The Translation Template Modeling Framework for Statistical Machine Translation

Shankar Kumar

with Bill Byrne

Center for Language and Speech Processing, Johns Hopkins University
& Machine Intelligence Laboratory, Cambridge University Engineering Department

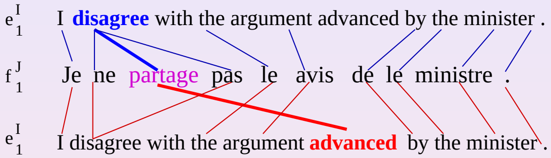
March 8, 2005

Introduction

- A statistical modeling framework for machine translation
  - The Generative Source-Channel Model
  - Implementation Using Weighted Finite State Transducers
  - Investigative Experiments
- Minimum Bayes-Risk Translation
- Recent Work
  - Phrase Reordering Model
  - Discriminative Training

Bitext word alignment and Loss functions

Alignment Error



- Measures # of non-NULL alignment links by which the candidate alignment differs from reference alignment
- Alignment Error Rate (Och and Ney '00)

$$
\begin{aligned}
AER(B, B') &= \frac{|B| + |B'| - 2|B \cap B'|}{|B| + |B'|} \\
&= \frac{10 + 10 - 2 * 9}{10 + 10} = \frac{2}{20} = 10\%
\end{aligned}
$$

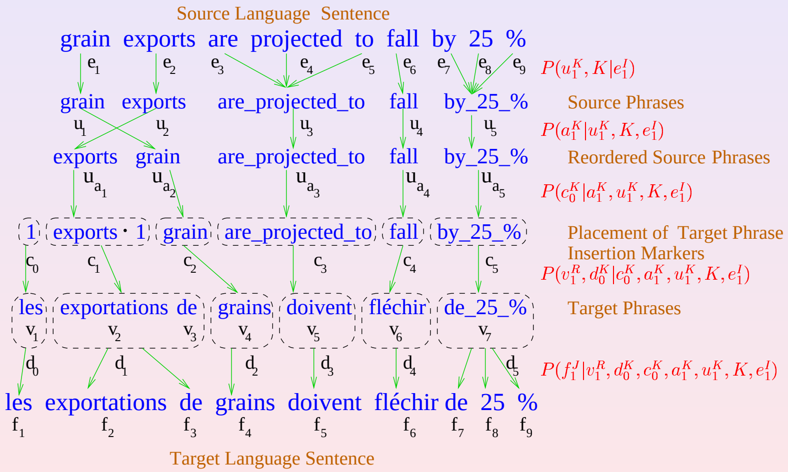
Outline

Translation Template Model

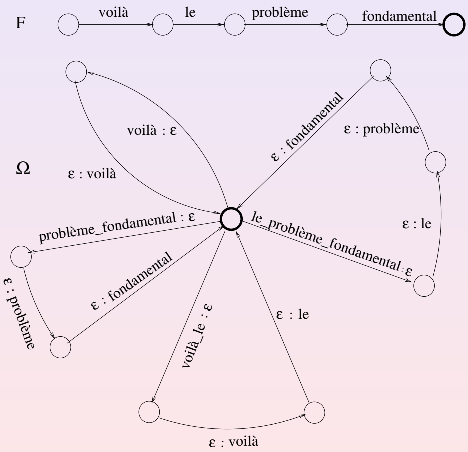
Refinements within the TTM Framework

- Minimum Bayes-Risk Translation
- A Phrase Reordering Model inside the TTM
- Discriminative Training

Generative Translation Process underlying the TTM



Target Phrase Segmentation Transducer Ω



Based on a Fixed Set of French phrases

Assume F is the sentence to be translated

Phrase sequences that could have generated F: $\Omega \circ F$

voilà_le problème_fondamental

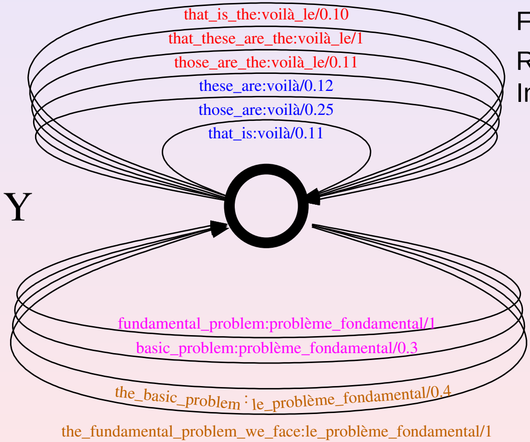
voilà le_problème_fondamental

The Phrase Pair Inventory

English Phrase u	French Phrase v	Phrase Transduction Probability $P(v\|u)$
hear_hear	bravo	0.8
	bravo_bravo	0.15
	ordre	0.05
terms_of_reference	mandat	0.8
	de_son_mandat	0.2

- Phrase Pair Inventory affects the performance of the TTM
  - Word Alignment Quality of underlying models
  - Coverage of phrases on the test set

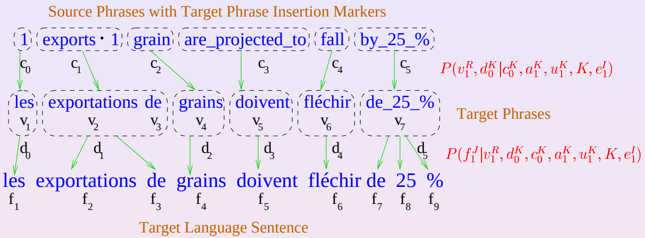
Phrase Translation Transducer Y



Map English phrases into French phrases

Realizes the Phrase-Pair Inventory

Phrase Translation with WFSTs



- Sequences c_0^K that could have generated F : $Y \circ \Omega \circ F$
- H1 - that_these_are_the fundamental_problem : voilà_le problème_fondamental
  $P(f_1^J | v_1^R) P(v_1^R, d_0^K | c_0^K) = 1 \times 1 = 1$
- H16 - that_is the_basic_problem: voilà le_problème_fondamental
  $P(f_1^J | v_1^R) P(v_1^R, d_0^K | c_0^K) = 1 \times 0.05 = 0.05$

An Overview of WFSTs for Alignment and Translation

Given a French sentence f_1^J to be translated into English, we build the following transducers (in this order)

- F to represent the French sentence
- Ω maps French phrases in our Phrase-Pair Inventory (PPI) to words in F
- Y maps English phrases to French phrases in Ω with probabilities given by the PPI
- Φ inserts French phrase insertion markers
- W maps English words to English phrases seen in Y

If f_1^J is to be aligned with an English sentence e_1^I, build E to represent e_1^I

- Build a n-gram backoff LM and compile it as a weighted acceptor G
- Assume a fixed phrase order model (monotone search)
  - I will discuss phrase reordering later

Bitext Word Alignment and Translation Via WFSTs

- TTM Joint Model :$P(f_1^J, v_1^R, d_0^K, c_0^K, a_1^K, u_1^K, K, e_1^I)$
- Map Alignment of a sentence pair f_1^J,e_1^I

$$\{\hat{K}, \hat{u}_1^{\hat{K}}, \hat{a}_1^{\hat{K}}, \hat{c}_0^{\hat{K}}, \hat{d}_0^{\hat{K}}, \hat{v}_1^{\hat{R}}\} = \underset{K, u_1^K, a_1^K, c_0^K, d_0^K, v_1^R}{\operatorname{argmax}} P(K, u_1^K, a_1^K, c_0^K, d_0^K, v_1^R | e_1^I, f_1^J)$$

- Map Translation of French sentence f_1^J

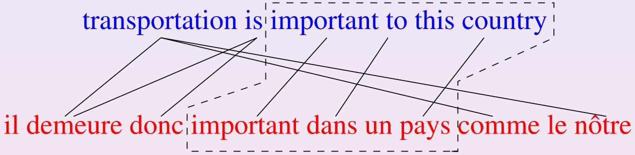
$$\{\hat{e}_1^{\hat{I}}, \hat{K}, \hat{u}_1^{\hat{K}}, \hat{a}_1^{\hat{K}}, \hat{c}_0^{\hat{K}}, \hat{d}_0^{\hat{K}}, \hat{v}_1^{\hat{R}}\} = \underset{e_1^I, K, u_1^K, a_1^K, c_0^K, d_0^K, v_1^R}{\operatorname{argmax}} P(K, u_1^K, a_1^K, c_0^k, d_0^K, v_1^R | f_1^J)$$

WFST Operations with Monotone Search

- Alignment
  1. Generate the alignment lattice: $\mathcal{B} = E \circ W \circ \Phi \circ Y \circ \Omega \circ F$
  2. MAP Alignment : least cost path in $\mathcal{B}$
- Translation
  1. Generate the translation lattice: $\mathcal{T} = G \circ U \circ \Phi \circ Y \circ \Omega \circ F$
  2. MAP Translation : least cost path in $\mathcal{T}$

Problems with Bitext Word Alignment under the TTM

Consider extraction of phrase pairs from word alignments



- Extracted Phrase-Pair Inventory is not rich enough to cover all the sentence-pairs
  - This pair is assigned a probability of zero under the model
  - In fact, most sentences from the training bitext have probability zero
- Solution:
  - TTM already allows insertions of target phrases
  - In addition, we allow deletions of source phrases

Source Phrase Deletion in Bitext Word Alignment

TTM Generative Process allowing insertions and deletions

transportation is important to this country

transportation is important_to_this_country

I3 transportation is important_to_this_country I3

il_demeure_donc ε ε important_dans_un_pays comme_le_nôtre

il demeure donc important dans un pays comme le nôtre

- Novel use of phrase-based translation models for alignment
- Word Alignments hypothesized by TTM are very accurate
- Allows development of parameter estimation procedures

Word Alignment Quality of Underlying Models

- Task: French-English Hansards (Train: 48K sent. pairs, Test: 500 sents)
- Build 4 nested subsets of bitext & train IBM translation models over each set. For each model :
  - Obtain word alignments over test bitext for which reference word alignments are available $\longrightarrow$ measure Alignment Error Rate
  - Obtain word alignments over a fixed 5K subset of the training bitext
  - Collect Phrase Pair Inventories over these word alignments
  - Construct the TTM under this inventory and translate the test set $\longrightarrow$ measure Translation Performance

# of sentence-pairs (K)	AER (%)	BLEU (%)
5	20.6	17.4
12	15.9	18.6
24	13.9	19.2
48	12.1	19.6

- More bitext improves alignment quality of the underlying models, and this in turn improves translation quality under the TTM

Coverage of the Test Set by the Phrase-Pair Inventory

- Task: French-English Hansards (Train: 48K sent. pairs, Test: 500 sents)
- Train IBM translation models on all bitext and obtain word alignments
- Collect Phrase Pair Inventories (PPIs) over 4 subsets of these word alignments
  - Alignment quality over these variable size inventories is held constant
- Coverage = % of phrases in the test set that exist in the PPI

# of sentence-pairs (K)	Test-Set Coverage (%)	BLEU (%)
5	20.8	19.6
12	26.8	20.8
24	31.4	21.5
48	36.0	22.3

- A higher coverage of the test set by PPI improves translation performance

Outline

Minimum Bayes-Risk Machine Translation

- Translation can be evaluated in various ways : BLEU, WER, Position Independent WER (PER)
- Given a translation loss function, we build Minimum Bayes-Risk decoders to optimize performance under the loss function
- Setup
  - A baseline translation model to give the probabilities over translations: $P(E|F)$
  - A set $\mathcal{E}$ of N-Best Translations of F
  - A Loss function $L(E, E')$ that measures the the quality of E' relative to E
- MBR Decoder

$$\hat{E} = \underset{E' \in \mathcal{E}}{\operatorname{argmin}} \sum_{E \in \mathcal{E}} L(E, E') P(E|F)$$

- MAP decoder is the MBR decoder under Sentence Error Loss function
- Related to ROVER used in ASR system combination

Performance of MBR Decoders

- Experimental Setup: TTM system trained on the TIDES 2004 Chinese-English Bitext (200M English words, 170M Chinese words)
- Test Set: Chinese-English NIST MT Task (2002) , 878 sentences, 1000-best lists

		Performance Metrics		
		BLEU (%)	mWER(%)	mPER (%)
	MAP(baseline)	28.5	63.9	41.8
M	BLEU	28.8	63.9	41.6
B	WER	28.8	63.3	41.3
R	PER	29.0*	63.5	41.0

- MBR Decoding allows translation process to be tuned for specific loss functions
- *On N-best lists borrowed from other research systems (ISI) we did not see this behavior

A WFST Local Phrase Reordering Model

- Our initial phrase order model reorders the English phrase sequence into the French phrase order
  Difficult to realize this model with WFSTs :
  - Alignment can only be done with a single English phrase segmentation
  - Can't be used in translation $\longrightarrow$ Employ Monotone Search
- Consider an alternate generative process
  - First generate a French phrase sequence in English phrase order
  - Next reorder this sequence into French phrase order under the Local Phrase Reordering Model
- Possible to realize with WFSTs both in alignment and translation
- Lose English phrase reordering process
- Reordering is prior to Insertion of Target Phrases

Experiments on the NIST Arabic-English Task

- Experimental Setup: TTM system trained on the TIDES 2004 Arabic-English Bitext (132M English words, 123M Arabic words)
- Test Sets: NIST 2002 (1043 sentences), NIST 2003 (663 sents), NIST 2004 (1353 sents)
- For Phrase-Pair (u, v), reordering parameter: $P(b = +1|v, u)$

	Eval02	Eval03	Eval04		
			News	Editorials	Speeches
No Reordering	38.1	40.3	40.1	30.7	35.7
Allow Reordering					
Single p	41.1	42.1	42.5	31.7	35.7
One p per phrase-pair	41.3	43.1	43.5	32.2	36.3

- Initial Experiments show good improvements by allowing reordering
- Investigating longer distance phrase reordering $\{0, +1, +2...\}$
- Interaction with higher-order n-gram LMs

A Discriminative Training Procedure

- Cast TTM as a log-linear model with scaling factors $\Lambda = \lambda_1^M$
  $P_{TTM}(E|F) = \prod_{m=1}^{M} p_m(E,F)^{\lambda_m}$
  λ's applied to WFSTs during decoding
- Minimum Error Training (Och 2003) : Estimate parameters of a log-linear model to reduce error count over a development set
- Minimize an Error Function $\mathcal{E}$ (BLEU) over a development corpus:

Results on Arabic-English:

	Eval02 (Dev)	Eval03 (Test)
Baseline	41.1	42.1
MET	43.3	44.9

N-best lists used for training

Multidimensional search in M dim space by Powell's algorithm

MET gives good improvement over a state-of-the-art baseline - shows how 'primitive' the discriminative training is in comparison with ASR ...

Conclusions

- Translation Template Model
  - A powerful generative source-channel model for SMT
  - Bitext word alignment and translation using standard optimized finite state operations
    - No need for a specialized decoder
  - Allows generation of N-best Lists and Lattices of Word Alignments and Translations
- Refinements in the TTM Framework
  - MBR decoders for translation allow translation to be tuned under specific loss functions
  - WFST Phrase Reordering Model extends the generative process underlying the TTM
  - Discriminative Training to estimate TTM scaling factors to optimize BLEU over a development set
- TTM Cookbook available for use in the MIL

Thank you!

References

P. .F. Brown et. al.
The mathematics of statistical machine translation: parameter estimation.
In *Computational Linguistics, 19(2)*, 1993

F. J. Och, C. Tillmann and H. Ney
Improved alignment models for statistical machine translation
In *Proceedings of EMNLP & VLC*, College Park, MD, USA, 1999

S. Kumar and W. Byrne.
Minimum Bayes-risk alignment of bilingual texts.
In *Proc. of EMNLP*, Philadelphia, PA, USA, 2002.

S. Kumar and W. Byrne.
A weighted finite state transducer implementation of the alignment template model for statistical machine translation.
In *Proc. of HLT-NAACL*, Edmonton, Canada, 2003.

F. J. Och
Minimum error rate training in statistical machine translation
In *Proc of ACL*, Sapporo, Japan, 2003.

S. Kumar and W. Byrne.
Minimum Bayes-risk decoding for statistical machine translation.
In *Proc. of HLT-NAACL*, Boston, MA, USA, 2004.

S. Kumar, Y. Deng and W. Byrne.
A weighted finite state transducer translation template model for statistical machine translation.
To appear in Journal of Natural Language Engineering, 2005.