Statistical Machine Translation Using Coercive Two-Level Syntactic Transduction

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 define, implement and evaluate a novel model for statistical machine translation, which is based on shallow syntactic analysis (part-of-speech tagging and phrase chunking) in both the source and target languages. It is able to model long-distance constituent motion and other syntactic phenomena without requiring a full parse in either language. We also examine aspects of lexical transfer, suggesting and exploring a concept of translation *coercion* across parts of speech, as well as a transfer model based on lemma-to-lemma translation probabilities, which holds promise for improving machine translation of low-density languages. Experiments are performed in both Arabic-to-English and French-to-English translation demonstrating the efficacy of the proposed techniques. Performance is automatically evaluated via the Bleu score metric.

1 Introduction

In this work we define, implement and evaluate a novel model for statistical machine translation (SMT).

Our goal was to produce a SMT system for translating foreign languages into English which utilizes some syntactic information 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 such a benefit. Allowing unconstrained translation reorderings at the word level generates a very large set of permutations that pose a difficult search problem at decoding time. We propose a model that makes use of shallow parses (text chunking) to support long-distance motion of phrases without requiring deeper analysis 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 and some interesting properties of lexical translation across parts of speech. One of the source languages we examine in this paper, 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 handle the above-described phenomenon effectively.

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. Much common local motion that is observed for French, such as adjective-noun swapping, is adequately 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, and hence longer distance constituent motion.

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 rotation 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 Door (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 lab has previously developed techniques for rapid training of such tools (Cucerzan and Yarowsky, 2002). Our translation experiments were carried out on two languages: Arabic and French. The Arabic training corpus was a subset of the United Nations (UN) parallel corpus which is being made available by the Linguistic Data Consortium. For French-English training, we used a portion of the Canadian Hansards. Both corpora utilized sentence-level alignments publicly distributed 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). French POS tagging was done using the trained French lexical tagger also provided with the fnTBL software. 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. Simple regular-expression-based phrase chunkers were developed by the authors for both Arabic and French, requiring less than a person-day each using existing multilingual learning tools.

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 (e.g., Arabic $\rightarrow$ English **and** English $\rightarrow$ Arabic) on our parallel corpora. The union of these bidirectional alignments was used to compute cross-language phrase correspondences by simple majority voting, and for purposes of estimating word translation probabilities, each link in this union 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 1.

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[1], PPNP[2], etc.) as given by the output of the phrase chunker. These are denoted in the top-level model presentation in Table 2(1). Given word alignment links, as described in Section 2, we compute phrasal alignments on training data. We contrain these to have cardinality
$(foreign) N \rightarrow 1 (English)$. Next, we collect counts over aligned phrase sequences and use the relative frequencies to estimate the probability distribution in Table 2(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. Table 2(3) encapsulates the remainder of the translation model, which is described below.

As an example, Table 3 shows the most probable aligned English phrase sequence generations given an Arabic simple sentence having the canonical VSO ordering. Also, note that all probabilities in the following

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

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

Arabic Example Sentence From Test Set	
(ARABIC)	*twSy Al- ljnp Al- sAdsp Al- jmEyp Al- EAmp b- AEtmAd m$rwE Al- mqrr Al- tAly :*
(PHR.-BRACKETED AR.)	*[twSy] [Al- ljnp Al- sAdsp] [Al- jmEyp Al- EAmp] [b- AEtmAd m$rwE Al- mqrr Al- tAly] [:]*
(AN ENG. GLOSS)	[recommends] [the committee the sixth] [the assembly the general] [to adoption draft the decision the following] [:]
(ENG. 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 1: 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 and French in this paper are italicized.

figures and tables are from the actual Arabic and French trained systems.

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 $\mathrm{VP}_{1,2}$	0.06

Table 3: 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 the corresponding 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 2. Each proposed English phrase in each of the phrase sequence possibilities, therefore, comes to the middle level of the translation model with access to the identity of the French phrases aligned to it. Phrase translation is implemented as shown in Table 4. 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 4 (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. Table 11 contains examples of some of these literal phrase translations from the French data.

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 4(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. Table 5 and Table 6 show 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-sequence, english-phrase-type), where a coarse POS is, for example, *N* instead of *NOUN-SG*. This is shown in Table 4(3).

In case further backoff is needed, as shown in Table 4(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 French and Arabic NP's.

Aligned English POS-tag Sequence Translation Probabilities (conditioned on Arabic POS-tag sequence from NP in example)	
$P(\mathrm{DT}_\emptyset\ \mathrm{JJ}_4\ \mathrm{NN}_2$	$\mid DET_1\ NOUN\text{-}SG_2\ DET_3\ ADJ_4, \mathrm{NP}) = 0.22$
$P(\mathrm{JJ}_4\ \mathrm{NN}_1$	$\mid DET_1\ NOUN\text{-}SG_2\ DET_3\ ADJ_4, \mathrm{NP}) = 0.20$
$P(\mathrm{DT}_\emptyset\ \mathrm{NN}_1$	$\mid DET_1\ NOUN\text{-}SG_2\ DET_3\ ADJ_4, \mathrm{NP}) = 0.13$
$P(\mathrm{DT}_\emptyset\ \mathrm{VBN}_4\ \mathrm{NNS}_2$	$\mid DET_1\ NOUN\text{-}SG_2\ DET_3\ ADJ_4, \mathrm{NP}) = 0.13$
$P(\mathrm{DT}_1\ \mathrm{NN}_2$	$\mid DET_1\ NOUN\text{-}SG_2\ DET_3\ ADJ_4, \mathrm{NP}) = 0.04$
$P(\mathrm{DT}_3\ \mathrm{JJ}_4\ \mathrm{NN}_2$	$\mid DET_1\ NOUN\text{-}SG_2\ DET_3\ ADJ_4, \mathrm{NP}) = 0.03$
$P(\mathrm{DT}_1\ \mathrm{VBN}_4\ \mathrm{NNS}_2$	$\mid DET_1\ NOUN\text{-}SG_2\ DET_3\ ADJ_4, \mathrm{NP}) = 0.03$
$P(\mathrm{DT}_\emptyset\ \mathrm{NN}_4\ \mathrm{NN}_2$	$\mid DET_1\ NOUN\text{-}SG_2\ DET_3\ ADJ_4, \mathrm{NP}) = 0.02$
$P(\mathrm{JJ}_4\ \mathrm{NNS}_2$	$\mid DET_1\ NOUN\text{-}SG_2\ DET_3\ ADJ_4, \mathrm{NP}) = 0.02$
$P(\mathrm{DT}_1\ \mathrm{JJ}_4\ \mathrm{NN}_2$	$\mid DET_1\ NOUN\text{-}SG_2\ DET_3\ ADJ_4, \mathrm{NP}) = 0.02$
$P(\mathrm{NN}_2$	$\mid DET_1\ NOUN\text{-}SG_2\ DET_3\ ADJ_4, \mathrm{NP}) = 0.02$
$P(\mathrm{NN}_4\ \mathrm{NN}_2$	$\mid DET_1\ NOUN\text{-}SG_2\ DET_3\ ADJ_4, \mathrm{NP}) = 0.02$

Table 5: From the running Arabic example, top English NP generations given an Arabic phrase *DET NOUN-SG DET ADJ*. Note: $\emptyset$ 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 The Basic Model

In the basic model of word generation, phrases may be translated directly as single atomic entities (as in Table 4(1)), or via phrasal decomposition to individual words translated independently, conditioned only on the source word and target POS. Word translation in the latter case

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

Table 2: Statement of the translation model at top level.

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

Table 4: 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[4].

Aligned English POS-tag Sequence Translation Probabilities (conditioned on Arabic POS-tag sequence from VP in example)	
$P($ VBZ$_1$	$\mid$ *VERB-IMP*$_1$, VP $)$ = .28
$P($ VBP$_1$	$\mid$ *VERB-IMP*$_1$, VP $)$ = .17
$P($ VBD$_1$	$\mid$ *VERB-IMP*$_1$, VP $)$ = .09
...	
$P($ MD$_\emptyset$ VB$_1$	$\mid$ *VERB-IMP*$_1$, VP $)$ = .06

Table 6: From the Arabic example, top English VP generations given an Arabic phrase *VERB-IMP*.

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 7 describes the backoff path for basic lexical transfer and presents a motivating example in the French word *droit*. Translation probabilities for one of the words in the example Arabic sentence can be found in Table 8.

4.3.2 Generation via a Lemma Model

To counter sparse data problems in estimating word translation probabilities, we also implemented a lemma-

Word Generation		
Examples		Model with Backoff Pathways
$P(\mathrm{W}_E \mid$ *droit* , NNS$)$		$P(\mathrm{W}_E \mid W_F, \mathrm{T}_{fine_E})$
rights	0.4389	p(rights $\mid$ *droit* , NNS)
benefits	0.0690	
people	0.0533	
laws	0.0188	
		$\downarrow$ (backoff if $C(W_F, \mathrm{T}_{fine_E}) = 0$)
$P(\mathrm{W}_E \mid$ *droit* , N$)$		$P(\mathrm{W}_E \mid W_F, \mathrm{T}_{coarse_E})$
right	0.4970	
law	0.1318	
rights	0.0424	p(rights $\mid$ *droit* , N)
property	0.0115	
		$\downarrow$ (backoff if $C(W_F, \mathrm{T}_{coarse_E}) = 0$)
$P(\mathrm{W}_E \mid$ *droit*$)$		$P(\mathrm{W}_E \mid W_F)$
right	0.2919	
entitled	0.0663	
law	0.0652	
the	0.0249	
to	0.0240	
rights	0.0210	p(rights $\mid$ *droit*)
		$\downarrow$ (backoff if $C(W_F) = 0$)
		$p($ *UNKNOWN_WORD* $\mid W_F) = 1$

Table 7: Description of the conditioning for different levels of backoff in the lexical transfer model. The example shows translations for the French word *droit* ("right") conditioned on decreasingly specific values. The progressively lower ranking of the correct translation as we move from fine, to coarse, to no POS, illustrates the benefit of conditioning generation on the English part of speech.

Arabic Word	English POS	English Wd.	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	questions	0.027
ljnp	NNS	representatives	0.024
ljnp	N	committee	0.577
ljnp	N	commission	0.227
ljnp	N	subcommittee	0.035

Table 8: From running example, translation probabilities for Arabic noun *ljnp*, "committee".

based model for word translation. Under this model, translation distributions are estimated by counting word alignment links between foreign and English lemmas, assuming a lemmatization of both sides of the parallel corpus as input. The form of the model is illustrated below:

$$
\begin{aligned}
&P(\ \mathrm{W}_E \mid W_F, T_{coarse_F}, \mathrm{T}_{fine_E}) = \\
&\quad P(\ \mathrm{W}_E \mid \mathrm{lemma}_E\ , T_{coarse_F}\ , \mathrm{T}_{fine_E}) * \\
&\quad P(\ \mathrm{lemma}_E \mid \mathit{lemma}_F\ , T_{coarse_F}\ , \mathrm{T}_{fine_E}\ ) * \\
&\quad P(\ \mathit{lemma}_F \mid W_F\ , T_{coarse_F}\ , \mathrm{T}_{fine_E}\ ) \\
&\quad\ \Downarrow \textbf{ approximated by} \\
&\quad P(\ \mathrm{W}_E \mid \mathrm{lemma}_E\ , \mathrm{T}_{fine_E}) * \\
&\quad P(\ \mathrm{lemma}_E \mid \mathit{lemma}_F\ , \mathrm{T}_{coarse_E}\ ) * \\
&\quad P(\ \mathit{lemma}_F \mid W_F\ , T_{coarse_F}\ )
\end{aligned}
$$

First, note that $P(\ \mathit{lemma}_F \mid W_F\ , T_{coarse_F}\ )$ is very simply a hard lemma assignment by the foreign language lemmatizer. Second, English word generation from English lemma and coarse POS ($P(\ \mathrm{W}_E \mid \mathrm{lemma}_E\ , \mathrm{T}_{fine_E})$) is programmatic, and can be handled by means of rules in conjunction with a lookup table for irregular forms. The only distribution here that must be estimated from data is $P(\ \mathrm{lemma}_E \mid \mathit{lemma}_F\ , \mathrm{T}_{coarse_E}\ )$. This is done as described above. Furthermore, given an electronic translation dictionary, even this distribution can be pre-loaded: indeed, we expect this to be an advantage of the lemma model, and an example of a good opportunity for integrating compiled human knowledge about language into an SMT system. Some examples of the lemma model combating sparse data problems inherent in the basic word-to-word models can be found in Table 9.

4.3.3 Coercion

Lexical coercion is a phenomenon that sometimes occurs when we condition translation of a foreign word on the word and the 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, for instance, a nominal form with similar meaning from a French adjective, or an adjectival realization of a French verb's meaning; some examples of this phenomenon are shown in Table 10. We find this coercion effect to be of interest because it identifies interesting associations of meaning. For example, in Table 10 "willing" and "ready" are both sensible ways to realize the meaning of the action "to accept" in a passive, descriptive mode. *droit* behaves similarly. Though the English verb "to right' or "to be righted" does not have the philosophical/judicial entitlement sense of the noun "right", we see that the model has learned to realize the meaning in an active, verbal form: e.g., VBG 'receiving" and VB "qualify".

5 Decoding

Decoding was implemented by constructing 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 1. The sentence-level (aligned phrase sequence generation) and phrase-level (aligned intra-phrase sequence generation) translation 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 1 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 2 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 dev-test 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.

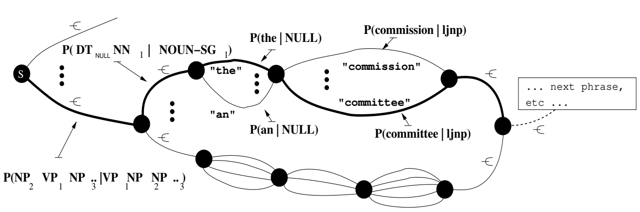

Figure 1: An illustration of the translation model structure for an Arabic test sentence.

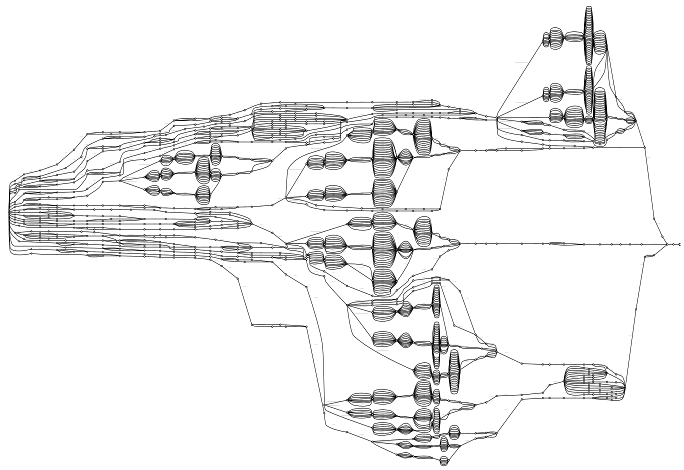

Figure 2: A portion of the translation model for an Arabic test sentence, compacted and aggressively pruned by path probability for display purposes.

6 Evaluation

Results Tables A and B below list evaluation results for translation on the Arabic and French test sets respectively. In each case, results for a comparison system – the Giza++ IBM Model 4 implementation (Och and Ney, 2000) with the ReWrite decoder (Marcu and Germann, 2002) – are included as a benchmark. Results were generated for training corpora of varying sizes. For Arabic, we ran our system on two large subsets of the UN corpus and evaluated on a 200-sentence held-out set (refer to Results Table A below). For the 150K sentence Arabic training set, Giza++ and the shallow syntax model achieved very similar performance. We were unable to obtain evaluation numbers for Giza++/ReWrite on the large Arabic training set, however, since its language model component has a vocabulary size limit which was exceeded in the larger corpus. In French we observed the systems to perform similarly on the small training sets we used (Results Table B). We performed some experiments in classifier combination using the two compatible (150K-training-sentence) Arabic systems, wherein a small devtest set was used to identify simple system combination parameters based on model confidence and sentence length. In situations where our system was confident we used its output, and used Giza++ output otherwise. We achieved a 3% boost in Bleu score over Giza++ performance on the evaluation set with these very simple classifier combination techniques, and anticipate that research in this direction – classifier combination of diversely trained SMT systems – could yield significant performance improvements.

	Bleu Score	
System	150K Trn. Sent.	500K Trn. Sent.
Giza++/ReWrite Decoder	0.17	*
2-level Syntax Model	0.17	0.18

Results Table A: Results comparison for Arabic to English translation on the UN corpus, with a 200-sentence evaluation set. Note that Giza++/ReWrite cannot be run for the 500K sentence training set; the CMU Language Modeling Toolkit, which ReWrite uses, has a vocabulary size limit which is exceeded in the 500K corpus.

	Bleu Score	
System	5K Trn. Sent.	20K Trn. Sent.
Giza++/ReWrite Decoder	0.08	0.11
2-level Syntax Model	0.08	0.09

Results Table B: Results comparison for French to English translation on the Canadian Hansards corpus (200-sentence evaluation set).

7 Conclusions

We have described and implemented an original syntax-based statistical translation model that yields baseline results which compete successfully with other state-of-the-art SMT models. This is particularly encouraging in that the authors are not well-versed in Arabic or French and it appears that the quality of the rule-based phrase chunkers we developed in a single person-day offers substantial room for improvement. We expect to be able to attain improved bracketings from native speakers and, in addition, via translingual projection of existing bracketers. Secondly, the lemma model we have proposed for lexical transfer provides an efficient framework for integrating electronic dictionaries into SMT models. Although we have at this time no large electronic dictionaries for either Arabic or French, efforts are underway to acquire electronic or scanned paper dictionaries for this purpose. We did evaluate the lemma models in isolation for French and Arabic without dictionary inclusion, but in each experiment the results did not differ significantly from the word-specific lexical transfer models, despite their substantially reduced dimensionality. We anticipate that the relatively seamless direct incorporation of dictionaries into the lemma-based models will be particularly effective for translating low-density languages, which suffer from data sparseness in the face of limited parallel text. Finally, we incorporated lexical translation coercion models into this full SMT framework, the induction of which is a phenomenon of interest in its own right.

8 Acknowledgements

This work was supported in part by NSF grant number IIS-9985033. In addition, we owe many thanks to colleagues who generously lent their time and insights. David Smith shared his tools for Arabic part-of-speech tagging and morphological analysis and answered many questions about the Arabic language. Thanks to Skankar Kumar and Sanjeev Khudanpur for numerous helpful discussions.

9 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.

S. Cucerzan and D. Yarowsky. 2002. Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day. *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL), Taipei, 2002*.

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

Word Translation Probabilities			
Word translation for *mangeait* conditioned on French Word, EnglishPOS			
mangeait	VBG	eating	1.00
mangeait	VB	go	0.50
mangeait	VB	anticipate	0.50
mangeait	VBD	were	1.00
mangeait	VBP	knelt	1.00
mangeait	NN	bill	1.00
Word translation for *mangeait* conditioned on French Word, English Coarse POS			
mangeait	V	eating	0.44
mangeait	V	were	0.22
mangeait	V	knelt	0.11
mangeait	V	go	0.11
mangeait	V	anticipate	0.11
mangeait	N	bill	1.00
Word translation for *mangeait* conditioned on French Word only			
mangeait		eating	0.29
mangeait		were	0.14
mangeait		go	0.07
mangeait		bill	0.07
Word translation for *mangeant* conditioned on French Word, EnglishPOS			
mangeant	RB	mostly	1.00
mangeant	JJ	final	1.00
mangeant	VBN	obtained	1.00
mangeant	VBG	eating	1.00
mangeant	WP	who	1.00
mangeant	IN	through	1.00
mangeant	NN	lard	1.00
mangeant	VBZ	eats	0.50
mangeant	VBZ	comes	0.50

Lemma Translation Probabilities			
Generation of a verb lemma given *manger*			
manger	V	eat	0.60
manger	V	feed	0.05
manger	V	have	0.04
Generation of a noun lemma given *manger*			
manger	N	meal	0.06
manger	N	trough	0.04
manger	N	loaf	0.04
manger	N	food	0.04
Generation of an adj. lemma given *manger*			
manger	J	hungry	0.33
Raw lemma translation probabilities (ignoring English Coarse POS)			
manger		eat	0.28
manger		to	0.03
manger		feed	0.03
manger		out	0.02
manger		have	0.02
manger		are	0.02
manger		,	0.02
manger		you	0.01
manger		meal	0.01

Table 9: **Direct generation** (word-to-word translation probabilities at the various levels of backoff) is contrasted with **lemma generation**. *Manger* ("to eat") is a relatively rare word in the Hansards. Note that due to low counts, the desired verb POS (target of generation) for "eat" may not have been observed as a translation in training data. In addition, in this situation, noisy word alignments may cause an incorrect translation to have similar estimated translation probability. This problem is addressed by the lemma model; note the much sharper probability distribution for verb lemmas given *manger*. Generation of English inflections given lemma and target POS is algorithmic (and irregular exceptions are handled via a lookup table).

French Wd.	Eng. POS	Eng. Wd.	Prob.
accepter	JJ	unacceptable	0.12
accepter	JJ	acceptable	0.12
accepter	JJ	willing	0.11
accepter	JJ	ready	0.03
accepter	NN	acceptance	0.09
accepter	NN	amendment	0.03
droit	VBN	entitled	0.66
droit	VBN	allowed	0.09
droit	VBN	denied	0.03
droit	VBN	given	0.02
droit	VBN	permitted	0.02
droit	VBN	justified	0.01
droit	VBN	qualified	0.01
droit	VBN	allotted	0.01
droit	VB	qualify	0.14
droit	VB	be	0.11
droit	VB	have	0.09
droit	VB	receive	0.08
droit	VB	get	0.07
droit	VB	expect	0.03
droit	VBG	receiving	0.11
droit	VBG	getting	0.08
droit	NNS	rights	0.44
droit	NNS	benefits	0.69

Table 10: Examples of word translation coercions. Coercions of the French verb *accepter* "to accept" and the French noun *droit* "right" (there is parallel polysemy between the two languages for this word, but the predominant sense in our corpus is the philosophical/judicial sense, as opposed to the direction).

Eng. Phrase Type	French Phrase	Eng. Phrase	Prob.
NP	*dans_le_cas_présent*	a_situation	0.25
NP	*dans_le_cas_présent*	the_subject_of_debate	0.25
NP	*dans_le_cas_présent*	the_position	0.25
NP	*dans_le_cas_présent*	it	0.25
VP	*dans_le_cas_présent*	should_apply	1.00
ADVP	*dans_le_cas_présent*	really	1.00
PPNP	*dans_le_cas_présent*	in_this_case	0.63
PPNP	*dans_le_cas_présent*	in_this_instance	0.04
PPNP	*dans_le_cas_présent*	in_this_actual_case	0.04
PPNP	*dans_le_cas_présent*	in_this_particular_case	0.04
PPNP	*dans_le_cas_présent*	in_that_case	0.04
PPNP	*dans_le_cas_présent*	in_the_present_circumstances	0.04
VP	*acceptons*	accept	0.48
VP	*acceptons*	agree	0.14
NP	*acceptons*	this_consent	1.00
PPNP	*par_an*	per_year	0.67
PPNP	*par_an*	in_each_year	0.03
PPNP	*par_an*	for_a_year	0.03
ADVP	*par_an*	annually	1.00
NP	*par_an*	a_year	0.79
NP	*par_an*	each_year	0.02
NP	*un_discours*	a_speech	0.83
NP	*un_discours*	an_address	0.05
VP	*un_discours*	to_speak	1.00

Table 11: Examples of direct phrase translations (see Table 4(1)), including some coercions.