new SLs provided that some necessary resources exist. Expected resources involving the SL include a syntactic parser and a simple bilingual SL–TL one-to-many translation dictionary without translation preferences or probabilities. Rich TL symbolic resources such as word lexical semantics and categorial variations are used to overgenerate multiple structural variations from a TL-glossed syntactic representation of SL sentences. This symbolic overgeneration accounts for possible translation divergences, cases where the underlying concept or "gist" of a sentence is realized differently in two languages such as *to put butter* and *to butter* (Dorr 1993b). Overgeneration is constrained by multiple statistical TL models including surface *n*-grams and structural *n*-grams. The statistical components in GHMT are limited to TL monolingual resources, which arguably makes it a simple form of a *hybrid system*.

In this article, we extend the hybrid nature of GHMT with bilingual (SL–TL) statistical translation resources by enriching its translation dictionary with phrase tables from a phrase-based statistical MT (PBSMT) system and by using a PBSMT system as a sub-component of GHMT. We also describe the details of retargeting GHMT to Arabic–English MT. The morphological richness of Arabic brings several challenges to the hybridization task. We conduct an extensive evaluation of multiple variant systems. Our evaluation shows that this new variant of GHMT—a primarily symbolic system extended with monolingual and bilingual statistical components—improves measurably over a basic GHMT system (with only monolingual TL statistical components). In addition, although it does not outperform the PBSMT system in terms of standard automatic evaluation measures, the hybrid GHMT system has a higher degree of grammaticality than the PBSMT system according to a qualitative analysis, where grammaticality is measured in terms of correct verb-argument realization. The hybrid GHMT system also outperforms the PBSMT system on an automatic evaluation isolating the performance on long-distance dependencies.

The remainder of this article is organized as follows. In Sect. 2, we discuss related work on MT hybridization. In Sect. 3, we describe GHMT in more detail. In Sect. 4, we discuss the issues involved in building the Arabic specific components of Arabic–English GHMT. Sect. 5 identifies the issues involved in extending GHMT with bilingual statistical components and resources. In Sect. 6, we present an extensive evaluation of several variant MT systems. In Sect. 7 we conclude, and offer some avenues for further work.

2 Related work

During the last few years, much interest has developed in the area of hybrid MT by researchers working on both symbolic and statistical approaches to MT. In the case of statistical approaches, hybridization has led to the incorporation of symbolic knowledge such as morphological information and syntactic structure. Our own research differs in that we are approaching the hybridization question from the opposite direction, i.e. how to incorporate statistical MT components into symbolic rule-based systems (Senellart 2006). Nonetheless, the research on statistical MT hybrids has influenced many of our decisions and directions. In the next two subsections, we primarily review this body of research. Other relevant related work is discussed in context in later sections.

2.1 Morphology-based approaches

Morphologically rich languages, such as Arabic or Czech, tend to have very large vocabularies resulting from the presence of large numbers of morphologically inflected forms, which leads to sparse computational models. The anecdotal intuition in the field is that reduction of data 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 statistical MT quality focused on morphologically rich languages such as German (Nießen and Ney 2004), Spanish, Catalan, and Serbian (Popović and Ney 2004), Czech (Goldwater and McClosky 2005), and languages with very limited linguistic resources such as Vietnamese (Nguyen and Shimazu 2006). These studies examined the effects of various kinds of tokenization, lemmatization and part-of-speech (POS) tagging and showed a positive effect on statistical MT 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) and Sadat and Habash (2006) reached similar conclusions on a much larger set of experiments including multiple preprocessing schemes reflecting different levels of morphological representation and multiple techniques for disambiguation and tokenization. They showed that specific preprocessing decisions can have a positive effect when decoding text with a different genre than that of the training data (in essence another form of data sparsity). They also demonstrated gains in MT quality through a combination of different preprocessing schemes. Additional similar results were reported using specific preprocessing schemes and techniques (Och 2005; El Isbihani et al. 2006; Riesa and Yarowsky 2006; Zollmann et al. 2006).

Within our approach, working with Arabic morphology is especially challenging for reasons discussed in more detail in Sect. 4.

2.2 Syntax-based approaches

More recently a number of statistical MT approaches have 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. Transformations were applied to the SL parse trees to make the order of the SL words and phrases closer to that of TL. The same reordering was done for a new SL sentence before decoding. A modest statistically significant improvement was demonstrated over basic PBSMT.

Xia and McCord (2004) described an approach for translation from French to English, where reordering rules were acquired automatically using SL and TL parses and word alignment. The reordering rules—context-free constituency representations—were mostly lexicalized, with marked heads. Source and target parses were used to constrain the word alignments that serve as input to rule extraction. This work showed that there is a positive effect to reordering. Zhang et al. (2007) described a similar approach that learned reordering rules for Chinese–English statistical MT using chunking (shallow parsing). Unlexicalized context-free syntactic chunk tags and POS tags were used on the SL side only. All possible learned reorderings were used to create a lattice that is taken as input to the decoder; the reordering rules were not applied to training data (unlike Xia and McCord 2004). Other work that has incorporated syntax into preprocessing for statistical MT includes that of Crego and Mariño (2007), a similar approach that combined reordering and decoding, and that of Habash (2007b), a method for learning reordering rules for Arabic–English statistical MT.

Quirk et al. (2005) used sub-graphs of dependency trees to deal with word-order differences between SL and TL. During training, dependency graphs on the SL side were projected onto the TL side using the alignment links between words in the two languages. During decoding, the different sub-graphs were combined in order to generate the most likely dependency tree.

Similar to Collins et al. (2005), Quirk et al. (2005) and Habash (2007b), our approach uses SL syntactic (specifically dependency) representations to capture generalizations about the SL text. Unlike them, we do not use or learn specific mappings between the syntactic structure of SL and TL. Instead, our approach maps SL to a syntactically language-independent representation that forms the basis for TL generation. Additionally, we extend our approach by adding statistically learned components to its symbolic core. In following this approach, we are in not in the field's mainstream that tends to extend statistical systems with symbolic components. There has been some recent research that attempts similar extensions to transfer-based systems. Font-Llitjós and Vogel (2007) used parallel data to automatically learn rule refinements. Dugast et al. (2009) selectively added dictionary entries learned from parallel data to their rule-based MT system. The work presented here is closer to Dugast et al. (2009).

Finally, a closely-related NLP area of research that models syntax and has inspired our approach is that of hybrid natural language generation (HNLG).[1]

Two particular HNLG systems are most relevant to the work presented here: Nitrogen/Halogen (Langkilde and Knight 1998a,b; Langkilde 2000) and FERGUS (Bangalore and Rambow 2000a). Nitrogen is a hybrid NLG system that uses n-gram LMs to rank paths in a symbolically overgenerated lattice of possible outputs. A later version of Nitrogen, Halogen, improves on time-space efficiency by compressing the search space into *forests*, which are compact non-redundant syntactically-derived

[1] The term "hybrid" is used here to refer to the use of statistical language models (LM) combined with linguistic generation rules. Earlier publications on GHMT used the word "hybrid" in a similar sense. Since then, the use of LMs has become common enough in symbolic systems that it hardly counts as a hybrid to use one.

representations of lattices (Langkilde 2000).[2] Although structural syntactic information is used in constructing forests, the only LM used in Halogen is a surface n-gram LM. FERGUS (Flexible Empiricist/Rationalist Generation Using Syntax) extends the use of n-gram LMs with a tree-based statistical model, structural n-gram (Sn-gram) model and a lexicalized tree-based syntactic grammar (Bangalore and Rambow 2000a). The use of Sn-grams for lexical selection was tested through an artificial expansion of words using WordNet supersynsets (Fellbaum 1998). The experiment showed that lexical choice was improved using structural LMs.

In terms of input complexity and the balance of symbolic and statistical components, our generation component, EXERGE (EXpansivE Rich Generation for English, Sect. 3.3), falls between the hybrid NLG systems Nitrogen and FERGUS. FERGUS requires the shallowest input (closest to the TL surface form) and employs the most statistical and symbolic power. Nitrogen's input is the deepest (semantic representation) and its resources the simplest (an overgenerating grammar and n-gram LM). EXERGE uses several symbolic and statistical components for generation. The most notable difference is the use of lexical semantic resources for structural expansion, a process unparalleled in FERGUS or Nitrogen/Halogen. EXERGE also uses Sn-grams for both lexical and structural selection. For surface realization, EXERGE implements a rule-based grammar in a linearization engine oxyGen (Habash 2000) that creates surface lattices, ranked using an n-gram LM.

3 Generation-heavy machine translation

GHMT is primarily a symbolic MT approach that comes out of the tradition of interlingual and transfer MT. However, it is itself neither interlingual nor transfer. GHMT diverges in that it has no interlingual or transfer lexicons for translating from SL to TL. In fact, as far as the basic lexical transfer component in GHMT, it is closer to word-based translation, except that this lexical translation is done on transfer-like tree representations, which are then extended using a target-side interlingual lexicon to provide alternative realizations.

GHMT extends the HNLG approach to handle translation divergences *without* the use of a deeper semantic representation or transfer rules. This is accomplished through the inclusion of structural and categorial expansion of SL syntactic dependencies in the symbolic overgeneration component. The overgeneration is constrained by linguistically motivated rules that utilize TL lexical semantic knowledge and subcategorization frames and is independent of SL preferences.

In this section, we begin with a discussion of translation divergences, which motivate the research and design behind GHMT. We then present an overview and a representative end-to-end Spanish–English example to explain the intuition behind the approach. Next we discuss the generation component in GHMT in more detail starting with the different generation resources and processes. Finally, we present a summary of results from a previously published Spanish–English GHMT system (Habash 2003a).

[2] Note that *forests* here should not be confused with *forests* often described in the parsing literature, where they refer to a compact representation of different parse structures.

Table 1 Translation divergence types

Divergence	Spanish/Arabic	English gloss	English translation	%
Categorial	X *tener hambre*	(X *have hunger*)	X *be hungry*	98
	ArAd X *AlðhAb*	(*wanted* X *the-going*)	X *wanted to go*	
Conflational	X *dar puñaladas a* Z	(X *give stabs to* Z)	X *stab* Z	83
	Alqý X *AlqbD ςlý* Y	(*placed* X *the-arrest on* Y)	X *arrested* Y	
Structural	X *entrar en* Y	(X *enter in* Y)	X *enter* Y	35
	Aςrb X *ςn* Y	(*expressed* X *about* Y)	X *expressed* Y	
Head	X *cruzar* Y *nadando*	(X *cross* Y *swimming*)	X *swim across* Y	8
Swapping	*Asrς* X *AlsbAHħ*	(*speed* X *the-swimming*)	X *swam fast*	
Thematic	X *gustar a* Y	(X *appeal to* Y)	Y *like* X	6
	Aςjb X Y	(*appeal-to* X Y)	Y *like* X	

We describe the details of the Arabic-specific component in the system we evaluate in this article in Sect. 4.

3.1 Translation divergences

A translation divergence occurs when the underlying concept or “gist” of a sentence is distributed over different words for different languages, which results in a lack of parallelism in the correspondence between SL and TL structures and contents. For example, the notion of *floating across a river* is expressed as *float across a river* in English and *cross a river floating* (*atravesó el río flotando*) in Spanish (Dorr 1993b). An investigation done by Dorr et al. (2002) found that divergences relative to English occurred in approximately 1 out of every 3 sentences in the Spanish TREC El Norte Newspaper corpus (Rogers 2000).[3] In what follows, we describe translation divergence types before turning to alternative approaches to handling them.

3.1.1 Translation divergence types

While there are many ways to classify divergences, we present them here in terms of five specific divergence *types* that can take place alone or in combination with other types of translation divergences. Table 1 presents these divergence archetypes with Spanish–English and Arabic–English examples.[4]

The divergence categories are described in more detail by Dorr et al. (2002), who report that 10.5% of Spanish sentences and 12.4% of Arabic sentences have at least one translation divergence.

[3] The TREC El Norte corpus consists of approximately 500 MB of Spanish newswire data from the Mexican El Norte newspaper and the Agence France Presse newswire for the period 1994–1996.

[4] All Arabic transliterations are provided in the Habash–Soudi–Buckwalter transliteration scheme (Habash et al. 2007). See the addendum at the end of the article.

Categorial divergence involves a translation that uses different parts of speech. *Conflation* involves the translation of two words using a single word that combines their meaning. This divergence type usually involves a single English verb being translated using a combination of a light verb,[5] and some other meaning-bearing unit such as a noun or a progressive manner verb. *Structural divergence* involves the realization of incorporated arguments such as subject and object as obliques (i.e. headed by a preposition in a prepositional phrase) or vice versa. *Head swapping* (or 'head switching') involves the demotion of the head verb and the promotion of one of its modifiers to head position. In other words, a permutation of semantically equivalent words is necessary to go from one language to the other. In Spanish, this divergence is typical in the translation of an English motion verb and a preposition as a directed motion verb and a progressive verb. Finally, *thematic divergence* occurs when the verb's arguments switch syntactic argument roles from one language to another (i.e. subject becomes object and object becomes subject).

The last column in Table 1 displays a percentage of occurrences of the specific divergence type, taken from the first 48 unique instances of Spanish–English divergences from the TREC El Norte corpus. Note that there is often overlap among the divergence types—the reason the percentages do not add up to 100—with the categorial divergence occurring almost every time there is any other type of divergence.[6]

This highlights the need for a systematic approach to handling divergences that addresses all their different types and the interactions between them rather than addressing specific cases one at a time.

3.1.2 Handling translation divergences

Since translation divergences require a combination of lexical and structural manipulations, they are most typically handled in symbolic MT systems through the use of transfer rules (Han et al. 2000; Lavoie et al. 2000).[7] A pure transfer approach is a brute force attempt to manually encode all translation divergences in a transfer lexicon (Dorr et al. 1999). Very large parsed and word-aligned bilingual corpora have also been used to automatically extract transfer rules (Watanabe et al. 2000; Lavoie et al. 2001).

Alternatively, more linguistically-sophisticated techniques that use lexical semantic knowledge to detect and handle divergences have been developed. One approach uses Jackendoff's Lexical Conceptual Structure (LCS) (Jackendoff 1983, 1990) as an IL (Dorr 1993b). LCS is a compositional abstraction with language-independent properties that transcend structural idiosyncrasies by providing a granularity of representation

[5] Semantically *light verbs* carry little meaning in their own right and cannot function as predicates without additional words (Grimshaw and Mester 1988). Common English light verbs include *give*, *do* or *have*.

[6] As such, categorial divergences are central to the problem of divergence detection. That is, approaches that focus on categorial divergence as a means of detecting *any* divergence are more likely to be successful than those that focus on other divergence type. That categorial variations are central to divergence detection is what motivated the construction of a large database (CatVar) of over 28K categorial variations (Habash and Dorr 2003). CatVar is discussed further in Sect. 3.3.1.2.

[7] No special treatment is given to translation divergences in basic statistical MT approaches although there have been some efforts on improving word alignment using linguistic knowledge about translation divergences (Dorr et al. 2002; Ayan et al. 2004).

much finer than syntactic representation. LCS has been used in several projects such as UNITRAN (Dorr 1993a) and ChinMT (Traum and Habash 2000; Habash et al. 2003). Important ideas and resources in LCS are used in our GHMT approach. Another approach enriches lexico-structural transfer at Mel'čuk's Deep Syntactic Structure (DSyntS) level (Mel'čuk 1988) with cross-linguistic lexical semantic features (Nasr et al. 1997). A major limitation of these interlingual and transfer approaches (whether using lexical semantics or corpus-based) is that they require a large amount of explicit symmetric knowledge for both SL and TL.

Our approach, GHMT, is closely related to HNLG (Langkilde and Knight 1998a; Bangalore and Rambow 2000a; Langkilde 2000) mentioned in Sect. 2. The idea is to combine symbolic and statistical knowledge in generation through a two-step process: (1) *symbolic overgeneration* followed by (2) *statistical extraction*. This approach has been used previously for generation from semantic representations (Langkilde and Knight 1998a) or from shallow unlabeled dependencies (Bangalore and Rambow 2000b). GHMT extends this earlier work by including structural and categorial expansion of SL syntactic dependencies as part of the symbolic overgeneration component to produce correct TL structures and contents, thus allowing us to handle the problem of translation divergence.[8]

The fact that GHMT does not require semantically analyzed SL representations or structural transfer lexicons makes it well-suited for handling translation divergences with relatively minimal lexical resources for the SL. The overgeneration is constrained by linguistically-motivated rules that utilize TL lexical semantics and is independent of the SL preferences. The generated lexico-structural combinations are then ranked by a statistical extraction/ranking component. Through this use of its resources, GHMT approximates interlingual behavior without fully utilizing an IL (Dorr and Habash 2002).[9]

3.2 GHMT: an overview of the framework

Figure 1 describes the different components of GHMT. There are three phases: *analysis*, *translation* and *generation*. The last phase is marked as EXERGE—EXpansivE Rich Generation for English—a (standalone) SL-independent generation module for English. These three phases are very similar to the *analysis-transfer-generation* or *analysis-IL-generation* paradigms of MT (Dorr et al. 1999). However, the GHMT phases are not symmetric. Analysis consists only of SL sentence parsing and is independent of English generation.

The output of analysis is a deep syntactic dependency that normalizes over syntactic phenomena such as passivization and morphological expressions of tense, number,

[8] Another corpus-based MT approach that relies heavily on TL language modeling is context-based MT (CBMT) (Carbonell et al. 2006), which does not require the presence of parallel corpora, although it needs a very large surface-based dictionary.

[9] Another non-interlingual non-transfer approach is Shake-and-Bake MT (Beaven 1992; Whitelock 1992) which overgenerates TL sentences and symbolically constrains the output by "parsing it." It differs from GHMT in two ways: (1) Shake-and-Bake is a purely symbolic approach and (2) it requires symmetric resources for SL and TL.

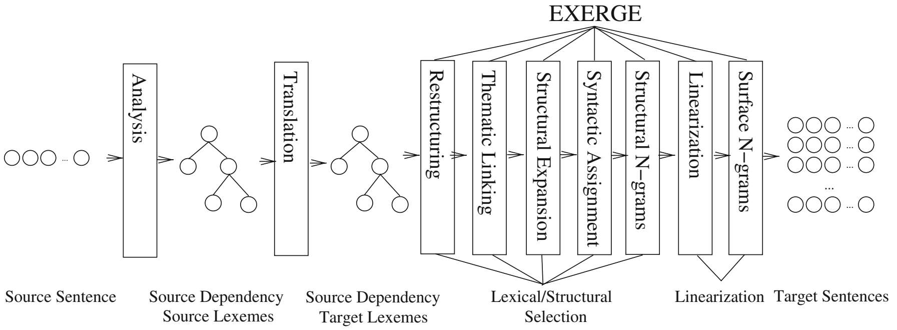

Fig. 1 Generation-heavy machine translation

etc. The dependency tree is labeled with minimal syntactic relations such as *subject* (subj), *object* (obj), *indirect object* (obj2), and *modifier* (mod). Moreover, the nodes must have lexemes and features from a pre-specified set of feature names and values (Habash 2003a). Translation converts the SL lexemes into bags of English lexemes. The dependency structure of the SL is maintained. The generation phase (EXERGE in Fig. 1) is where most of the work is done to manipulate the input lexically and structurally and produce English sequences.

The GHMT analysis step is by most considerations quite SL-heavy and requires a rich resource (a parser) to be available. This requirement in GHMT is a relative weakness but it is generally commonly expected in symbolic approaches. The source-poverty aspect that GHMT addresses is that of the lexical translation component, which is much simpler than complex transfer and interlingual lexicons.

We will use a Spanish representative example although our focus later is on the translation from Arabic. The Spanish sentence *Maria puso la mantequilla en el pan (Maria put the butter on the bread)* is analyzed to produce a dependency tree, a representation describing the syntactic relations among the words in the sentence:

(1) (puso :subj Maria :obj (mantequilla :mod la) :mod (en :obj (pan :mod el)))

This dependency tree specifies that *Maria* is the subject of the verb *puso* and that *mantequilla* is the object. In the translation step, each of the Spanish words in the dependency tree are mapped into sets of English words:

(2) (lay/locate/place/put/render/set/stand :subj Maria :obj (butter :mod the) :mod (on/in/into/at :obj (bread/loaf :mod the)))

During generation, different variants of (2) are expansively created using lexical semantic information and other English-specific heavy resources. The following are only a few of these variants:

(3) (put :subj Maria :obj (butter :mod the) :mod (on :obj ((bread loaf) :mod the)))
(lay :subj Maria :obj (butter :mod the) :mod (at :obj ((bread loaf) :mod the)))
(butter :subj Maria :obj ((bread loaf) :mod the))
(bread :subj Maria :obj (butter :mod the))
. . .

The first two examples show little difference in structure from the Spanish structure in (2), but the last two are very different. In the linearization step, the dependency trees in (3) are converted into a word lattice compressing multiple possible sentences:

(4) (OR
  (SEQ Maria put the butter (OR on into) (OR bread loaf))
  (SEQ Maria laid the butter (OR at into) (OR bread loaf))
  (SEQ Maria buttered the (OR bread loaf))
  (SEQ Maria breaded the butter) . . .)

These different sequences are then ranked using a statistical LM. The overgenerated variants score higher than direct word translations, e.g. the top-ranked output in this example is *Maria buttered the bread*. The associated scores are the logarithms of the probabilities from the LM.

(5) −47.0841 Maria buttered the bread
−48.7334 Maria breaded the butter
−51.3784 Maria buttered the loaf
−54.1280 Maria put the butter on bread

3.3 The generation component: EXERGE

The generation component consists of seven steps (marked EXERGE in Fig. 1). The first five are responsible for lexical and structural selection and the last two are responsible for realization. Initially, the SL syntactic dependency (now with TL words) is converted into a thematic dependency (restructuring and thematic linking steps). This is followed by structural expansion, where semantics are modeled through an exploration of structural variations within the thematic dependency (discussed in detail below—Sect. 3.3.2.3). The fourth step maps the thematic dependency to a target syntactic dependency (semantic-syntax interface). The fifth step filters the resulting forest of syntactic dependencies using a structural lexeme-based *n*-gram LM. The resulting trees are then linearized. The linearization step models syntactic ordering and morphological form generation. The surface lattice is then passed through a surface *n*-gram LM to produce a ranked *n*-best list of outputs.

The next section describes the generation resources followed by a detailed explanation of the generation sub-modules. The specific implementation of the generation component discussed here is focused on English as an output language. Extending this approach to some other TL requires providing the rich linguistic resources described next for that language.

3.3.1 Generation resources

In addition to using off-the-shelf resources, such *n*-gram LMs (Stolcke 2002) and oxyGen's English generation grammar (Habash 2000), the generation component utilizes four important resources created with GHMT in mind: a word-class lexicon, a categorial-variations lexicon, a syntactic-thematic linking map, and a structural *n*-gram LM. We also use a surface *n*-gram LM. Each of these resources is discussed below.

3.3.1.1. Word-class lexicon The word-class lexicon is a reduced version of the English LCS Verb and Preposition Database (henceforth, LCS-DB) (Dorr 2001), which links English verbs and prepositions with their subcategorization frames, LCS representations and thematic role grids. The LCS-DB is organized around Levin-style classes that distinguish among different word senses (Levin 1993) and contains example sentences and verb lists for each class. In the case of verbs, there are 511 verb classes for 3,131 verbs, totaling 8,650 entries. An example is shown here:

```
(6)  (DEFINE-WCLASS
     :NUMBER ``V.13.1.a.ii''
     :NAME ``Give - No Exchange''
     :SENTENCES (``He !!+ed the car to John'' ``He !!+ed
      John the car'')
     :POS V
     :THETA_ROLES (``ag/obl+th/obl+goal/obl/to''
      ``ag/obl+goal/obl+th/obl'')
     :LCS_PRIMS (cause go possessional)
     :WORDS (feed give pass pay peddle refund render
      repay serve))
```

In this example, the NUMBER and NAME entries refer to the Levin verb class number and name. The SENTENCE, POS and WORDS entries specify example sentences, the part-of-speech, and member words in the class, respectively. Whereas the full LCS representation (structure and content) is used in the LCS-DB, the word class lexicon only contains the LCS main primitives (LCS_PRIMS entry) for each structure and the associated thematic roles (THETA_ROLES entry). For this specific example of *verbs of giving with no exchange*, there are two theta role grids: the first, *ag/obl* + *th/obl* + *goal/obl/to*, specifies that verbs in this class require, in the following surface order, an obligatory (obl) agent (ag), an obligatory theme (th) and an obligatory goal generated with the preposition *to*; the second theta role grid is a variant with a different surface order (agent-goal-theme). The LCS primitives indicate that the verbs in this class involve a causative (cause) movement (go) in the possessional field, i.e. they involve *an agent causing a theme to go to a goal possessionally*.

In the case of prepositions, there are 43 preposition classes, for 125 prepositions, totaling 444 entries. An example with the class of temporal prepositions is shown here:

```
(7)  (DEFINE-WCLASS
     :NUMBER ``P.8''
     :NAME ``Preposition Class P.8''
     :POS P
     :THETA_ROLES (time)
     :LCS_PRIMS (path temporal)
     :WORDS (until to till from before at after))
```

Note that these entries are only required for the TL (English in our case). There are no equivalent entries for any SL.

3.3.1.2. Categorial-variation database (CatVar) This is a database of uninflected words (lexemes) and their categorial variants.[10] The database was developed using a combination of resources and algorithms including LCS-DB (Dorr 2001), the Brown Corpus section of the Penn Treebank (Marcus et al. 1994), an English morphological analysis lexicon developed for PC-Kimmo (ENGLEX) (Antworth 1990), Nomlex[11] (Macleod et al. 1998) and the Porter stemmer (Porter 1980). The database contains 28,305 clusters for 46,037 words. The following is an excerpt:

```
(8)  (:V (hunger) :N (hunger) :AJ (hungry))
     (:V (validate) :N (validation validity) :AJ (valid))
     (:V (cross) :N (crossing cross) :P (across))
```

A more detailed discussion of the construction and evaluation of CatVar is available in (Habash and Dorr 2003).

3.3.1.3. The Syntactic-Thematic Linking Map This is a large matrix extracted from LCS-DB (Dorr 2001). It relates syntactic relations to thematic roles. Syntactic relations include 125 prepositions in addition to *:subj* (subject), *:obj* (object) and *:obj2* (indirect object). These are mapped to varying subsets of the 20 different thematic roles used in our system, e.g. *ag* (agent), *instr* (instrument), *th* (theme), *exp* (experiencer), *loc* (location), *src* (source), *perc* (perceived), *poss* (possessed), *purp* (purpose), *ben* (benefactor) and *prop* (proposition). The total number of links is 341 pairs. The following is an excerpt:

```
(9)  (:subj -> ag instr th exp loc src goal perc poss)
     (:obj2 -> goal src th perc ben)
     (across -> goal loc)
     (in_spite_of -> purp)
     (in -> loc perc goal poss prop)
```

3.3.1.4. Structural n-gram model Statistical *surface n*-gram LMs capturing patterns of local co-occurrence of contiguous words in sentences have been used in various hybrid implementations of NLG and MT systems (Brown and Frederking 1995; Knight and Hatzivassiloglou 1995; Langkilde and Knight 1998a; Bangalore and Rambow 2000a; Habash et al. 2003). Other types of LMs that capture long-distance relationships, such as probabilistic context-free grammars (PCFG) or lexicalized syntax models, have been used in the parsing community with impressive improvements in parsing correctness (Charniak 1997; Collins 1997; Sima'an 2000; Charniak and Johnson 2001). In comparison, only one large-scale system built with NLG in mind uses a structural LM (Bangalore and Rambow 2000a). Additionally, the IBM Air Travel Reports system, which implements a dependency *n*-gram LM, uses templates and focuses on travel reports only (Ratnaparkhi 2000). A study by Daumé et al. (2002)

[10] An investigation of the existence of such a resource showed that none was available at the time it was created (ca. 2002). Since then, the WordNet project has added such links but only for Nouns and Verbs (Fellbaum 1998).

[11] An English Verb-Noun list extracted from Nomlex was provided by Bonnie Dorr and Greg Marton.

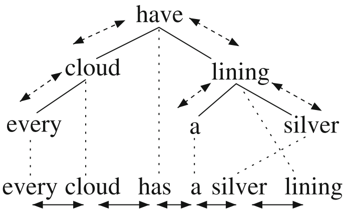

Fig. 2 S*n*-grams vs. *n*-grams

used the Charniak parser (Charniak 2000) as a lexicalized syntax model for generation purposes. They demonstrated the usefulness of these models in a variety of NLG tasks.

Whereas syntax models address both parent-child relationships and sisterhood relationships, the S*n*-gram LM characterizes the relationship between words in a dependency representation of a sentence without taking into account the overall structure at the phrase level. In other words, an independence in the behavior of the children relative to each other (their sisterhood relationships) is assumed in S*n*-grams.

Figure 2 exemplifies the differences between S*n*-grams (dashed arrows) and *n*-grams (solid arrows). In addition to capturing long-distance relationships between words (e.g. *have/has* and *lining*), S*n*-grams are based on uninflected lexemes, not on inflected surface forms. Therefore S*n*-grams can model more general relationships between lexical items. Moreover, the effect of S*n*-grams is only seen on lexical selection whereas the *n*-gram statistical ranking determines both lexical selection and linearization. Therefore, the two models are complementary in many ways.

The S*n*-gram LM is created using 500,000 parsed sentences (15 million words) from the English newswire text in the English Gigaword corpus (Graff 2003b). The text is parsed using Connexor's English parser (Tapanainen and Jarvinen 1997). The resulting set of parse trees is traversed and parent-child instance (lexeme) pairs are counted. The LM is basically the collection of these counts (totaling 1.7 million bigrams for 116,000 lexemes).[12] The use of S*n*-grams in a Spanish–English GHMT system for pruning the search space decreases runtime by 60% with no loss in translation quality (Habash 2004).

3.3.1.5. Surface n-gram model In addition to S*n*-grams, we use a surface trigram LM trained on approximately 340 million words of English newswire text from the English Gigaword corpus. The LM is implemented using the SRILM toolkit (Stolcke 2002).

3.3.2 Generation processes

The EXERGE module consists of seven generation sub-modules: *input normalization/restructuring*, *thematic linking*, *structural expansion*, *syntactic assignment*, *structural n-gram filtering*, *linearization*, and *statistical extraction* (surface *n*-gram modeling). Each of these is discussed below.

[12] In this work, we only consider the value of $n = 2$ (structural bigram).

3.3.2.1. Input normalization and restructuring The input syntactic dependency is converted into a *normalized* dependency representation in which prepositions are treated as *syntactic relation* edge labels rather than nodes. Edges in this normalized dependency have two simultaneous specifications: (1) syntactic relation (as in *:subj*, *:obj*, or a preposition or list of prepositions) and (2) thematic role (as in *ag*, *th*, or *goal*). See, for instance, example sentence (10) in Sect. 3.3.2.3.

While the initial syntactic relations are extracted from the input dependency parse, the thematic roles are generated in the thematic linking step (discussed next). The TL syntactic relations are reassigned in the later step of syntactic assignment (Sect. 3.3.2.4).

The normalized input is then inspected for any violations of wellformedness. When such cases are detected, they are corrected by restructuring the dependency. For example, if an SL verb that dominates another verb is translated into an auxiliary, such as the Spanish verb *poder* translating into 'can/could', the dependency is restructured to make the dependent verb the dominating head and the auxiliary its child.

3.3.2.2. Thematic linking The next step in generation is to turn the normalized SL syntactic dependency (with TL words) into a thematic dependency by assigning thematic roles to all edges. This step deceivingly resembles SL analysis in other MT approaches. However, the crucial difference is that linking is applied to TL words using TL resources with no knowledge of SL syntactic-thematic linking preferences (except through the choice of TL words used in the translation step). This is a *loose* linking algorithm that results in many possible syntactic-thematic links. Prepositions are treated as syntactic relations that constrain the option of thematic roles that can be assigned to their objects. For example, if a certain SL preposition is translated as the English prepositions 'to/toward/at', it is safe to assume that the object of the preposition is *goal* or *loc* but not *src* or *purp*. Words that are not assigned a thematic role receive the default role *mod* (modifier). This step makes use of the word-class lexicon (Sect. 3.3.1.1) and the syntactic-thematic linking map (Sect. 3.3.1.3).

3.3.2.3 Structural expansion This step overgenerates alternative structural configurations of the thematic dependencies. Two operations are applied here: *conflation* and *head swapping*. Lexical-semantic information from the word-class lexicon (theta grids and LCS primitives) is used to determine the conflatability and head-swappability of combinations of nodes in the trees. This step makes use of the word-class lexicon (Sect. 3.3.1.1) and the categorial variation database (Sect. 3.3.1.2).

For each argument of a given verb in the tree, the head verb (V_{head}) and argument (Arg) pair are checked for *conflatability*. A pair is *conflatable* if: (1) there exists a verb V_{conf} that is a categorial variation of Arg; (2) V_{conf} and V_{head} both share the same LCS primitives; and (3) V_{conf} can assign the same thematic roles that are assigned by V_{head} except for the role assigned to Arg. For example, the Spanish sentence *Yo le di puñaladas a Juan (I gave stabs to Juan)* results in the following thematic dependency tree after linking is done:

(10) (give :subj:ag (I) :obj:th (stab) :to/at:goal (Juan))

The theme *stab* has a verb categorial variation that belongs to two different verb classes, the *poison verbs* (as in *crucify, electrocute*, etc.) and the *swat verbs* (as in *bite, claw*, etc.). Only the first class shares the same LCS primitives as the verb *give* `(cause go)`. Moreover, the verb *stab* requires an agent and a goal. Therefore, a conflated instance is created in this case that can be generated later as *I stabbed Juan*:

(11) (stab :subj:ag (I) :to/at:goal (Juan))

Unlike conflation, head swapping is restricted to head-modifier pairs. Every such pair's swappability is determined by the following criteria: (1) there exists a verb V_{child} that is a categorial variation of the modifier; (2) there is a categorial variation of V_{head} that can become a child of V_{child} such as a noun, adjective, adverb or even a preposition; and (3) all the arguments before the swapping retain their thematic roles regardless of whether they move with the swapped verb or not. For example, the German *ich esse gern (I eat likingly)* would result in the following thematic dependency tree after linking is done:

(12) (eat :subj:th (I) :mod:mod (like))

Here the modifier *like* and the main verb *eat* can be swapped to produce *I like eating* or *I like to eat* from the following:

(13) (like :subj:th (I) :mod:mod (eat))

If the demoted verb can become a preposition, the swapping is more complicated since prepositions are syntactic labels (not nodes) in the dependency. For example, the Spanish *Juan cruzó el río nadando (Juan crossed the river swimming)* results in the following thematic dependency tree after linking is done:

(14) (cross :subj:th (Juan) :obj:loc (river) :mod:mod (swim))

The modifier *swim* is itself a verb, and the main verb *cross* has a prepositional categorial variation *across* which can assign the thematic role *:loc* to *river*. The head swapping results in the following dependency:

(15) (swim :subj:th (Juan) :across:loc (river))

3.3.2.4. Syntactic assignment During this step, thematic dependencies are turned into full TL syntactic dependencies. Syntactic positions are assigned to thematic roles using the verb class subcategorization frames associated with verbal heads in the thematic dependencies. Different alternations associated with a single class are also generated. Class category specifications are enforced by selecting appropriate categorial variations for different arguments. For example, the main verb for the Spanish *tengo hambre (I have hunger)* translates into 'have/own/possess/be'. For the last verb (*be*), there are different classes that have different specifications for the verb's second argument: a noun or an adjective. This, of course, results in *I am hungry* and *I am hunger* in addition to *I (have/possess/own) hunger*. A later step determines which sequence is more likely.

3.3.2.5. Structural n-gram filtering and expansion Sn-grams (see Sect. 3.3.1.4) are used in EXERGE for (1) lexical selection and (2) structural selection. First, Sn-grams are used to prune the ambiguous nodes in the forest of syntactic dependencies produced after the structural expansion and syntactic assignment steps. This pruning is motivated by the need to control the size of the word lattices passed on to the n-gram LM, which tends to be the most expensive step in the whole system. For each tree T in the forest, a bottom-up dynamic programming algorithm (Cormen et al. 2001) is used to calculate the maximum (joint) frequency of (word, parent$_{\text{word}}$) over all $words$ in the $nodes$ of T.[13] Once the scoring is completed, selecting the best unambiguous tree using the dynamic programming tables is straightforward.

Secondly, since the symbolic resources used in the structural expansion phase focus on verbs only, an Sn-gram-driven approach is used to expand the structure of *noun phrases*. This addresses cases where the direct modification relationship between two English nouns is expressed in Spanish using a preposition.[14] This process is done as follows. For every parent-child pair of nominal nodes linked by a $preposition$, the pair is determined to prefer a direct modification relation over $preposition$ if the Sn-gram frequency of (child, parent) is higher than the frequency of (preposition, parent) or the frequency of (child, preposition). For example, the preposition in the Spanish *el mundo* **en** *desarrollo (the world* **in** *development/developing)* is replaced by a direct modification relationship since the Sn-gram (*developing world*) is more common than (*world in*) and (*in development/developing*) in English.[15] Structural variations in noun-noun modification are common in translation between English and other languages, e.g. Japanese (Tanaka and Baldwin 2003).

The use of Sn-grams in EXERGE for both lexical and structural choice in a large-scale translingual setting is a major difference from FERGUS's use of Sn-grams for lexical choice only in a monolingual setting. Nitrogen does not use Sn-grams.

3.3.2.6. Linearization In this step, a rule-based linearization grammar is used to convert the sentence syntactic representation into a word lattice that encodes the different possible linear-order realizations of the sentence. The grammar is implemented using the linearization engine oxyGen (Habash 2000; Habash et al. 2003) and makes use of the morphological generation component of the generation system Nitrogen (Langkilde and Knight 1998b).

3.3.2.7. Statistical extraction The final step extracts an n-best list of sentences from the word lattice using the trigram LM. This step is done completely using the n-gram LM (Sect. 3.3.1.5) without passing any probabilities from previous steps. Most of the previous steps are symbolic, except for Sn-gram filtering and expansion, and thus do not have associated probabilities.

[13] In a different version of the system, the conditional probability, $P(\text{word}|\text{parent}_{\text{word}})$, is used with no significant effect. This is consistent with findings in the parsing community (Johnson 2001).

[14] The technique presented here for structural selection using Sn-grams can be used in reverse to allow translation of direct noun modification in the SL to prepositional modification in English.

[15] A relevant discussion of the translation of noun-noun compounds from Spanish to English is available in Aymerich (2001).

3.4 Summary of previous results: Spanish–English GHMT

A Spanish–English GHMT system was implemented by Habash (2003b). It was evaluated using test sets from three corpora: the United Nations (UN) corpus (Graff 1994), the FBIS corpus[16] and the Bible (Resnik et al. 1999). These corpora were selected to cover a wide range of genres so that the behavior of the evaluated systems could be examined under different conditions. This is important because resource poverty forces systems to be trained or built using whatever resources are available, which may not necessarily be the same as what needs to be translated. GHMT was evaluated against Systran,[17] an IBM Model 4 statistical MT system (Brown et al. 1993), and a gisting baseline (Resnik 1997). The statistical MT system was trained on the UN corpus. The UN corpus was also used for building the LMs used. A BLEU-based (Papineni et al. 2002) automatic evaluation showed that although GHMT scores lower than IBM Model 4 on the UN genre (IBM Model 4 training genre), GHMT has a higher degree of robustness and scores higher when tested on text with a new genre (Bible). Systran and gisting were the best and worst, respectively, for all genres. A qualitative evaluation focusing on the output's lexical choice, information loss, grammaticality and translation-divergence handling was conducted. It showed that, compared to the statistical MT system, GHMT made lexical choice errors that were less different from human references, lost fewer constituents and information, and handled translation of tense and pro-drop restoration much better.

4 Retargetability to Arabic–English GHMT

As described earlier, the English-targeted GHMT system can be used with a new SL given that an appropriate analysis and a word-based translation lexicon are provided. In this section, we use Arabic as an example and describe the steps that have to be accomplished in order to adapt GHMT to Arabic.

4.1 Arabic linguistic issues

Arabic is a morphologically rich language whose automatic processing faces many challenges at the levels of orthography, morphology and syntax (Habash 2007a).

First, Arabic is written with optional diacritics that mostly represent short vowels and consonantal doubling. Absence of diacritics leads to a high degree of ambiguity.

Second, Arabic words are morphologically complex with a large set of morphological features such *person, number, gender, voice, aspect, mood,* and *case*. Arabic features are realized using both concatenative (affixes and stems) and templatic (root and patterns) morphology with a variety of morphological, phonological and orthographic adjustments. In addition, Arabic has a set of very common clitics that are written attached to the word, e.g. the conjunction *w*+ 'and', the preposition *l*+ 'to/for',

[16] The FBIS corpus is a multilingual news corpus distributed by the Linguistic Data Consortium (LDC2003E14).

[17] http://www.systransoft.com/.

the definite article *Al+* 'the' and a range of pronominal clitics that can attach to nouns and verbs. For instance, the word *wlmktbAthm* 'and for their libraries' can be analyzed as *wa+li+mak.tab+At+hm* 'and+for+library+plural+their'. The stem *maktab* can be further decomposed templatically to include a trilateral root *k-t-b* 'writing-related' and the pattern *ma12a3*, often associated with 'location'.[18] Much work has been done on Arabic tokenization for statistical MT (Habash and Sadat 2006). We do not investigate different tokenizations in this article. We discuss our tokenization of choice in Sect. 4.2.

Third, Arabic is syntactically quite different from English. For example, Arabic verb subjects may be: (a) pro-dropped (verb conjugated), (b) pre-verbal, or (c) post-verbal. Only the pre-verbal cases are directly comparable to English, while the pro-dropped cases require generating the correct subject and the post-verbal cases require reordering. An added challenge is that identifying the correct verb-subject configuration is not always easy for automatic parsers. Another example of a syntactic difference between Arabic and English is that Arabic adjectival modifiers typically follow their nouns. Finally, Arabic has a special nominal construction called *Idafa*, in which possessive modifiers immediately follow what they modify without connecting particles. For example, *the student's book* or *the book of the student* can only be expressed in Arabic as *ktAb AlTAlb* '[lit. book the-student]'.[19]

We are aware of two published approaches on Arabic symbolic MT: one transfer approach (Sharaf 2002; Alsharaf et al. 2004) and one interlingual approach (Soudi et al. 2002; Abdel-Monem et al. 2003; Soudi 2004). Two of the top Arabic MT companies using symbolic MT systems are Apptek[20] and Sakhr.[21]

4.2 Arabic analysis

Below, we describe the necessary steps for processing an Arabic input sentence into an acceptable input format for EXERGE. All of the steps described here occur within the "Analysis" block in Fig. 1.

The core component of Arabic analysis is the statistical parser of (Bikel 2002) trained out-of-the-box on the Penn Arabic Treebank (PATB) part 1 (Maamouri et al. 2004). The parser assumes the same kind of tokenization used in the PATB as input[22] and it produces unlabeled syntactic constituencies. These two features of this off-the-shelf component, in addition to the issue of efficiency of parsing long Arabic sentences,[23] are the reasons behind the application of additional pre- and post-processing steps that we discuss next. The input sentences must first be tokenized and POS-tagged

[18] In templatic morphology, the pattern (or template) provides positions (marked here with digits) for root radicals to be inserted.

[19] This is only true for Modern Standard Arabic. We do not discuss Arabic dialects and their challenges here.

[20] http://www.apptek.com/.

[21] http://www.sakhr.com/.

[22] PATB tokenization splits off conjunctions, particles and pronominal clitics, and normalizes the word stem.

[23] The complexity of Bikel's parser is $O(n^5)$.

before being parsed. We additionally chunk the sentences to ensure that their length is computationally manageable for parsing. The output of the parser is modified to turn it from an unlabeled surface-form constituency into a labeled lexeme-based dependency.

4.2.1 Pre-parsing

The pre-parsing phase involves two processes: one at the word level (tokenization and POS tagging) and one at the sentence level (chunking).

4.2.1.1. Tokenization and POS tagging For tokenization, we use the PATB tokenization scheme, which is most compatible with statistical parsers trained on the PATB (Maamouri et al. 2004). We preprocess the text using the MADA (Morphological Analysis and Disambiguation of Arabic) tool (Habash and Rambow 2005), together with TOKAN, a general tokenizer for Arabic (Habash 2007a). MADA uses the ALMORGEANA (Arabic Lexeme-based Morphological Analysis and Generation) system (Habash 2007a), which itself uses the databases of the Buckwalter Arabic Morphological Analyzer (Buckwalter 2002). For the POS tagset, we use the PATB collapsed tagset (24 tags) often used for parsing Arabic (Diab et al. 2004).[24]

4.2.1.2. Chunking We employ a rule-based segment chunker implemented with Perl regular expressions to address two issues. First, 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 news by-lines requires template handling in analysis and generation, which needs to be updated depending on the genre. We choose to preserve SL 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 all of which use POS information: Arabic conjunction proclitic *w*/CC 'and', numbers (CD), punctuation (PUNC), and the subordinating conjunction *An*/IN 'that'.[25] On average, sentences have 3.3 chunk separators.

4.2.2 Post-parsing

The post-parsing phase involves three processes: mapping constituency to dependency, relation labeling, and lexeme selection.

4.2.2.1. Constituency to dependency We convert constituencies to dependencies using modified head-percolation rules from the Bikel parser (Habash and Rambow 2004) applied with the Const2Dep tool (Hwa 2001). We additionally apply the following restructuring operations to the dependency tree:

[24] Since this research was conducted, newer collapsed sets have been shown to have better performance for Arabic parsing than the 24-tag set we used (Kulick et al. 2006).

[25] The Arabic conjunction proclitic *w*/CC 'and' is deleted altogether in the sentence-initial position, where it often appears.

- Idafa handling: noun-noun Idafa constructions are modified to include an intervening node that has no surface value but is glossed to ‘of/’s/*empty*’.
- Al+ handling: the untokenized proclitic *Al*+ ‘the’ is turned into a separate node that is a dependent of the word to which it is attached.
- Feature mapping: Arabic-specific features are mapped to language-independent features used in EXERGE. For example, the untokenized proclitic *s*+ ‘will’ is mapped to the feature `tense:future` and the Arabic perfective aspect verb feature is turned into `tense:past`.

4.2.2.2. Relation labeling As discussed above, an Arabic subject may be: (a) pro-dropped (verb conjugated), (b) pre-verbal (full conjugation with verb), or (c) post-verbal (third-person 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 case (c), where the adjacent noun is a subject. A verb may 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, we fill the subject position with an empty category (*PRO*) and mark the voice feature as passive. In principle, Arabic’s rich case system can account for the different configurations and can also allow many variations in order, but since most cases are diacritical (and thus optionally written), that information is not always available. Arabic prose writers generally avoid such syntactic acrobatics. We use heuristics based on Arabic syntax to determine the relation of the verb to its children.

4.2.2.3. Lexeme selection MADA is a morphological disambiguation tool that makes no sense-disambiguation choices.[26] Therefore, multiple lexemes are still available as ambiguous options at the tree nodes. In some cases, the parser overrides the POS tag that is chosen initially by MADA. As a result, morphological analysis must be revisited. The ALMORGEANA system (Habash 2007a) is reapplied on the tokenized words and then analyses are filtered using the criteria below. If 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. 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 is trained on PATB (part 1 and part 2)[27] and 1 million words from Arabic Gigaword (Graff 2003a) disambiguated using MADA. The lexemes are chosen based on their unigram frequencies. Lexeme ties are broken with feature unigram frequencies.

[26] Recent improvements on MADA allow better lexeme selection (Roth et al. 2008).

[27] Later parts of PATB were not available at the time of conducting this experiment.

Table 2 Entries in the BAMA stem lexicon (BAMA-LEX)

Lexeme	Stem	Category	English gloss	POS
kuwfiy˜	kuwfiy˜	Nall	of/from Kufa (Iraq);Kufic	`ADJ`
kAtib_1	kAtib	N/ap	writer;author	
	kAtib	N/ap	clerk	
	kut˜Ab	N	authors;writers	
	katab	Nap	authors;writers	
katab-u_1	katab	PV	write	
	k.tub	IV	write	
	kutib	PV_Pass	be written;be fated;be destined	
	k.tab	IV_Pass_yu	be written;be fated;be destined	

4.3 Arabic lexical translation

As was mentioned earlier, the GHMT framework requires the existence of a lexeme-based dictionary. For Arabic, there are few available dictionaries, and only one fits our needs in terms of availability to researchers and structural consistency across lexemes (dictionary entries), namely the lexicon of the Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter 2002). In this section we describe the necessary steps to transform this resource into something readily usable by our GHMT system.

The main challenge of using the BAMA lexicon (BAMA-LEX) is that its entries are organized around word stems, which are surface realizations of underlying lexemes inflected for various features such as number, aspect and voice. The English glosses in BAMA-LEX are paired with stems, not lexemes. The lexemes appear in the lexicon as comments that mark sections of related stems. Table 2 illustrates the entries associated with three lexemes: *kuwfiy˜_1*, *kAtib_1* and *katab-u_1*. Each entry in BAMA-LEX consists of four components (columns two to five in Table 2). The first component (column two) contains the diacritized stems.[28] The second component (column three) specifies an ortho-morphological category that controls what affixes can be attached to the stem. The third component (column four) contains one or more English glosses. The last component (column five) is optional. It 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.

[28] The actual lexicon also includes the undiacritized form, which we exclude here for presentation purposes.

Table 3 Entries in BAMA-LEX after our cleaning operations

Lexeme	POS	English glosses
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

- Gloss depluralization. A plural gloss is discarded if the singular form of the gloss is used for a different stem of the same lexeme. For example, the gloss “writers” for a plural stem of the lexeme *kAtib_1* in Table 2 is removed since the singular form “writer” appears under a different stem of the same lexeme.
- 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.

Table 3 presents the entries in our final lexicon that correspond to those in Table 2.

5 Statistical extensions to GHMT

The goal of extending GHMT with statistical components is to exploit GHMT’s focus on phrase-structure generation (global level) together with a PBSMT system’s robustness (local phrases). One particular case in Arabic that we investigate 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 pre-verbal position is easy for GHMT (given a correct parse) but can be hard for a PBSMT system, especially with subject noun phrases exceeding the system’s distortion limit (i.e. the number of words that can be jumped over). An example is shown next. Bolding and italics are used to mark the verb and subordinating conjunction that surround the subject noun phrase (NP) (12 words) in Arabic and what they map to in English, respectively.

(16) **[V AςIn]** [NP-SBJ Almnsq AlςAm lmšrwς Alskħ AlHdyd byn dwl mjls AltςAwn Alxlyjy HAmd xAjh] *[SUB An . . .]*

(17) [NP-SBJ The general coordinator of the railroad project among the countries of the Gulf Cooperation Council, Hamid Khaja,] **[V announced]** *[SUB that . . .]*

The main challenge in integrating statistical and symbolic systems is finding a common definition for the basic units being manipulated. Here we discuss the challenges and solutions involved in defining basic tokens (words) and phrases in a GHMT system integrated with statistical components.

5.1 Words

Since Arabic is a morphologically rich language, many different possible representations of Arabic morphological tokens can be, and have been, used in different resources for Arabic NLP (Habash 2007a).

For statistical MT, in principle, it does not matter what level of morphological representation is used, as 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. Shallower representations tend to maintain distinctions among morphological forms that might not be relevant for translation, thus increasing the sparsity of the data. This point interacts with the MT language pair. For example, normalizing subject inflections of Arabic verbs when translating to a morphologically poor language like English might be desirable since it reduces sparsity without potentially affecting translation quality. If the TL is morphologically richer, such as French, that would not be the case. This, of course, may not be a problem when large amounts of training data are available. Additionally, transforming the training text to deeper representations comes at a cost since selecting a deeper representation involves some degree of morphological disambiguation, a task that is typically neither cheap nor foolproof (Habash and Rambow 2005).

Symbolic MT approaches tend to capture more abstract generalizations about the languages they translate in contrast 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/punctuation tokenization. However, a statistical parser (such as Collins 1997 or Bikel 2002) trained out-of-the-box on PATB (Maamouri et al. 2004) assumes the same kind of tokenization it uses. This means that a separate tokenizer is needed to convert input text to this representation (Diab et al. 2004; Habash and Rambow 2005).

An additional issue with a treebank-trained statistical parser is that its input/output is in a normalized segmentation format 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 SL and TL. 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 mainly interested in MT.

The common representation we use to bring together the statistical and symbolic components is the PATB tokenization, which is in the format expected by the parser (Bikel 2002) and has been shown to produce a competitive statistical MT performance (Habash and Sadat 2006; Sadat and Habash 2006). Deeper representations involving lexemes and features are used in GHMT (but limited by/respecting the PATB tokenization).

5.2 Phrases

The second challenge for integrating statistical MT components into GHMT is that the notion of a *phrase* (anything beyond a single word) is radically different. On one hand, the traditional concept of *phrase* refers more or less to a linguistic constituent. Systems that use syntactic parsers, like GHMT, follow this definition and assume a phrase has a linguistic structure that defines it and its behavior in a bigger context. On the other hand, PBSMT systems redefine the concept of a *phrase* to mean any sequence of words with or without an underlying linguistic structure (Koehn et al. 2003). From the point of view of building MT systems, both kinds of *phrases* come with problems.

Statistical phrases are created from alignments potentially containing errors, which affect phrase extraction and may result in *jagged* phrase edges, a similar problem to *boundary friction* often cited in Example-based MT (Brown et al. 2003). 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 hallucinations*, i.e. suboptimal phrase translations resulting from bad alignments. For example, in the phrase table we use, there is an entry translating [*AlswdAn w*], literally 'Sudan and', into 'enterprises and banks'. Of course, when alignments are good, they provide solid translations that include enough context to guide when they are used.

Linguistic phrases come with a different set of problems. Since parsing technology for Arabic is still behind English,[29] many linguistic phrases are misparsed creating *symbolic hallucinations* that affect the rest of the system. A common example is the incorrect attachment of a sentential modifier, e.g. a prepositional phrase, to an NP rather than to an S node.

We investigate two variants of a basic approach to using statistical phrases in the GHMT system. The PBSMT system we use is Pharaoh (Koehn 2004a). 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. We considered more complex solutions that use statistical phrases covering parts of a linguistic phrase, but decided against them as they would require major changes to the implemented structure of the GHMT system.

In one variant of our approach (GHMT + Phrase Table, henceforth GHPHT), 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. This approach is comparable to other research on adding statistical components in symbolic systems (Dugast et al. 2009).

In a second variant, (GHMT + Pharaoh, henceforth GHPHAR), we use Pharaoh to translate the subtree projections for all the subtrees in the input sentence. These

[29] The parser we use for this article is among the best available at the time, yet its performance for Arabic is in the low 70s for percentage as measured by the labeled constituency PARSEVAL F-1 score (Black et al. 1991).

translations are added as alternatives to the basic GHMT system. Results comparing these two variants and a few others are described in Sect. 6.

6 Evaluation

Below we evaluate the two Arabic–English extended GHMT variants we presented in Sect. 5 and compare them to four other systems and baselines. We present the specific variants evaluated and various experimental results and relevant discussions.

6.1 Systems and resources

We compare six system variants:

- GIST is a simple gisting system that produces a sausage lattice from the English glosses in BAMA-LEX.[30] Arabic word order is preserved and English realization is limited to the variants provided in BAMA-LEX.
- GHMT is the system described in Sect. 4. The lexical translation is limited to BAMA-LEX.
- GHPHT is a variant of GHMT that uses a statistical phrase table as support for a multi-word surface dictionary (see Sect. 5).
- GHPHAR is a second variant discussed in Sect. 5. It uses Pharaoh to generate subtree phrases.
- PHARBW is the Pharaoh PBSMT system trained on the basic training set in addition to the entries in BAMA-LEX.
- PHARAOH is the Pharaoh PBSMT system trained only on the basic training set.

We use the standard NIST MTEval datasets for the years 2003, 2004 and 2005 (henceforth MT03, MT04 and MT05, respectively).[31] MT03 and MT05 are known to behave similarly but MT04 is different in that it is a mixed genre data set: out of 200 documents in that set, 100 are news, 50 speeches and 50 editorials. The training data we use is all Arabic news. For the fully statistical MT systems (PHARBW and PHARAOH), the 2002 MTEval test set is used for parameter estimation using *minimum error rate training* (MERT) (Och 2003). No MERT was done for the other variants.

We use an Arabic–English parallel corpus of about 5 million words to train the translation model.[32] We use the default set of features in Pharaoh, which includes the conditional translation probabilities in both directions, the lexical word translation probabilities, phrase penalty, distance-based distortion, word penalty, and LM probability (Koehn 2004a).

[30] A lattice is a sausage lattice if for each pair of vertices v_i and v_j, if there is an arc from v_i to v_j, then all arcs from v_i have to end in v_j. The name 'sausage' refers to the shape of the lattice which resembles a sausage.

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

[32] The Arabic–English corpus is distributed by the Linguistic Data Consortium: http://www.ldc.upenn.edu. The parallel text includes Arabic News, eTIRR, English translation of the Arabic Treebank, and Ummah.

Table 4 System comparison points

Component	Gist	GHMT	GHPhT	GHPhar	PharBw	Pharaoh
Tokenizer		X	X	X	X	X
Parser		X	X	X		
BAMA-LEX	X	X	X	X	X	
Catvar		X	X	X		
Sn-grams		X	X	X		
Full PhT				X	X	X
Projective PhT			X	X	X	X
LM	X	X	X	X	X	X
TrueCaser	X	X	X	X	X	X
MERT					X	X

PhT is an abbreviation for *phrase table*

For Arabic preprocessing, the PATB scheme is used (Habash and Sadat 2006; Habash 2007a). English preprocessing simply includes lower-casing, separating punctuation from words and splitting off "'s". All systems use the same surface trigram LM, trained on approximately 340 million words of English newswire text from the English Gigaword corpus and implemented using the SRILM toolkit (Stolcke 2002). As a post-processing step, the translations of all systems are true-cased. True casing is done using an in-house system trained on the English LM data and implemented using the *disambig* tool from the SRILM toolkit (Stolcke 2002) to map uncased English to cased English.

Table 4 summarizes the main variation points among the six evaluated systems.

6.2 Results and discussions

We conduct four sets of evaluations to explore different aspects of the data sets and system variants. The first is an automatic full system evaluation. The second is an automatic genre-specific evaluation. Although automatic metrics (especially BLEU) are commonly used in the field to report and compare MT performance, we are aware of their shortcomings (Callison-Burch et al. 2006). Therefore, we present additional evaluations that are more sensitive. Specifically, the third evaluation is a qualitative evaluation of certain linguistic phenomena, while the fourth evaluation is an automatic evaluation that uses rich linguistic information (English parses). We also examine different sub-scores of the different automatic metrics to obtain a more complete picture of the different systems' behavior variations.

6.2.1 Full system evaluation

For the purpose of the full system evaluation, we use three automatic evaluation metrics:

- **BLEU** computes the n-gram precision of matching n-grams between the MT output and a set of human reference translations. The different n-gram precision values for all n are averaged by taking the geometric mean. Additionally, a system-level brevity penalty is applied when translations are shorter than the human references. See (Papineni et al. 2002) for more details.
- **NIST** computes a variant of BLEU using the arithmetic mean of weighted n-gram precision values (Doddington 2002).
- **METEOR** computes a score that combines unigram-precision, unigram-recall, and a measure of fragmentation that is designed to directly capture how well-ordered the matched words in the MT output are in relation to the reference. Unigram matches are computed using multiple types of matching (surface forms, stemmed forms, and WordNet synonyms) (Banerjee and Lavie 2005). METEOR puts a bigger weight on *recall* as opposed to *precision*, unlike the other two metrics that are only precision-based. We expect this to give us different insights into the results.

Scores for BLEU and METEOR range between 0 (worst) and 1 (best). NIST scores are positive with 0 being the worst possible. Both BLEU and NIST are used with 1- to 4-g and with case sensitivity. All scores are computed against four references.

The results of the full systems are presented in Table 5. The row marked with "±" specifies the confidence interval of the BLEU score.[33]

In terms of BLEU and NIST scores, the lowest performing system is GIST, as expected. GHMT, using only BAMA-LEX and no other bilingual 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. GHPHT and GHPHAR provide substantial improvements over GHMT. In GHPHT, only 54.6% of subtrees find a match in the phrase table, as opposed to GHPHAR 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 statistical components. However, the scores are still lower than the fully statistical system. We discuss the differences further in Sect. 6.2.3. The primarily statistical systems PHARAOH and PHARBW outperform all other systems. PHARAOH does better than PHARBW for MT03 and MT05 but not for MT04. For all three data sets, the pair-wise differences among all the systems are statistically significant (except for the differences between the pair PHARAOH and PHARBW).

The METEOR scores are generally in line with the BLEU and NIST scores, meaning that BLEU or NIST improvements also result in METEOR improvements, except in the case of GIST and GHMT, which are indistinguishable for METEOR. This perhaps highlights a weakness in METEOR, which deemphasizes higher n-gram surface-form

[33] The confidence intervals are computed using bootstrap resampling (Koehn 2004b), where the samples are drawn from the test set which is divided into many blocks and BLEU is computed on each block separately. The mean and variance of these block BLEU numbers are then used to compute a confidence interval around the mean. The spread of the interval is then transferred to the BLEU score computed on the entire test set (Press et al. 2002; Och 2003).

Table 5 True-cased results of various systems on NIST MTEval test sets

Test set	Metric	GIST	GHMT	GHPHT	GHPHAR	PHARBW	PHARAOH
MT03	BLEU	0.0811	0.1479	0.2362	0.3379	0.4128	0.4162
	±	0.0115	0.0143	0.0212	0.0286	0.0164	0.0160
	NIST	5.1846	6.0528	7.3213	8.2569	9.9205	9.9300
	METEOR	0.4452	0.4402	0.5071	0.5746	0.6766	0.6774
MT04	BLEU	0.0651	0.1402	0.2110	0.2777	0.3546	0.3522
	±	0.0068	0.0086	0.0150	0.0211	0.0219	0.0208
	NIST	4.3904	6.0935	7.0981	7.5834	9.2038	9.1291
	METEOR	0.4385	0.4496	0.4958	0.5517	0.6182	0.6157
MT05	BLEU	0.0607	0.1450	0.2313	0.3239	0.3935	0.3960
	±	0.0057	0.0078	0.0099	0.0184	0.0156	0.0161
	NIST	4.7259	6.2636	7.4836	8.3687	9.6980	9.6615
	METEOR	0.4438	0.4628	0.5243	0.5911	0.6667	0.6633

Table 6 Breakdown of the METEOR scoring statistics

Test set	Parameter	GIST	GHMT	GHPHT	GHPHAR	PHARBW	PHARAOH
MT03	Score	0.4452	0.4402	0.5071	0.5746	0.6766	0.6774
	Matches	8352	7742	8287	8793	10324	10373
	Chunks	6363	5420	5177	4539	4999	5024
	Precision	0.5745	0.5428	0.6224	0.7042	0.7507	0.7518
	Recall	0.5712	0.5302	0.5729	0.6086	0.7138	0.7146
MT04	Score	0.4385	0.4496	0.4958	0.5517	0.6182	0.6157
	Matches	18894	18444	19248	20034	22532	22453
	Chunks	14161	12839	12247	10808	12095	11985
	Precision	0.5904	0.5687	0.6407	0.7212	0.7515	0.7483
	Recall	0.5517	0.5379	0.5621	0.5876	0.6621	0.6584
MT05	Score	0.4438	0.4628	0.5243	0.5911	0.6667	0.6633
	Matches	14769	14470	15315	15960	18070	18010
	Chunks	11267	10137	9629	8058	8943	8861
	Precision	0.5874	0.5702	0.6479	0.733	0.7577	0.7559
	Recall	0.5687	0.5576	0.5937	0.6222	0.7047	0.7001

precision (remember that GIST's output are *uninflected* lexemes). When we examine the specific METEOR sub-scores (Table 6), we can see some of the differences between the two. While GIST consistently outperforms GHMT in terms of precision and recall, GHMT has a large reduction in the number of chunks relative to GIST. This is indicative of better word order. In fact, this reduction continues across the systems and reaches its lowest with the GHPHAR system. However, it is not clear whether the last two systems do not have additional reductions because they do not model syntax or because their number of matches is much higher.

Table 7 Percentage of sentences with better sentence-level BLEU score in column over row

	GIST	GHMT	GHPHT	GHPHAR	PHARBW	PHARAOH
MT03						
GIST		68.33	79.79	85.82	97.74	98.34
GHMT	30.92		74.21	86.73	93.51	94.42
GHPHT	19.91	20.66		73.60	85.82	87.78
GHPHAR	13.88	9.20	21.27		67.87	67.12
PHARBW	2.11	6.33	13.42	28.66		43.59
PHARAOH	1.66	5.43	11.76	26.85	39.52	
MT04						
GIST		19.22	64.23	89.43	95.34	95.34
GHMT	80.64		82.19	95.34	98.15	98.08
GHPHT	35.40	16.41		75.61	87.51	85.59
GHPHAR	10.42	3.33	21.29		67.26	65.26
PHARBW	4.58	1.77	12.27	30.08		41.69
PHARAOH	4.58	1.85	14.12	30.67	45.82	
MT05						
GIST		74.53	87.41	91.48	97.73	98.30
GHMT	25.28		78.69	88.83	93.47	93.66
GHPHT	12.59	18.09		75.19	85.98	86.36
GHPHAR	8.33	8.43	22.06		66.10	65.15
PHARBW	2.27	6.53	13.92	31.34		44.98
PHARAOH	1.70	6.34	13.64	30.78	42.61	

Finally, we present an oracle-based analysis and evaluation by comparing the sentence-level BLEU scores for every pair of systems we have in Table 7. The results show that, on average, almost one-third of the GHPHAR sentences outperform their PHARAOH counterpart. This is not a trivial result given the differences between these systems and it highlights their complementarity. The potential BLEU scores resulting from selecting the higher scoring sentences are presented in Table 8. On average, GHPHAR can add almost four BLEU points (or around 10% relative) to the PHARAOH score.

6.2.2 Genre evaluation

The MTEval 2004 data set is special in that it consists of a mixture of genres. We investigate the difference in behavior of our evaluated systems over different genres. Table 9 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 5 (including statistical significance). The difference across genres is very clear, with the news subset performing at a similar score level to that of the MT03 and MT05 test sets in Table 5. One big difference here from the full evaluation is the consistently higher METEOR score of PHARBW over PHARAOH.

Table 8 MT03, 04 and 05 oracle combination results in BLEU (basic system scores in parentheses)

MT03	GHMT (0.1479)	GHPHT (0.2362)	GHPHAR (0.3379)	PHARBW (0.4128)	PHARAOH (0.4162)
GIST (0.0811)	0.1661	0.2525	0.3622	0.4359	0.4391
GHMT (0.1479)		0.2577	0.3614	0.4370	0.4403
GHPHT (0.2362)			0.3695	0.4409	0.4431
GHPHAR (0.3379)				0.4596	0.4580
PHARBW (0.4128)					0.4639
MT04	GHMT (0.1402)	GHPHT (0.2110)	GHPHAR (0.2777)	PHARBW (0.3546)	PHARAOH (0.3522)
GIST (0.0651)	0.0872	0.1881	0.2939	0.3694	0.3677
GHMT (0.1402)		0.1763	0.2916	0.3697	0.3680
GHPHT (0.2110)			0.3018	0.3768	0.3759
GHPHAR (0.2777)				0.3879	0.3856
PHARBW (0.3546)					0.3963
MT05	GHMT (0.1450)	GHPHT (0.2313)	GHPHAR (0.3239)	PHARBW (0.3935)	PHARAOH (0.3960)
GIST (0.0607)	0.1622	0.2465	0.3443	0.4146	0.4169
GHMT (0.1450)		0.2509	0.3452	0.4151	0.4176
GHPHT (0.2313)			0.3545	0.4187	0.4207
GHPHAR (0.3239)				0.4384	0.4368
PHARBW (0.3935)					0.4421

This clearly must come from the additional BAMA-LEX translations not in PHARAOH. We can only see this effect using METEOR which abstracts away from surface forms.

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 different subsets as measured against the basic training data are as follows: news (2.02%), speeches (2.01%) and editorials (2.34%). The differences are very small. This indicates that it is the style/use difference that is the biggest contributor to the difference in scores.

The fact that we see similar genre-score differences in GIST and GHMT as in PHARAOH contradicts our hypothesis that GHMT is more genre-independent than statistical MT approaches. We believe this is because our Arabic linguistic resources are biased toward the news genre. For example, the Arabic treebank used for training the parser is only in the news genre. BAMA-LEX potentially also has some internal bias toward the news genre because it was developed in tandem with the Arabic treebank.

Table 9 Genre-specific true-cased results of various systems on NIST MT04 subsets

Genre	Metric	GIST	GHMT	GHPHT	GHPHAR	PHARBW	PHARAOH
News	BLEU	0.0817	0.1617	0.2582	0.3434	0.4266	0.4244
	±	0.0072	0.0101	0.0142	0.0155	0.0174	0.0167
	NIST	4.8989	6.358	7.6143	8.3132	9.7206	9.6796
	METEOR	0.4754	0.4929	0.5580	0.6187	0.6913	0.6564
Speech	BLEU	0.0429	0.1276	0.1821	0.2447	0.3088	0.3043
	±	0.0112	0.0106	0.0150	0.0205	0.0198	0.0205
	NIST	3.2993	5.3923	6.2022	6.6354	7.8796	7.7164
	METEOR	0.4029	0.4213	0.4585	0.5145	0.5652	0.5201
Editorial	BLEU	0.0575	0.1144	0.1542	0.1914	0.2704	0.2703
	±	0.0104	0.0127	0.0151	0.0196	0.0204	0.0196
	NIST	3.7633	4.9751	5.4724	5.4608	7.2344	7.1812
	METEOR	0.4071	0.3993	0.4169	0.4613	0.5308	0.4875

6.2.3 Qualitative evaluation

Automatic evaluation systems are often criticized for not capturing linguistic subtleties. This is clearly apparent in the field's moving back toward the use of human-informed evaluation metrics such as the Human Translation Error Rate (HTER) (Snover et al. 2006). We conduct a fully human evaluation of verb and subject realization in 16 random documents from MT04. The documents contained 95 sentences (3,440 words) and reflect the distribution of genres in the MT04 test set. We compare three systems: GHMT, GHPHAR and PHARAOH.

The evaluation is 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 realized or not realized 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, e.g. the Arabic verb *qrrt* 'it decided' translated to 'its decision.' 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 10 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 statistical MT systems

Table 10 Verb and subject realization in 16 documents from MT04

Genre	Verb Count	GHMT		GHPHAR		PHARAOH	
		Verbs Seen	Realized Subject	Verbs Seen	Realized Subject	Verbs Seen	Realized Subject
News	107	92	57	90	64	87	62
		86.0%	62.0%	84.1%	71.1%	81.3%	71.3%
Speech	69	57	28	56	33	44	23
		82.6%	49.1%	81.2%	58.9%	63.8%	52.3%
Editorial	67	52	29	45	29	46	26
		77.6%	55.8%	67.2%	64.4%	68.7%	56.5%
All	243	201	114	191	126	177	111
		82.7%	56.7%	78.6%	66.0%	72.8%	62.7%

(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 GHPHAR 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. This most likely 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 PBSMT translation of their projected subtree produces the right verb-subject pair in the correct relative order.

6.2.4 Span-specific syntactic dependency precision evaluation

We present here another evaluation of the three systems compared above, that allows us to isolate the performance difference on long-distance constructions.

We compute the translational structural bigram (SB) precision of each system against four references. An SB is a structural *n*-gram (Sect. 3.3.1.4.) specifying a parent and child pair in a dependency representation. This is a variant on BLEU, which has been explored by different researchers (Liu and Gildea 2005; Giménez and Màrquez 2007; Owczarzak et al. 2007; Habash and Elkholy 2008). We parse all of the translated sentences in GHMT, GHPHAR, and PHARAOH in addition to all the references for all the studied test sets. We use the MICA dependency parser (Nasr and Rambow 2006). The translational SB precision is computed as follows: for each SB in the translated output, we consider a match to be present if the same SB appears in any of the references. Similar to BLEU, we employ "clipping" to prevent more matches than the maximum number of matchable instances in a single reference sentence.

Table 11 shows, for each test set and MT system, the total count of SBs, the number of matching SBs and the percentage of matching SBs. The last set of rows (MT0X) combines the results for all test sets. For each MT system, four sets of scores are presented for different surface span size ranges. The surface span of an SB is the length of the sequence of words bordered by and including the parent and child words in the

Table 11 Precision of matching parent–child syntactic dependencies in translation against four references

	GHMT			GHPHAR			PHARAOH		
MT03									
ALL	13,648	2,355	17.3%	12,221	4,552	37.2%	14,308	6,080	42.5%
[1, 7]	12,173	2,149	17.7%	11,020	4,278	38.8%	12,860	5,772	44.9%
[8, ∞]	1,475	206	14.0%	1,201	274	22.8%	1,448	308	21.3%
[20, ∞]	411	105	25.5%	313	103	32.9%	340	90	26.5%
MT04									
ALL	32,053	5,406	16.9%	29,198	10,067	34.5%	31,454	11,564	36.8%
[1, 7]	28,274	4,951	17.5%	25,996	9,452	36.4%	28,062	10,954	39.0%
[8, ∞]	3,779	455	12.0%	3,202	615	19.2%	3,392	610	18.0%
[20, ∞]	1,113	196	17.6%	937	233	24.9%	989	181	18.3%
MT05									
ALL	24,023	4,242	17.7%	21,364	8,139	38.1%	25,067	10,046	40.1%
[1, 7]	21,249	3,874	18.2%	19,100	7,621	39.9%	22,465	9,499	42.3%
[8, ∞]	2,774	368	13.3%	2,264	518	22.9%	2,602	547	21.0%
[20, ∞]	923	195	21.1%	697	214	30.7%	719	157	21.8%
MT0X									
ALL	69,724	12,003	17.2%	62,783	22,758	36.2%	70,829	27,690	39.1%
[1, 7]	61,696	10,974	17.8%	56,116	21,351	38.0%	63,387	26,225	41.4%
[8, ∞]	8,028	1,029	12.8%	6,667	1,407	21.1%	7,442	1,465	19.7%
[20, ∞]	2,447	496	20.3%	1,947	550	28.2%	2,048	428	20.9%

We report results for all span sizes (ALL), spans less than 8 ([1, 7]), spans larger than or equal to 8 ([8, ∞]) and spans larger than or equal to 20 ([20, ∞])

SB. For example, the surface span of SB (have-lining) in Fig. 2 is four. The different ranges are: all span sizes (ALL), spans less than 8 ([1, 7]), spans larger than or equal to 8 ([8, ∞]) and spans larger than or equal to 20 ([20, ∞]).

The 'ALL' score is consistent with other automatic scores discussed earlier. However, when we consider different spans, we see a different pattern. The number of total SBs is generally comparable across different MT systems because parsers can always find a reading of some sort. For instance, a missing verb in a translation can lead the parser to consider some noun in the sentence as the verb and to produce an unlikely parse of it. SBs have a Zipfian distribution with 47% of all SBs having surface span of 2 (surface bigrams), 20% are trigrams, and so on. 89% of all SBs are 7-grams or less. We think that parses that produce longer-span SBs that are more consistent with translation references show a higher degree of grammaticality than those producing fewer matches. Although PHARAOH outperforms the other systems for the shorter spans in terms of total precision, GHPHAR outperforms for spans longer than 7 ([8, ∞]). The precision scores for very long spans ([20, ∞]) are even bigger in terms of precision and actual matches compared to PHARAOH.

7 Conclusions and future work

We presented the challenges and details of an implementation of an Arabic–English GHMT system extended with statistical MT 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 and automatic evaluation of long-distance dependency translation show that our hybrid approach (GhPhar) outperforms both the symbolic and the statistical approaches on those metrics.

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. Our evaluation 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.

Finally, although GHMT development has focused on English as a TL, the algorithmic core of GHMT is essentially language-independent. The English focus is due to the availability of the English lexical resources necessary for this generation-heavy approach. We are interested in extending this approach to target other languages in the future.

Arabic transliteration

Arabic transliterations are provided in the Habash-Soudi-Buckwalter (HSB) transliteration scheme (Habash et al. 2007). This scheme extends Buckwalter's transliteration scheme (Buckwalter 2002) to increase its readability while maintaining the 1-to-1 correspondence with Arabic orthography as represented in standard encodings of Arabic, i.e. Unicode, CP-1256, etc. The following table maps Arabic letters to their HSB transliterations. The Buckwalter transliteration is indicated in parentheses when different from HSB (Table 12).

Table 12 The Habash-Soudi-Buckwalter transliteration scheme (Buckwalter transliteration in parentheses where different)

ء	’	آ	Ā (|)	أ	Â (>)	ؤ	ŵ (&)	إ	Ǎ (<)	ئ	ŷ (})
ا	A	ب	b	ة	ħ (p)	ت	t	ث	θ (v)	ج	j
ح	H	خ	x	د	d	ذ	ð (*)	ر	r	ز	z
س	s	ش	š ($)	ص	S	ض	D	ط	T	ظ	Ď (Z)
ع	ς (E)	غ	γ (g)	ف	f	ق	q	ك	k	ل	l
م	m	ن	n	ه	h	و	w	ى	ý (Y)	ي	y
ـَ	a	ـُ	u	ـِ	i	ـً	ã (F)	ـٌ	ũ (N)	ـٍ	ĩ (K)
ـّ	~	ـْ	. (o)								

Acknowledgements This work has been supported, in part, by Army Research Lab Cooperative Agreement DAAD190320020, NSF CISE Research Infrastructure Award EIA0130422, Office of Naval Research MURI Contract FCPO.810548265, DoD Contract MDA904-96-C-1250, ONR MURI Contract FCPO.810548265, Mitre Contract 010418-7712, the GALE program of the Defense Advanced Research Projects Agency, Contracts No. HR0011-06-2-0001 and HR0011-06-C-0023, and the Human Language Technology Center of Excellence. Any opinions, findings, conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the sponsors. We would like to thank Necip Fazil Ayan and Nitin Madnani for help with some of the experiments done. We would like thank Kishore Papineni for providing us with the confidence interval computation code. We would like to thank Owen Rambow, Srinivas Bangalore and Alexis Nasr for providing us with their MICA parser. We would also like to thank Amy Weinberg for helpful conversations.

References

Abdel-Monem A, Shaalan K, Rafea A, Baraka H (2003) A proposed approach for generating Arabic from interlingua in a multilingual machine translation system. In: Proceedings of the 4th conference on language engineering. Cairo, Egypt, pp 197–206

Alsharaf H, Cardey S, Greenfield P, Shen Y (2004) Problems and solutions in machine translation involving Arabic, Chinese and French. In: Proceedings of the international conference on information technology. Las Vegas, NA, pp 293–297

Antworth E (1990) PC-KIMMO: a two-level processor for morphological analysis. Dallas Summer Institute of Linguistics, Dallas, TX

Ayan NF, Borr B, Habash N (2004) Multi-align: combining linguistic and statistical techniques to improve alignments for adaptable MT. In: Proceedings of the conference of the Association for Machine Translation in the Americas (AMTA-2004). Washington DC, USA, pp 17–26

Aymerich J (2001) Generation of noun-noun compounds in the Spanish–English machine translation system SPANAM. In: Proceedings of the eighth machine translation summit (MT SUMMIT VIII). Santiago de Compostela, Spain

Banerjee S, Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. Ann Arbor, MI, pp 65–72

Bangalore S, Rambow O (2000a) Corpus-based lexical choice in natural language generation. In: ACL 2000: 38th annual meeting of the association for computational linguistics. Hong Kong, China, pp 464–471

Bangalore S, Rambow O (2000b) Exploiting a probabilistic hierarchical model for generation. In: Proceedings of the 18th international conference on computational linguistics. Saarbrücken, Germany, pp 42–48

Beaven J (1992) Shake and bake machine translation. In: Proceedings of fifteenth [sic] international conference on computational linguistics. Nantes, France, pp 603–609

Bikel D (2002) Design of a multi-lingual, parallel-processing statistical parsing engine. In: Proceedings of HLT 2002, second international conference on human language technology conference. San Diego, CA, pp 178–182

Black E, Abney S, Flickinger D, Gdaniec C, Grishman R, Harrison P, Hindle D, Ingria R, Jelinek F, Klavans J, Liberman M, Marcus M, Roukos S, Santorini B, Strzalkowski T (1991) A procedure for quantitatively comparing the syntactic coverage of English grammars. In: Proceedings of the 1991 DARPA speech and natural language workshop. Pacific Grove, CA, Morgan Kaufmann, pp 306–311

Brown R, Frederking R (1995) Applying statistical English language modeling to symbolic machine translation. In: Proceedings of the sixth international conference on theoretical and methodological issues in machine translation. Leuven, Belgium, pp 221–239

Brown P, Della-Pietra S, Della-Pietra V, Mercer R (1993) The mathematics of machine translation: parameter estimation. Comput Linguist 19(2):263–311

Brown RD, Hutchinson R, Bennett PN, Carbonell JG, Jansen P (2003) Reducing boundary friction using translation-fragment overlap. In: MT Summit IX, Proceedings of the ninth machine translation summit. New Orleans, LA, pp 24–31

Buckwalter T (2002) Buckwalter Arabic morphological analyzer version 1.0. Linguistic Data Consortium Catalog No.: LDC2002L49

Callison-Burch C, Osborne M, Koehn P (2006) Re-evaluating the role of BLEU in machine translation research. In: Proceedings of the 11th conference of the European chapter of the association for computational linguistics (EACL'06). Trento, Italy, pp 249–256

Carbonell J, Klein S, Miller D, Steinbaum M, Grassiany T, Frey J (2006) Context-based machine translation. In: Proceedings of the 7th conference of the association for machine translation in the Americas: visions for the future of machine translation. Cambridge, MA, pp 19–28

Charniak E (1997) Statistical parsing with a context-free grammar and word statistics. In: Proceedings of the AAAI. Providence, RI, pp 598–603

Charniak E (2000) A maximum-entropy-inspired parser. In: Proceedings of the 1st North American chapter of the association for computational linguistics conference. Seattle, WA, pp 132–139

Charniak E, Johnson M (2001) Edit detection and parsing for transcribed speech. In: Proceedings of the second meeting of the North American chapter of the association for computational linguistics. Pittsburgh, PA, pp 118–126

Collins M (1997) Three generative, lexicalised models for statistical parsing. In: 35th annual meeting of the association for computational linguistics and 8th conference of the European chapter of the association for computational linguistics, proceedings of the conference. Madrid, Spain, pp 16–23

Collins M, Koehn P, Kucerova I (2005) Clause restructuring for statistical machine translation. In: 43rd annual meeting of the association for computational linguistics. Ann Arbor, MI, pp 531–540

Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to algorithms, 2nd edn. The MIT Press, Cambridge, MA

Crego JM, Mariño JB (2007) Syntax-enhanced N-gram-based SMT. In: Machine translation Summit XI, proceedings. Copenhagen, Denmark, pp 111–118

Daumé H III, Knight K, Langkilde-Geary I, Marcu D, Yamada K (2002) The importance of lexicalized syntax models for natural language generation tasks. In: Proceedings of the international natural language generation conference (INLG-02). New York, NY, pp 9–16

Diab M, Hacioglu K, Jurafsky D (2004) Automatic tagging of Arabic text: from raw text to base phrase chunks. In: Proceedings of the 5th meeting of the North American chapter of the association for computational linguistics/human language technologies conference (HLT-NAACL04). Boston, MA, pp 149–152

Doddington G (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the second international conference on human language technology research. San Francisco, CA, pp 138–145

Dorr BJ (1993a) Interlingual Machine translation: a parameterized approach. Artif Intell 63(1 & 2): 429–492

Dorr BJ (1993b) Machine translation: a view from the Lexicon. The MIT Press, Cambridge, MA

Dorr BJ (2001) LCS verb database. Technical Report Online Software Database, University of Maryland, College Park, MD (with Mari Olsen and Nizar Habash and Scott Thomas). http://www.umiacs.umd.edu/~bonnie/LCS_Database_Docmentation.html

Dorr BJ, Habash N (2002) Interlingua approximation: a generation-heavy approach. In: Workshop on interlingua reliability, fifth conference of the association for machine translation in the Americas, AMTA-2002. Tiburon, CA, pp 1–6

Dorr BJ, Jordan PW, Benoit JW (1999) A survey of current research in machine translation. In: Zelkowitz M (ed) Advances in computers. Academic Press, London, pp 1–68

Dorr BJ, Pearl L, Hwa R, Habash N (2002) DUSTer: a method for unraveling cross-language divergences for statistical word-level alignment. In: Proceedings of the 5th conference of the association for machine translation in the Americas (AMTA-02). Springer-Verlag, Berlin/Heidelberg, pp 31–43

Dugast L, Senellart J, Koehn P (2009) Selective addition of corpus-extracted phrasal lexical rules to a rule-based machine translation system. In: MT Summit XII, proceedings of the twelfth machine translation summit. Ottawa, ON, Canada, pp 222–229

El Isbihani A, Khadivi S, Bender O, Ney H (2006) Morpho-syntactic Arabic preprocessing for Arabic to English statistical machine translation. In: Proceedings of the NAACL workshop on statistical machine translation. New York, NY, pp 15–22

Fellbaum C (1998) WordNet: an electronic lexical database. MIT Press, Cambridge, MA. http://www.cogsci.princeton.edu/~wn (2000, September 7)

Font-Llitjós A, Vogel S (2007) A walk on the other side: adding statistical components to a transfer-based translation system. In: Proceedings of the workshop on syntax and structure in statistical translation at the human language technology conference of the North American chapter of the association for computational linguistics. Rochester, NY, pp 72–79

Giménez J, Màrquez L (2007) Linguistic features for automatic evaluation of heterogenous MT systems. In: Proceedings of the second workshop on statistical machine translation. Prague, Czech Republic, pp 256–264

Goldwater S, McClosky D (2005) Improving statistical MT through morphological analysis. In: HLT/EMNLP 2005, proceedings of human language technology conference and conference on empirical methods in natural language processing. Vancouver, BC, Canada, pp 676–683

Graff D (1994) UN parallel text (Spanish-English). Linguistic Data Consortium Catalog No. LDC94T4A

Graff D (2003a) Arabic Gigaword. Linguistic Data Consortium Catalog No. LDC2003T12

Graff D (2003b) English Gigaword corpus. Linguistic Data Consortium Catalog No. LDC2003T05

Grimshaw J, Mester A (1988) Light verbs and theta-marking. Linguist Inq 19:205–232

Habash N (2000) oxyGen: a language independent linearization engine. In: AMTA-2000, fourth conference of the association for machine translation in the Americas: envisioning machine translation in the information future. Cuernavaca, Mexico, pp 68–79

Habash N (2003a) Generation heavy hybrid machine translation. Ph.D. thesis, University of Maryland, College Park, MD

Habash N (2003b) Matador: a large scale Spanish-English GHMT system. In: MT Summit IX, proceedings of the ninth machine translation summit. New Orleans, LA, pp 149–156

Habash N (2004) The use of a structural N-gram language model in generation-heavy hybrid machine translation. In: Belz A, Evans R, Piwek P (eds) Natural language generation, third international conference, INLG 2004. Springer-Verlag, Berlin, Heidelberg, NY, pp 61–69

Habash N (2007a) Arabic morphological representations for machine translation. In: van den Bosch A, Soudi A, Neumann G (eds) Arabic computational morphology: knowledge-based and empirical methods. Springer, Dordrecht, The Netherlands pp 263–285

Habash N (2007b) Syntactic preprocessing for statistical MT. In: Machine translation summit XI, proceedings. Copenhagen, Denmark, pp 215–222

Habash N, Dorr BJ (2002) Handling translation divergences: combining statistical and symbolic techniques in generation-heavy machine translation. In: Machine translation: from research to real users, 5th conference of the association for machine translation in the Americas, AMTA 2002, proceedings. Springer-Verlag, Berlin Heidelberg, New York, pp 84–93

Habash N, Dorr BJ (2003) A categorial variation database for English. In: HLT-NAACL: human language technology conference of the North American chapter of the association for computational linguistics, Vol. 1. Edmonton, AL, Canada, pp 96–102

Habash N, Elkholy A (2008) SEPIA: surface span extension to syntactic dependency precision-based MT evaluation. In: Proceedings of the NIST metrics for machine translation workshop at the association for machine translation in the Americas conference, AMTA-2008. Waikiki, HI

Habash N, Rambow O (2004) Extracting a tree adjoining grammar from the Penn Arabic Treebank. In: Proceedings of Traitement Automatique du Langage Naturel (TALN-04). pp 277–284. Fez, Morocco

Habash N, Rambow O (2005) Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In: 43rd annual meeting of the association for computational linguistics (ACL'05). Ann Arbor, MI, pp 573–580

Habash N, Sadat F (2006) Arabic preprocessing schemes for statistical machine translation. In: Proceedings of the 7th meeting of the North American chapter of the association for computational linguistics/human language technologies conference (HLT-NAACL06). New York, NY, pp 49–52

Habash N, Dorr BJ, Traum D (2003) Hybrid natural language generation from lexical conceptual structures. Mach Transl 18:81–127

Habash N, Soudi A, Buckwalter T (2007) On Arabic transliteration. In: van den Bosch A, Soudi A, Neumann G (eds) Arabic computational morphology: knowledge-based and empirical methods. Springer, Dordrecht, The Netherlands, pp 15–22

Han C, Lavoie B, Palmer M, Rambow O, Kittredge R, Korelsky T, Kim N, Kim M (2000) Handling structural divergences and recovering dropped arguments in a Korean/English machine translation system. In: AMTA-2000, fourth conference of the association for machine translation in the Americas: envisioning machine translation in the information future. Cuernavaca, Mexico, pp 40–53

Hwa R (2001) Const2Dep Tool. http://www.cs.cmu.edu/afs/cs/user/alavie/MTEval/code/hwc/const2dep/

Jackendoff R (1983) Semantics and cognition. The MIT Press, Cambridge, MA

Jackendoff R (1990) Semantic structures. The MIT Press, Cambridge, MA

Johnson M (2001) Joint and conditional estimation of tagging and parsing models. In: Association for computational linguistics, 39th annual meeting and 10th conference of the European chapter, proceedings of the conference. Toulouse, France, pp 314–321

Knight K, Hatzivassiloglou V (1995) Two-level, many-paths generation. In: 33rd annual meeting of the association for computational linguistics (ACL-95). Cambridge, MA, pp 252–260

Koehn P (2004a) Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In: Proceedings of the 6th biennial conference of the association for machine translation in the Americas. Washington, DC, pp 115–124

Koehn P (2004b) Statistical significance tests for machine translation evaluation. In: Proceedings of the 2004 conference on empirical methods in natural language processing conference. Barcelona, Spain, pp 388–395

Koehn P, Och F, Marcu D (2003) Statistical phrase-based translation. In: HLT-NAACL: human language technology conference of the North American chapter of the association for computational linguistics. Edmonton, AL, Canada, pp 127–133

Kulick S, Gabbard R, Marcus M (2006) Parsing the Arabic Treebank: analysis and improvements. In: Proceedings of the Treebanks and linguistic theories conference. Prague, Czech Republic, pp 31–42

Langkilde I (2000) Forest-based statistical sentence generation. In: 1st meeting of the North American chapter of the association for computational linguistics, proceedings. Seattle, WA, pp 170–177

Langkilde I, Knight K (1998a) Generating word lattices from abstract meaning representation. Technical report, Information Science Institute, University of Southern California, Marina del Rey, CA

Langkilde I, Knight K (1998b) Generation that exploits corpus-based statistical knowledge. In: COLING-ACL 98, 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics, proceedings of the conference. Montreal, QC, Canada, pp 704–710

Lavoie B, Kittredge R, Korelsky T, Rambow O (2000) A framework for MT and multilingual NLG systems based on uniform lexico-structural processing. In: 6th applied natural language processing conference, proceedings of the conference. Seattle, WA, pp 63–67

Lavoie B, White M, Korelsky T (2001) Inducing lexico-structural transfer rules from parsed bi-texts. In: Proceedings of the 39th annual meeting of the association for computational linguistics—DDMT workshop. Toulouse, France, pp 17–24

Lee Y-S (2004) Morphological analysis for statistical machine translation. In: Proceedings of the 5th meeting of the North American chapter of the association for computational linguistics/human language technologies conference (HLT-NAACL04). Boston, MA, pp 57–60

Levin B (1993) English verb classes and alternations: a preliminary investigation. University of Chicago Press, Chicago, IL

Liu D, Gildea D (2005) Syntactic features for evaluation of machine translation. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. Ann Arbor, MI, pp 25–32

Maamouri M, Bies A, Buckwalter T, Mekki W (2004) The Penn Arabic Treebank: building a large-scale annotated Arabic Corpus. In: NEMLAR conference on Arabic language resources and tools. Cairo, Egypt, pp 102–109

Macleod C, Grishman R, Meyers A, Barrett L, Reeves R(1998) NOMLEX: a lexicon of nominalizations. In: Proceedings of EURALEX'98. Liège, Belgium, pp 187–193

Marcus MP, Santorini B, Marcinkiewicz MA (1994) Building a large annotated Corpus of English: the Penn Treebank. Comput Linguist 19(2):313–330

Mel'čuk I (1988) Dependency syntax: theory and practice. State University of New York Press, Albany, NY

Nasr A, Rambow O (2006) Parsing with lexicalized probabilistic recursive transition networks. In: Yli-Jyrä A, Karttunen L, Karhumäki J (eds) Finite-state methods and natural language processing, vol 4002 of lecture notes in computer science. Springer-Verlag, Berlin/Heidelberg, pp 156–166

Nasr A, Rambow O, Palmer M, Rosenzweig J (1997) Enriching lexical transfer with cross-linguistic semantic features (or how to do interlingua without interlingua). In: Proceedings of the 2nd international workshop on interlingua. San Diego, CA

Nguyen TP, Shimazu A (2006) Improving phrase-based statistical machine translation with morphosyntactic transformation. Mach Transl 20(3):147–166

Nießen S, Ney H (2004) Statistical machine translation with scarce resources using morpho-syntactic information. Comput Linguist 30(2):181–204

Och FJ (2003) Minimum error rate training for statistical machine translation. In: 41st annual meeting of the association for computational linguistics. Sapporo, Japan, pp 160–167

Och FJ (2005) Google system description for the 2005 NIST MT evaluation. In: MT Eval workshop (unpublished talk)

Owczarzak K, van Genabith J, Way A (2007) Labelled dependencies in machine translation evaluation. In: Proceedings of the second workshop on statistical machine translation. Prague, Czech Republic, pp 104–111

Papineni K, Roukos S, Ward T, Zhu W-J (2002) BLEU: a method for automatic evaluation of machine translation. In: 40th annual meeting of the association for computational linguistics. Philadelphia, PA, pp 311–318

Popović M, Ney H (2004) Towards the use of word stems and suffixes for statistical machine translation. In: Proceedings of the 4th international conference on language resources and evaluation (LREC). Lisbon, Portugal, pp 1585–1588

Porter M (1980) An algorithm for suffix stripping. Program 14(3):130–137

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2002) Numerical recipes in C++. Cambridge University Press, Cambridge, UK

Quirk C, Menezes A, Cherry C (2005) Dependency treelet translation: syntactically informed phrasal SMT. In: 43rd annual meeting of the association for computational linguistics. Ann Arbor, MI, pp 271–279

Ratnaparkhi A (2000) Trainable methods for surface natural language generation. In: Proceedings of the 1st annual North American association of computational linguistics (NAACL-2000). Seattle, WA, pp 194–201

Resnik P (1997) Evaluating multilingual gisting of web pages. AAAI symposium on natural language processing for the world wide web, Stanford, CA

Resnik P, Olsen M, Diab M (1999) The bible as a parallel corpus: annotating the book of 2000 tongues. Comput Humanit 33:129–153

Riesa J, Yarowsky D (2006) Minimally supervised morphological segmentation with applications to machine translation. In: Proceedings of the 7th conference of the association for machine translation in the Americas: visions for the future of machine translation. Cambridge, MA, pp 185–192

Rogers W (2000) TREC Spanish corpus. Linguistic Data Consortium catalog no. LDC2000T51

Roth R, Rambow O, Habash N, Diab M, Rudin C (2008) Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In: 46th annual meeting of the association for computational linguistics: human language technologies, proceedings of the conference, short papers. Columbus, OH, pp 117–120

Sadat F, Habash N (2006) Combination of Arabic preprocessing schemes for statistical machine translation. In: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics. Sydney, Australia, pp 1–8

Senellart J (2006) Boosting linguistic rule-based MT system with corpus-based approaches. In: Presentation. GALE PI Meeting. Boston, MA

Sharaf M (2002) Implications of the agreement features in (English to Arabic) machine translation. Master's thesis, Al-Azhar University, Cairo, Egypt

Sima'an K (2000) Tree-gram parsing: lexical dependencies and structural relations. In: 38th annual meeting of the association for computational linguistics (ACL'00). Hong Kong, China, pp 37–44

Snover M, Dorr BJ, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation error rate with targeted human annotation. In: Proceedings of the 7th conference of the association for machine translation in the Americas: visions for the future of machine translation. Cambridge, MA, pp 223–231

Soudi A (2004) Challenges in the generation of Arabic from interlingua. In: Proceedings of Traitement Automatique des Langues Naturelles (TALN-04). Fez, Morocco, pp 343–350

Soudi A, Cavalli-Sforza V, Jamari A (2002) A prototype English-to-Arabic interlingua-based MT system. In: Proceedings of the third international conference on language resources and evaluation: workshop on Arabic language resources and evaluation: status and prospects. Las Palmas de Gran Canaria, Spain, pp 18–25

Stolcke A. (2002) SRILM—an extensible language modeling toolkit. In: Proceedings of the international conference on spoken language processing (ICSLP), vol 2. Denver, CO, pp 901–904

Tanaka T, Baldwin T (2003) Translation selection for Japanese–English noun-noun compounds. In: MT Summit IX, proceedings of the ninth machine translation summit. New Orleans, LA, pp 378–385

Tapanainen P, Jarvinen T (1997) A non-projective dependency parser. In: Proceedings of the 5th conference on applied natural language pro cessing. Washington, DC, pp 64–71

Traum D, Habash N (2000) Generation from lexical conceptual structures. In: Proceedings of the workshop on applied interlinguas, North American association of computational linguistics/applied natural language processing conference, NAACL/ANLP-2000. Seattle, WA, pp 34–41

Vauquois B (1968) A survey of formal grammars and algorithms for recognition and transformation in machine translation. In: IFIP congress-68. Edinburgh, UK, pp 254–260

Watanabe H, Kurohashi S, Aramaki E (2000) Finding structural correspondences from bilingual parsed corpus for corpus-based translation. In: Proceedings of the 18th international conference on computational linguistics, vol 2. Saarbrücken, Germany, pp 906–912

Whitelock P (1992) Shake-and-bake translation. In: Proceedings of fifteenth [sic] international conference on computational linguistics. Nantes, France, pp 784–791

Xia F, McCord M (2004) Improving a statistical MT system with automatically learned rewrite patterns. In: Proceedings of the 20th international conference on computational linguistics (COLING 2004). Geneva, Switzerland, pp 508–514

Zhang Y, Zens R, Ney H (2007) Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In: Proceedings of the workshop on syntax and structure in statistical translation at the human language technology conference of the North American chapter of the association for computational linguistics. Rochester, NY, pp 1–8

Zollmann A, Venugopal A, Vogel S (2006) Bridging the inflection morphology gap for Arabic statistical machine translation. In: Proceedings of the human language technology conference of the NAACL, companion volume: short papers. New York, NY, pp 201–204