Finite-State Transducer-based Statistical Machine Translation using Joint Probabilities

Srinivas Bangalore, Stephan Kanthak, Patrick Haffner

{srini, skanthak, haffner}@research.att.com

Introduction

MT viewed here to consist of two major subproblems:

- Lexical choice - generate target words
- Reordering - find the right target word order

Generative Model, FST-based (AT&T Evaluation System):

- Training
- Decoding
- Reordering

Discriminative Models for MT:

- Sequential Maximum Entropy (MaxEnt) Lexical Choice
- Bag-Of-Words (BOW) Maximum Entropy Lexical Choice

FST-based MT (Training)

Generative Sequence Classifier

Approach for training the model:

1. Create word alignment for bilingual sentence-aligned corpus
2. Pair aligned phrases, reorder words within target phrases and create sequences of bilingual (-> joint) phrase tuples from the word alignment
3. Estimate n-gram language model (LM) on sequence of bilingual phrase tuples
4. Convert LM into a weighted finite-state transducer (WFST) by splitting bilingual phrases into input and output labels

See also: Bangalore (2000), Vidal (2004), Kanthak (2004), Crego (2004)

FST-based MT (Bilanguage Example)

Chinese: 你 想 在 咖 啡 里 加 奶 油 和 糖 吗？

English: Would you like cream and sugar in your coffee ?

Bilanguage: (你, Would you) (想, like) (在, in) (咖, your) (啡, coffee) (里,) (加,) (奶, cream) (油,) (和, and) (糖, sugar) (吗,) (？, ?)

FST-based MT (Decoding)

Using the WFST *T* estimated from bilingual phrases/tuples, decoding a source sentence *F* into the best target translation *E* can be performed by FST-*composition* followed by best-path search:

$$E^* = \pi_1(best(F \circ T))$$

Using an additional model *W* to penalize word insertions:

$$E^* = \pi_1(best(F \circ T \circ W))$$

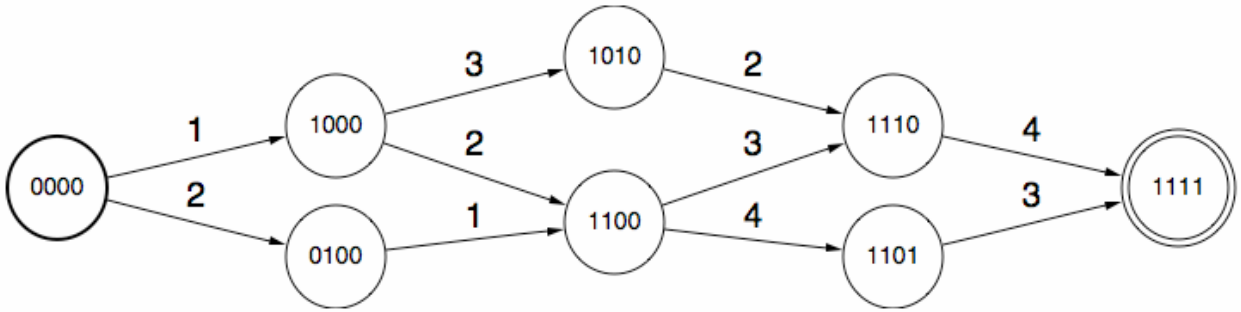
Additionally using global reordering:

Source: $E^* = \pi_1(best(perm(F) \circ T \circ W))$

Target: $E^* = best(perm(\pi_1(n\text{-}best(F \circ T \circ W))) \circ LM_E)$

FST-based MT (Permutation)

Constraint permutations: *local* (N = 4, P = 2):



Properties:

- *Not* finite-state => if at all possible, only applicable to statically compiled speech-to-speech translation FSTs for small *P*
- Only feasible with on-demand exploration of the automaton
- State-space complexity $O(2^N)$ for $P > N$, and $O(N * 2^P)$ for $P \leq N$
- Memoization during decoding is the limiting factor

See also: Kanthak (2005)

Sequential MaxEnt Lexical Choice Model

Idea: model *P(E|F)* directly

- Obvious solution: discriminative sequence classifier, e.g. CRF
- Problem: CRFs yet only applied to either small tasks or by making crude assumptions

Use conditional MaxEnt with independence assumptions instead:

$$P(E \mid F) = \prod_{i}^{I} P(e_i \mid \Phi(F,i))$$

- Trained directly on the bilanguage constructed in the FST approach
- Feature functions used here: source words in context

See also: Bangalore, Haffner (ICSLP 2006)

Sequential MaxEnt Lexical Choice Model

High complexity of the classifier O(N*F*C):

- Here N = 100k, F = 100k, C = 10k
- Reduce complexity by using binary 1-vs-other classifiers (another independence assumption), trick to train the classifiers in parallel
- We use AT&T's highly optimized, scalable large-margin classifier implementation *LLAMA*: e.g. in this case, training of the classifier is about 40x faster compared to LIBSVM
- Still pretty slow in decoding as all classifiers have to be evaluated at each source position

Bag-Of-Words MaxEnt Lexical Choice Model

Sequential MaxEnt Model suffers from:

- Improper alignment, e.g. from GIZA++
- Early and fixed reordering decisions due to the bilanguage
- Independence assumptions

Bag-Of-Words MaxEnt Lexical Choice Model:

- Just different parameterization of the sequential MaxEnt model
- Decoding:

$$BOW^*(F,\theta) = \{e \mid P(e \mid \Phi(F)) > \theta\}$$

Properties:

- Doesn't use alignment
- No independence assumptions

BOW Lexical Choice Model (Refinements)

Length modelling:

- Produce larger bag than sentence length (tuned by cutoff parameter)
- Allow for word deletions in reordering phase (additional global deletion penalty)

Reordering:

- Exactly the same as in FST and sequential MaxEnt model
- Interprets bag-of-words as sequence of words => window size parameter has no meaning anymore

Comparison: Sequential and BOW MaxEnt

	Sequential Classifier	**BOW Classifier**
Output target	Target word for each source position *i*	Target word given a source sentence
Input features	*BOgram(F,i-d,i+d):* bag of n-grams in source sent. in the interval *[i-d, i+d]*	*BOgram(F,0,|F|):* bag of n-grams in source sentence
Probabilities	$P(e_i \mid BOgram(F,i-d,i+d))$ Independence assumption between the labels	*P(BOW(E) | BOgram(F,0,|F|))*
Number of classes	One per target word or phrase	
Training samples	One per source token	One per sentence pair
Preprocessing	Source/target word alignment	Source/target sentence alignment

Corpus Statistics

Statistics of the supplied corpora for the IWSLT
Chinese -> English Speech Translation Task

	Training		Dev 2005		Dev 2006	
	Chinese	**English**	**Chinese**	**English**	**Chinese**	**English**
Sentences	46,311		506		489	
Running Words	351,060	376,615	3,826	3,897	5,214	6,362*
Vocabulary	11,178	11,232	931	898	1,136	1,134*
Singletons	4,348	4,866	600	538	619	574*
OOVs [%]	-	-	0.6	0.3	0.9	1.0
ASR WER [%]	-	-	-	-	25.2	-

* Statistics collected only on the first of multiple references

Experimental Setup

Chinese data split into sequences of characters

Punctuation marks:

- Automatically inserted using MaxEnt classifiers
- 6 classes: ; , . ? ! and none
- Model trained on IWSLT official training data

Target reordering:

- Used for all 3 approaches consistenly
- 4-gram
- Language model trained *only* on English part of IWSLT training data

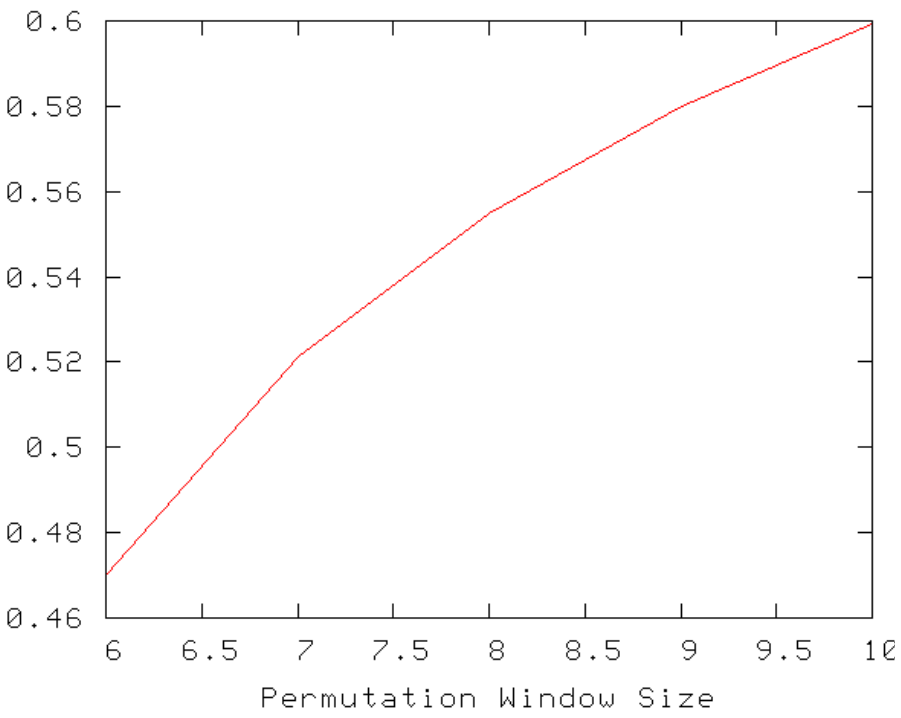
Experimental Results

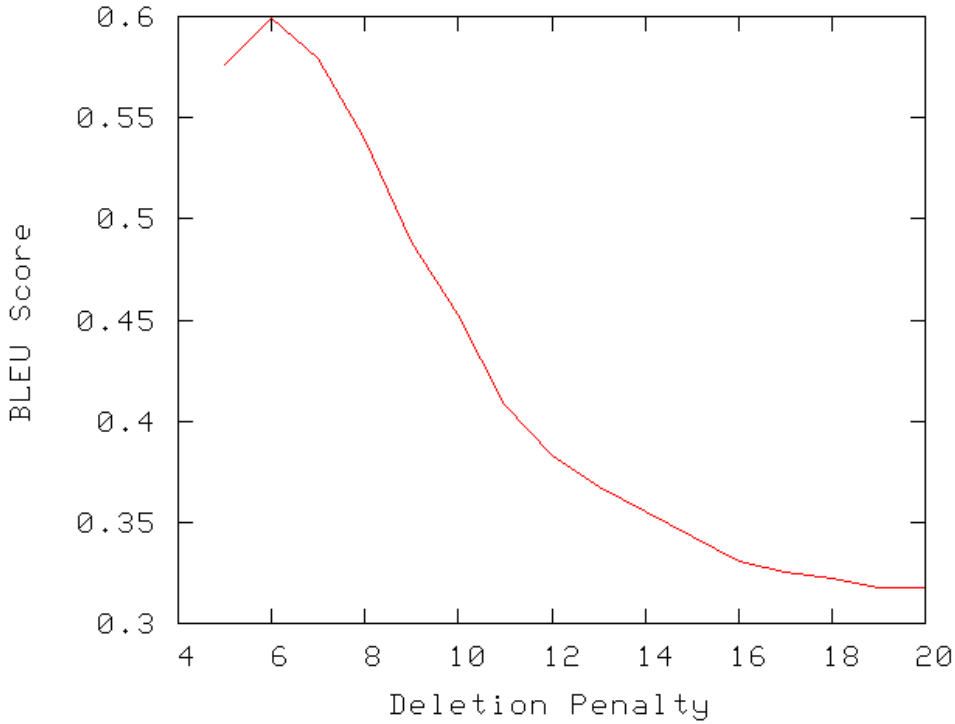
Comparison of mBLEU scores for the 3 different translation approaches on the IWSLT Chinese -> English Speech Translation Test Corpora

	Dev 2005	Dev 2006		Eval 2006	
	Text	Text	ASR	Text	ASR
FST	51.8	19.5	16.5	14.4	12.3
SeqMaxEnt	53.5	19.4	16.3	-	-
BOWMaxEnt	59.9	19.3	16.6	-	-
FST w/06	-	22.3*	-	16.0	-

* average of 10-fold split on Dev 2006, rest was added to training with weight 25

Parameters in BOW Lexical Choice Model





- Performance currently limited by permutation window size
- Also no pruning applied

Conclusion & Outlook

This talk was about:

- AT&T's generative FST model
- 2 new discriminative models for lexical choice
- Bag-of-Words model does not rely on word alignment at all
- Discriminative models superior to generative FST approach

Future work:

- More features for both discriminative approaches
- Better reordering framework for BOW model

Thank you for your attention!

Questions please.