A Two-Level Syntax-Based Approach to Arabic-English Statistical Machine Translation

Charles Schafer and David Yarowsky
Center for Language and Speech Processing / Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218 USA
{*cschafer,yarowsky*}*@cs.jhu.edu*

Abstract

We formulate an original model for statistical machine translation (SMT) inspired by characteristics of the Arabic-English translation task. Our approach incorporates part-of-speech tags and linguistically motivated phrase chunks in a 2-level shallow syntactic model of reordering. We implement and evaluate this model, showing it to have advantageous properties and to be competitive with an existing SMT baseline. We also describe cross-categorial lexical translation coercion, an interesting component and side-effect of our approach. Finally, we discuss the novel implementation of decoding for this model which saves much development work by constructing finite-state machine (FSM) representations of translation probability distributions and using generic FSM operations for search. Algorithmic details, examples and results focus on Arabic, and the paper includes discussion on the issues and challenges of Arabic statistical machine translation.

1 Introduction

In this work we define, implement and evaluate a novel model for statistical machine translation (SMT), which is motivated by considerations of Arabic syntactic ordering as they affect Arabic-to-English translation.

Our goal was to produce a SMT system for translating foreign languages, and Arabic in particular, into English by utilizing some information about syntax in both the foreign language and English without, however, requiring a full parse in either language. Some advantages of not relying on full parses include that (1) there is a lack of availability of parsers for many languages of interest; (2) parsing time complexity represents a potential bottleneck for both model training and testing.

Intuitively, the explicit modeling of syntactic phenomena should be of benefit in the machine translation task; the ability to handle long-distance motion in an intelligently constrained way is a salient example of this. Allowing unconstrained translation reorderings at the word level generates a very large set of permutations, creating a difficult search problem at decode time. We propose a model that makes use of shallow parses (text chunking) to allow long-distance motion of phrases while ignoring deeper issues of syntax. The resources required to train this system on a new language are minimal, and we gain the ability to model long-distance movement as well as some interesting properties of lexical translation across parts of speech. Arabic has a canonical sentence-level order of Verb-Subject-Object, which means that translation into English (with a standard ordering of Subject-Verb-Object) commonly requires motion of entire phrasal constituents, which is not true of French-to-English translation, to cite one language pair whose characteristics have wielded great influence in the history of work on statistical machine translation. A key motivation for and objective of this work was to build a translation model and feature space to effectively handle the above-described phenomenon.

2 Prior Work

Statistical machine translation, as pioneered by IBM (e.g. Brown et al., 1993), is grounded in the noisy channel model. And similar to the related channel problems of speech and handwriting recognition, the original SMT language pair French-English exhibits a relatively close linear correlation in source and target sequence. Most common non-sequential motion that is observed, in terms of adjective-noun swapping, is well modeled by the relative-position-based distortion models of the classic IBM approach. Unfortunately, these distortion models are less effective for languages such as Japanese or Arabic, which have substantially different top-level sentential word orders from English.

Wu (1997) and Jones and Havrilla (1998) have sought to more closely tie the allowed motion of constituents between languages to those syntactic transductions supported by the independent rota-

tion of parse tree constituents. Yamada and Knight (2000, 2001) and Alshawi et al. (2000) have effectively extended such syntactic transduction models to fully functional SMT systems, based on channel model tree transducers and finite state head transducers respectively. While these models are well suited for the effective handling of highly divergent sentential word orders, the above frameworks have a limitation shared with probabilistic context free grammars that the preferred ordering of subtrees is insufficiently constrained by their embedding context, which is especially problematic for very deep syntactic parses.

In contrast, Och et al. (1999) have avoided the constraints of tree-based syntactic models and allow the relatively flat motion of empirically derived phrasal chunks, which need not adhere to traditional constituent boundaries.

Our current paper takes a middle path, by grounding motion in syntactic transduction, but in a much flatter 2-level model of syntactic analysis, based on flat embedded noun-phrases in a flat sentential constituent-based chunk sequence that can be driven by syntactic bracketers and POS tag models rather than a full parser, facilitating its transfer to lower density languages. The flatter 2-level structures also better support transductions conditioned to full sentential context than do deeply embedded tree models, while retaining the empirically observed advantages of translation ordering independence of noun-phrases.

Another improvement over Och et al. and Yamada and Knight is the use of the finite state machine (FSM) modelling framework (e.g. Bangalore and Riccardi, 2000), which offers the considerable advantage of a flexible framework for decoding, as well as a representation which is suitable for the fixed two-level phrasal modelling employed here.

Finally, the original cross-part-of-speech lexical coercion models presented in Section 4.3.3 have related work in the primarily-syntactic coercion models utilized by Dorr and Habash (2002) and Habash and Dorr (2003), although their induction and modelling are quite different from the approach here.

3 Resources

As in other SMT approaches, the primary training resource is a sentence-aligned parallel bilingual corpus. We further require that each side of the corpus be part-of-speech (POS) tagged and phrase chunked. Our translation experiments were carried out using the United Nations Arabic-English parallel corpus made available (with sentence alignments) by the Linguistic Data Consortium.

POS tagging and phrase chunking in English were done using the trained systems provided with the fnTBL Toolkit (Ngai and Florian, 2001); both were trained from the annotated Penn Treebank corpus (Marcus et al., 1993). For Arabic, we used a colleague's POS tagger and tokenizer (clitic separation was also performed prior to POS tagging), which was rapidly developed in our laboratory. Phrase segmentation was achieved via a simple decision list of chunk join/split decisions, based on variable-length right and left context patterns, as illustrated in Table 1. The highest-ranked matching pattern was used at each decision point (between any two contiguous words we consider there to be a decision point, at which a binary decision must be made between *split* and *join*). Each such segmentation decision was made in isolation, that is, with no global maximization.

Context Rules		Context Rules		Context Rules
N+DN		D+N		MN+N
N|DA		R|N		N+NNN
N|D		N+N		N+NND
DN|D	. . .	DN|N	. . .	N+NNA
NN|D		AN+N		N+NN
A|N		RN+N		VN+A

Table 1: A small sample of the phrase segmentation patterns used. | means insert a phrase chunk boundary at this point. + mean join left and right context into a single phrase at this point. N,D,A,V etc. refer to Arabic core parts of speech.

A further input to our system is a set of word alignment links on the parallel corpus. These are used to compute word translation probabilities and phrasal alignments. The word alignments can in principle come from any source: a dictionary, a specialized alignment program, or another SMT system. We used alignments generated by Giza++ (Och and Ney, 2000) by running it in both directions on our parallel corpus. The union of these bidirectional alignments was used to compute cross-language phrase correspondences (alignments) by simple plurality voting. Specifically, each Arabic phrase in the training corpus was allowed to vote on the single English phrase to which it was most strongly aligned. Each word alignment link between a token in the Arabic phrase and a token in an English phrase was counted as a unit vote. Ties among English phrases with equal scores were broken by taking the leftmost such English phrase. The

resulting phrasal alignments were taken as hard decisions, and while individual word alignment links violating the induced phrase alignments were still used in calculating word translation probabilities, they were ignored with respect to the alignment model[1]. For purposes of estimating word translation probabilities, each link in the union of word alignments was treated as an independent instance of word translation.

4 Translation Model

Now we turn to a detailed description of the proposed translation model. The exposition will give a formal specification and also will follow a running example throughout, using one of the actual Arabic test set sentences. This example, its gloss, system translation and reference human translation are shown in Table 3.

The translation model (TM) we describe is trained directly from counts in the data, and is a direct model, not a noisy channel model. It consists of three nested components: (1) a sentence-level model of phrase correspondence and reordering, (2) a model of intra-phrase translation, and (3) models of lexical transfer, or word translation. We make a key assumption in our construction that translation at each of these three levels is independent of the others.

4.1 Sentence Translation

As mentioned, both the foreign language and English corpora are input with "hard" phrase bracketings and labeled with "hard" phrase types (e.g., NP, VP[2], PPNP[3], etc.). These are denoted in the top-level model presentation in Table 4 (1). Given word alignment links, as described in Section 3, we compute phrasal alignments on training data. We contrain these to have cardinality (*foreign*)$N \leftrightarrow 1$(*English*). Next, we collect counts over aligned phrase sequences and use the relative frequencies to estimate the probability distribution in Table 4 (2). Particularly for smaller training corpora, unseen foreign-language phrase sequences are a problem, so we implemented a simple backoff method which assigns probability to translations of unseen foreign-language phrase sequences[4].

Table 4 (3) encapsulates the remainder of the translation model, which is described below.

As an example, see Table 2 for the most probable aligned English phrase sequence generations given an Arabic simple sentence having the canonical VSO ordering.

Arabic Phrase Sequence	Aligned English Phrase Sequence	Prob.
VP_1 NP_2 NP_3	NP_2 VP_1 NP_3	0.23
VP_1 NP_2 NP_3	VP_1 NP_2 PP_3	0.10
VP_1 NP_2 NP_3	NP_3 $VP_{1,2}$	0.06

Table 2: Top learned sentence-level reorderings for Arabic, for canonical Arabic simple sentence structure VP (verb) NP (subject) NP (object). Subscripts in English phrase sequence are alignments to positions in Arabic phrase sequence.

4.2 Phrase Translation

Given an Arabic test sentence, a distribution of aligned English phrase sequences is proposed by the sentence-level model described in the previous section and in Table 4. Each proposed English phrase in each of the phrase sequence possibilities, therefore, comes to the phrase translation level of the model with access to the identity of the Arabic phrase(s) aligned to it. Phrase translation is implemented as shown in Table 8. The phrase translation model is structured with several levels of backoff: if no observations exist from training data for a particular level, the model backs off to the next-more-general level. In all cases, generation of an English phrase is conditioned on the foreign phrase as well as the type (NP, VP, etc.) of the English phrase.

Table 8 (1) describes the initial phrase translation model. It comes into play if the precise sequence of foreign words has been observed aligning to an English phrase of the appropriate type. In the example, we are trying to generate an NP given the Arabic word string "*Al- ljnp Al- sAdsp*" (literally: "the committee the sixth"). If this has been observed in data, then that relative frequency distribution serves as the translation probability distribution. The following table (Table 5) contains examples of some

[1]Note that the described phrasal alignment procedure results in an (Arabic↔English) N↔1 cardinality of phrasal correspondences. This is an attribute of the current implementation, but there is no inherent requirement to respect this particular cardinality. One avenue of future enhancement will be to explore modifying or eliminating this constraint.

[2]VP in our parlance is perhaps more properly called a verb chunk: it consists of a verb, its auxiliaries, and contiguous adverbs.

[3]PPNP consists of a NP with its prepositional head attached.

[4]Using heuristics, Arabic phrase chunks and English phrase chunks were clumped into segments (e.g., a segment might be NP PPNP). Arabic segment to aligned English segment translation probabilities were estimated from counts (e.g., NP_1 $PPNP_2 \rightarrow NP_2$ NP_1 with probability 0.1). A sentence-level segment reordering probability distribution was estimated separately.

Arabic Example Sentence From Test Set	
ARABIC:	*twSy Al- ljnp Al- sAdsp Al- jmEyp Al- EAmp b- AEtmAd m$rwE Al- mqrr Al- tAly :*
BRACKETED ARABIC:	*[twSy] [Al- ljnp Al- sAdsp] [Al- jmEyp Al- EAmp] [b- AEtmAd m$rwE Al- mqrr Al- tAly] [:]*
GLOSS:	[recommends] [the committee the sixth] [the assembly the general] [to adoption draft the decision the following] [:]
MT OUTPUT:	[the sixth committee] [recommends] [the general assembly] [in the adoption of the following draft resolution] [:]
REFERENCE TRANS.:	the sixth committee recommends to the general assembly the adoption of the following draft decision :

Table 3: An Arabic translation from the test set. We revisit portions of this example throughout the text. All Arabic strings in this paper are rendered in the reversible Buckwalter transliteration. In addition, all words or symbols referring to Arabic are italicized.

Top-level Definition of Translation Model		
Example Instantiation of Model Variables		Model Description
P(the sixth committee recommends the general assembly .. | *twSy Al- ljnp Al- sAdsp Al- jmEyp Al- EAmp ..*) =		P(english_words | *foreign_words*) =
P([*twSy*]$_{VP_1}$ [*Al- ljnp Al- sAdsp*]$_{NP_2}$ [*Al- jmEyp Al- EAmp*]$_{NP_3}$ .. | *twSy Al- ljnp Al- sAdsp Al- jmEyp Al- EAmp ..*)	(1)	P(*foreign_bracketing* , *foreign_phrase_sequence* | *foreign_words*)
$*P$(NP_2 VP_1 NP_3 $PPNP_4$ $PUNC_5$ | *VP*$_1$ *NP*$_2$ *NP*$_3$ *PPNP*$_4$ *PUNC*$_5$)	(2)	P(english_phrase_sequence , phrase_alignment_matrix | *foreign_phrase_sequence*)
$*P$([the sixth committee]$_{NP_2}$ [recommends]$_{VP_1}$ [the general assembly]$_{NP_3}$ .. | [*twSy*]$_{VP_1}$ [*Al- ljnp Al- sAdsp*]$_{NP_1}$ [*Al- jmEyp Al- EAmp*]$_{NP_2}$ .. , NP_2 VP_1 NP_3 $PPNP_4$ $PUNC_5$)	(3)	P(english_words , english_bracketing , english_phrase_sequence | *foreign_words* , *foreign_bracketing* , *foreign_phrase_sequence* , english_phrase_sequence , phrase_alignment_matrix)

Table 4: Statement of the translation model at top level.

of these literal phrase translations from the Arabic data.

Type	Arabic Phrase	English Phrase	Prob.
NP	Al- AtfAq	the agreement	0.593
NP	Al- AtfAq	agreement	0.268
NP	Al- AtfAq	an agreement	0.041
NP	Al- AtfAq	the compact	0.031
NP	Al- AtfAq	this agreement	0.010
NP	Al- AtfAq	the form of the agreement	0.010
NP	Al- AtfAq	the accord	0.010
NP	Al- AtfAq	the largest agreement	0.005
NP	Al- AtfAq	the standard agreement	0.005
PPNP	Al- AtfAq	in the agreement	0.313
PPNP	Al- AtfAq	by the agreement	0.313
PPNP	Al- AtfAq	with the agreement	0.187
PPNP	Al- AtfAq	before the compact	0.063
PPNP	Al- AtfAq	to the accord	0.063
PPNP	Al- AtfAq	to agreement	0.063
VP	Al- AtfAq	agree	0.321
VP	Al- AtfAq	to agree	0.226
VP	Al- AtfAq	agreed	0.094
VP	Al- AtfAq	agreeing	0.057
VP	Al- AtfAq	could agree	0.019
VP	Al- AtfAq	be agreed	0.019
VP	Al- AtfAq	establishes	0.019
VP	Al- AtfAq	is understood	0.019
VP	Al- AtfAq	cannot agree	0.019
VP	Al- AtfAq	will have to be agreed	0.019
VP	Al- AtfAq	are agreed	0.019

Table 5: Literal phrase translations learned by the system, including some coercions across phrase type (NP → NP, PPNP, VP). Translation probability of the English phrase is conditioned on the English phrase type and the Arabic phrase. Examples are all for the Arabic phrase *Al- AtfAq* ("the agreement").

The next stage of backoff from the above, literal level is a model that generates aligned English POS tag sequences given foreign POS tag sequences: details and an example can be found in Table 8 (2). The sequence alignments determine the position in English phrase and the part-of-speech into which we translate the foreign word. Again, translation is also conditioned on the English phrase type. See Figure 1 for the most probable aligned English sequence generations for two of the phrases in the example sentence.

If there were no counts for (foreign-POS-sequence, english-phrase-type) then we back off to counts collected over (foreign-coarse-POS-equence, english-phrase-type), where a coarse POS is, for example, *N* instead of *NOUN-SG*. This is shown in Table 8 (3).

In case further backoff is needed, as shown in Table 8 (4), we begin stripping POS-tags off the "less significant" (non-head) end of the foreign POS-sequence until we are left with a phrase sequence that has been seen in training, and from this a corresponding English phrase distribution is observable. We define the "less significant" end of a phrase to be the end if it is head-initial, or the beginning if it is head-final, and at this point ignore issues such as nested structure in Arabic NP's.

Finally, we should note here that word generation from NULL alignments is allowed in some cases. As a practical matter, some phrases observed

in training data are so deficient in word-alignment links (due to the noisy and incomplete word alignments available) that they must be discarded with heuristics from training the POS-sequence alignments. For example, it doesn't make much sense to generate an English phrase with 4 nouns from an Arabic phrase with 4 nouns, with only one word alignment link between a single Arabic-English noun pair. However, we take phrase pairs with unaligned English determiners, prepositions, modals, etc. (essentially closed-class words) and allow generation from a list of such possible NULL generations based only on P(english-word | english-POS).

Phrase Translation Examples

P(*aligned Eng. POS sequence* | DET$_1$ NOUN-SG$_2$ DET$_3$ ADJ$_4$,NP)

.22	DT$_0$ JJ$_4$ NN$_2$
.20	JJ$_4$ NN$_2$
.13	DT$_0$ NN$_2$
.13	DT$_0$ VBN$_4$ NNS$_2$
	⋮
.02	NN$_4$ NN$_2$

P(*aligned Eng. POS sequence* | VERB-IMP$_1$,VP)

.28	VBZ$_1$
.17	VBP$_1$
.09	VBD$_1$
	⋮
.06	MD$_0$ VB$_1$

Figure 1: From the running Arabic example, (1) top English NP generations given an Arabic phrase *DET NOUN-SG DET ADJ*; (2) top English VP generations given an Arabic phrase *VERB-IMP*. Note: 0 denotes a null alignment (generation from null). Generation from a null alignment is allowed for specified parts of speech, such as determiners and prepositions.

4.3 Lexical Transfer

4.3.1 Word Translation Model

In the word generation model, phrases may be translated directly as single atomic entities (as in Table 8 (1)), or via phrasal decomposition to individual words translated independently, conditioned only on the source word and target POS. Word translation is done in the context that the model has already proposed a sequence of POS tags for the phrase. Thus we know the English POS of the word we are trying to generate in addition to the foreign word that is generating it. Consequently, we condition translation on English POS as well as the foreign word. Table 6 describes the backoff path for basic lexical transfer and presents a motivating example in the Arabic word *mrdwd*. Additionally, translation probabilities for one of the words in the example Arabic sentence can be found in Table 7.

Word Generation	
Examples	Model with Backoff Pathways
$P(W_E \mid mrdwd, NNS)$	$P(W_E \mid W_F, T_{fine_E})$
returns 0.43	$p(returns \mid mrdwd, NNS)$
wages 0.14	
rewards 0.07	
proceeds 0.07	
	↓ (backoff if $C(W_F, T_{fine_E}) = 0$)
$P(W_E \mid mrdwd, N)$	$P(W_E \mid W_F, T_{coarse_E})$
return 0.27	
returns 0.16	$p(returns \mid mrdwd, N)$
wages 0.05	
benefit 0.03	
	↓ (backoff if $C(W_F, T_{coarse_E}) = 0$)
$P(W_E \mid mrdwd)$	$P(W_E \mid W_F)$
return 0.14	
returns 0.08	$p(returns \mid mrdwd)$
financial 0.06	
yield 0.04	
	↓ (backoff if $C(W_F) = 0$)
	$p($ *UNKNOWN_WORD* $\mid W_F) = 1$

Table 6: Description of the conditioning for different levels of backoff in the lexical transfer model. The example shows translations for the Arabic word *mrdwd* (financial meaning of "revenue/return on investment") conditioned on decreasingly specific values. The progressively lower probability and ranking of the desired plural noun translation as we move from fine, to coarse, to no POS, illustrates the benefit of conditioning generation on the English part of speech.

Translation Probabilities for *"ljnp"*			
Arabic Word	**English POS**	**English Word**	**Prob.**
ljnp	NN	committee	0.591
ljnp	NN	commission	0.233
ljnp	NN	subcommittee	0.035
ljnp	NN	acc	0.013
ljnp	NN	report	0.005
ljnp	NN	ece	0.004
ljnp	NN	icrc	0.004
ljnp	NN	aalcc	0.004
ljnp	NN	escap	0.004
ljnp	NN	escwa	0.004
ljnp	NN	eca	0.003
ljnp	NNS	members	0.088
ljnp	NNS	recommendations	0.033
ljnp	NNS	copuos	0.033
ljnp	NNS	representatives	0.024
...			
ljnp	NNS	commissions	0.008
Arabic Word	**English Coarse POS**	**English Word**	**Prob.**
ljnp	N	committee	0.577
ljnp	N	commission	0.227
ljnp	N	subcommittee	0.035

Table 7: From running example, translation probabilities for Arabic noun *ljnp*, "committee".

Phrase Translation Model with Backoff Pathways

Example Instantiations		Model Statement
$P($ the sixth committee | *Al- ljnp Al- sAdsp* , NP $) =$		
$P($ the sixth committee | *Al- ljnp Al- sAdsp* , NP $)$	(1)	$P(\mathrm{W}_{E_1}\ \mathrm{W}_{E_2}\ ..\ \mathrm{W}_{E_n}\ \|\ W_{F_1}\ W_{F_2}\ ..\ W_{F_m}\ ,\ \mathrm{phr\_type}_E)$
$\downarrow$		$\downarrow$ (backoff if $C(\ W_{F_1}\ W_{F_2}\ ..\ W_{F_m}\ ,\ \mathrm{phr\_type}_E) = 0$)
$P($ DT$_1$ JJ$_4$ NN$_2$ | *DET$_1$ NOUN-SG$_2$ DET$_3$ ADJ$_4$* , NP $)$	(2)	$P(\mathrm{T}_{fine_{E_1}}\ \mathrm{T}_{fine_{E_2}}\ ..\ \mathrm{T}_{fine_{E_n}}\ ,\ \Xi_i\ \|\ T_{fine_{F_1}}\ T_{fine_{F_2}}\ ..\ T_{fine_{F_m}}\ ,\ \mathrm{phr\_type}_E)$
$*P($ the | *Al-* , DT $)$		$*P(\mathrm{W}_{E_1}\ \|\ W_{F_{\Xi_i(1)}}\ ,\ T_{fine_{E_1}})$
$*P($ committee | *ljnp* , NN $)$		$*P(\mathrm{W}_{E_2}\ \|\ W_{F_{\Xi_i(2)}}\ ,\ \mathrm{T}_{fine_{E_2}})$
$*P($ sixth | *sAdsp* , JJ $)$		$*..*P(\mathrm{W}_{E_n}\ \|\ \mathrm{W}_{F_{\Xi_i(n)}}\ ,\ \mathrm{T}_{fine_{E_n}})$
$\downarrow$		$\downarrow$ (backoff if $C(\ T_{fine_{F_1}}\ T_{fine_{F_2}}\ ..\ T_{fine_{F_m}}\ ,\ \mathrm{phr\_type}_E) = 0$)
$P($ DT$_1$ JJ$_4$ NN$_2$ | $D_1\ N_2\ D_3\ A_4$, NP $)$	(3)	$P(\mathrm{T}_{fine_{E_1}}\ \mathrm{T}_{fine_{E_2}}\ ..\ \mathrm{T}_{fine_{E_n}}\ ,\ \Xi_i\ \|\ T_{coarse_{F_1}}\ T_{coarse_{F_2}}\ ..\ T_{coarse_{F_m}}\ ,\ \mathrm{phr\_type}_E)$
$*P($ the | *Al-* , DT $)$		$*P(\mathrm{W}_{E_1}\ \|\ W_{F_{\Xi_i(1)}}\ ,\ \mathrm{T}_{fine_{E_1}})$
$*P($ committee | *ljnp* , NN $)$		$*P(\mathrm{W}_{E_2}\ \|\ W_{F_{\Xi_i(2)}}\ ,\ \mathrm{T}_{fine_{E_2}})$
$*P($ sixth | *sAdsp* , JJ $)$		$*..*P(\mathrm{W}_{E_n}\ \|\ W_{F_{\Xi_i(n)}}\ ,\ \mathrm{T}_{fine_{E_n}})$
$\downarrow$		$\downarrow$ (backoff if $C(\ T_{coarse_{F_1}}\ T_{coarse_{F_2}}\ ..\ T_{coarse_{F_m}}\ ,\ \mathrm{phr\_type}_E) = 0$)
$P($? | $D_1\ N_2\ D_3$, NP $)$	(4)	$P(\mathrm{T}_{fine_{E_1}}\ \mathrm{T}_{fine_{E_2}}\ ..\ \mathrm{T}_{fine_{E_n}}\ ,\ \Xi_i\ \|\ T_{coarse_{F_1}}\ T_{coarse_{F_2}}\ ..\ T_{coarse_{F_{m-1}}}\ ,\ \mathrm{phr\_type}_E)$
$*$ ...		$*$ ...
$\downarrow$		$\downarrow$ (backoff if $C(\ T_{coarse_{F_1}}\ T_{coarse_{F_2}}\ ..\ T_{coarse_{F_{m-1}}}\ ,\ \mathrm{phr\_type}_E) = 0$)
$P($? | $D_1\ N_2$, NP $)$	(4)	$P(\mathrm{T}_{fine_{E_1}}\ \mathrm{T}_{fine_{E_2}}\ ..\ \mathrm{T}_{fine_{E_n}}\ ,\ \Xi_i\ \|\ T_{coarse_{F_1}}\ T_{coarse_{F_2}}\ ..\ T_{coarse_{F_{m-2}}}\ ,\ \mathrm{phr\_type}_E)$
$*$ ...		$*$ ...
$\downarrow$		$\downarrow$ (backoff if $C(\ T_{coarse_{F_1}}\ T_{coarse_{F_2}}\ ..\ T_{coarse_{F_{m-2}}}\ ,\ \mathrm{phr\_type}_E) = 0$)
....		....

Where:
W_{E_r} is the rth word in the English phrase ;
W_{F_r} is the rth word in the Arabic phrase ;
$\mathrm{phr\_type}_E$ is the type (NP,VP,...) of the English phrase ;
$C()$ represents the occurrence count in training data of the phenomenon at hand ;
$\mathrm{T}_{fine_{E_r}}$ is the fine-grained POS-tag of the rth word in the English phrase ;
$\mathrm{T}_{fine_{F_r}}$ is the fine-grained POS-tag of the rth word in the Arabic phrase ;
$\mathrm{T}_{coarse_{E_r}}$ and $\mathrm{T}_{coarse_{F_r}}$ have the corresponding meaning, for coarse-grained POS-tags ;
Ξ_i is the alignment matrix, representing positions in the Arabic phrase aligning to positions in the English phrase, for the pair i of aligned Arabic and English POS-tag sequences ;
$\Xi_i(r)$ is a function taking the position r in the English phrase and returning the position in the Arabic phrase to which it aligns ;
finally, "* ..." indicates a product of word translation probabilities which was omitted from the figure for space reasons.

Table 8: The phrase translation model, with backoff. Examples on the left side are from one of the Arabic test sentences. (1) is the direct, lexical translation level. (2) - (4) constitute the backoff path to handle detailed phenomena unseen in the training set. (2) is a model of fine POS-tag reordering and lexical generation; (3) is similar, but conditions generation on *coarse* POS-tag sequences in the foreign language. (4) is a model for progressively stripping off POS-tags from the "less significant" end of a foreign sequence. The idea is to do this until we reach a subsequence that has been seen in training data, and which we therefore have a distribution of valid generatons for. The term Ξ_i in (2) - (4) is a position alignment matrix. At all times, we generate not just an English POS-tag sequence, but rather an **aligned** sequence. Similarly, in the lexical transfer probabilities shown in this table, there is a function $\Xi_i()$ which takes an English sequence position index and returns the (unique) foreign word position to which it is aligned. At present, the model allows $1 \leftrightarrow \mathrm{N}$ cardinalities (Arabic $\leftrightarrow$ English) for word generation.

4.3.2 Lexical Coercion

Lexical coercion is a phenomenon that sometimes occurs when we condition translation of a foreign word on the word and the target (English) part-of-speech. We find that the system we have described frequently learns this behavior: specifically, the model learns in some cases how to generate e.g. a nominal form with similar meaning from an Arabic adjective, or an adjectival realization of an Arabic verb's meaning. Note the examples in Table 9. We find the coercion effect to be of note because it turns up interesting associations of meaning. For example, referring to the table, "yield" is a sensible way to realize the meaning of the word *mrdwd* (revenue/return on investment) in an active, verbal form. Similarly for *hdfA* (goal/objective/target). The system learned to coerce the nominal idea of an "objective" into the verb forms "undertaking" and "targeted".

Ar. Wd.	Eng. POS	Eng. Wd.
mrdwd	NN	return
mrdwd	VB	yield
hdfA	NN	objective
hdfA	VBN	targeted
hdfA	VBG	undertaking
<tlAf	NN	destruction
<tlAf	VBN	destroyed
<tlAf	VBG	vandalizing

Table 9: Examples of learned lexical coercion across parts of speech. Each example is the top-ranked choice of $P(W_E|W_F, T_{fine_E})$. <tlAf means "destruction"; refer to Section 4.3.2 for the other definitions.

5 Decoding

Decoding was implemented by constructing weighted finite-state machines (FSMs) **per evaluation sentence** to encode relevant portions (for the individual sentence in question) of the component translation distributions described above. Operations on these FSMs are performed using the AT&T FSM Toolkit (Mohri et al., 1997). The FSM constructed for a test sentence is subsequently composed with a FSM trigram language model created via the SRI Language Modeling Toolkit (Stolcke, 2002). Thus we use the trigram language model to implement rescoring of the (direct) translation probabilities for the English word sequences in the translation model lattice.

We found that using the finite-state framework and the general-purpose AT&T toolkit greatly facilitates decoder development by freeing the implementation from details of machine composition and best-path searching, etc.

The structure of the translation model finite-state machines is as illustrated in Figure 2. The sentence-level (aligned phrase sequence generation) and phrase-level (aligned intra-phrase sequence generation) reodering probabilities are encoded on epsilon arcs in the machines. Word translation probabilities are placed onto arcs emitting the word as an output symbol (in the figure, note the arcs emitting "committee", "the", etc.). The FSM in Figure 2 corresponds to the Arabic example sentence used throughout this paper. In the portion of the machine shown, the (best) path which generated the example sentence is drawn in bold. Finally, Figure 3 is a rendering of the actual FSM (aggressively pruned for display purposes) that generated the example Arabic sentence; although labels and details are not visible, it may provide a visual aid for better understanding the structure of the FSM lattices generated here.

As a practical matter in decoding, during translation model FSM construction we modified arc costs for output words in the following way: a fixed bonus was assigned for generating a "content" word translating to a "content" word. Determining what qualifies as a content word was done on the basis of a list of content POS tags for each language. For example, all types of nouns, verbs and adjectives were listed as content tags; determiners, prepositions, and most other closed-class parts of speech were not. This implements a reasonable penalty on undesirable output sentence lengths. Without such a penalty, translation outputs tend to be very short: long sentence hypotheses are penalized *de facto* merely by containing many word translation probabilities. An additional trick in decoding is to use only the N-best translation options for sentence-level, phrase-level, and word-level translation. We found empirically (and very consistently) in devtest experiments that restricting the syntactic transductions to a 30-best list and word translations to a 15-best list had no negative impact on Bleu score. The benefit, of course, is that the translation lattices are dramatically reduced in size, speeding up composition and search operations.

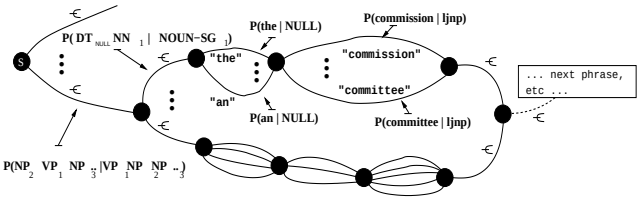

Figure 2: An illustration of the translation model structure for an Arabic test sentence. (a) The arcs immediately exiting the start state correspond to different sentence-level reordering possibilities. These arcs have, as attached weights, the probability of the particular sentence-level reordering designated. (b) Other arcs in the machine correspond to phrase-level reorderings, as shown in the figure. For each of the Arabic phrases in the test sentence, there will be a distribution over possible English reorderings / POS-tag sequence generations, and arcs in the machine corresponding to different reordering/generation choices, with associated probabilities. (c) Finally, word translation is also represented by arcs in the machine. These are not epsilon arcs, but rather arcs that emit the particular English word translation in question, and which have the appropriate word-to-word translation probability attached to the arc.

6 Evaluation

Table 10 below lists evaluation results for translation on the Arabic test set. Results for a compari-

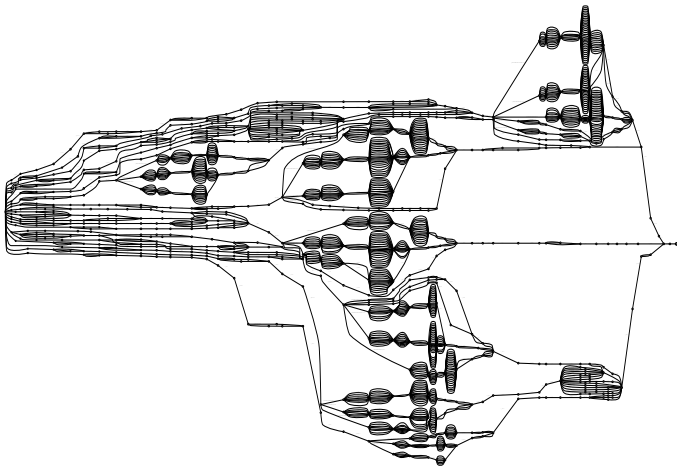

Figure 3: A portion of the translation model for an Arabic test sentence, aggressively pruned by path probability (pruning was performed for display purposes only: search for decoding was performed on **unpruned** FSMs). This is presented to further illustrate the structure of the FSMs used for decoding.

son system – the Giza++ IBM Model 4 implementation (Och and Ney, 2000) with the ReWrite decoder (Marcu and Germann, 2002) – are included as a baseline. For the Arabic UN corpus, we trained our system on a large subset of the UN corpus and evaluated on a 200-sentence held-out set. For this 150K sentence Arabic training set, Giza++ and the shallow syntax model achieved very similar performance. Results are scored via the Bleu metric proposed by Papineni et al. (2001).

	Bleu Score
System	150K Trn. Sent.
Giza++/ReWrite Decoder	0.17
2-level Syntax Model	0.17

Table 10: Results comparison for Arabic-English translation on UN corpus. (200-sentence evaluation set)

7 Conclusions

This paper has presented an original model for statistical machine translation inspired by and tailored to the syntactic divergences and other characteristics of Arabic-English statistical machine translation. The two-level syntactic transduction model supports both sentence-level and intra-phrase structural reordering, as well as a word translation component which benefits from empirically induced cross-part-of-speech lexical coercion. Current performance of this original full SMT model matches that of an existing, widely utilized SMT baseline approach.

References

H. Alshawi, S. Bangalore, and S. Douglas. 2000. Learning dependency translation models as collections of finite state head transducers *Computational Linguistics*, 26(1), 45–60.

S. Bangalore and G. Riccardi. 2000. Stochastic finite-state models for spoken language machine translation. In *Proceedings of the Workshop on Embedded Machine Translation Systems.*, pp. 52–59.

P. Brown, S. Della Pietra, V. Della Pietra and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 12(2), 263–312.

B. Dorr and N. Habash. 2002. Interlingua approximation: A generation-heavy approach. In Proceedings of AMTA-2002.

W. A. Gale and K. W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. In *29th Annual Meeting of the ACL*, Berkeley, CA.

N. Habash and B. Dorr. 2003. A categorial variation database for English. In *Proceedings of NAACL-HLT 2003*

D. Jones and R. Havrilla. 1998. Twisted pair grammar: Support for rapid development of machine translation for low density languages. In Proceedings of AMTA98, pp. 318–332.

D. Marcu and U. Germann. 2002. *The ISI ReWrite Decoder Release 0.7.0b.* http://www.isi.edu/licensed-sw/rewrite-decoder/.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, Vol. 19.

M. Mohri, F. Pereira, and M. Riley. 1997.
ATT General-purpose finite-state machine software tools.
http://www.research.att.com/sw/tools/fsm/.

G. Ngai and R. Florian. Transformation-based learning in the fast lane. In *Proceedings of North Americal ACL 2001*, pages 40-47, June 2001.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447.

F.J. Och, C. Tillmann, H. Ney. Improved Alignment Models for Statistical Machine Translation. In *Proceedings of EMNLP 1999*, pp. 20-28.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901-904. Denver, CO, USA. http://www.speech.sri.com/projects/srilm/.

D. Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3), 377–404.

K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. In *Proceedings of ACL-2001*, pp. 523–529.

K. Yamada and K. Knight. 2002. A decoder for syntax-based statistical MT In *Proceedings of ACL-2002*, pp. 303–310.