Minimum Bayes-Risk Decoding for Statistical Machine Translation

May 4, 2004

Shankar Kumar and Bill Byrne

Center for Language and Speech Processing

Department of Electrical and Computer Engineering

The Johns Hopkins University,

Baltimore, MD 21218

Introduction

- Statistical Machine Translation systems can be evaluated using a variety of metrics
  - Different aspects of translation quality: BLEU, NIST, WER, PER, ..
  - Application specific criteria: usefulness for IR, summarization etc
- Maximum Likelihood techniques are used in decision processes of most current Statistical MT systems
  - Do not explicitly take into account evaluation criteria
- Minimum Bayes-Risk Decoding
  - Automatic systems tuned for desired evaluation criteria
  - Formulation in Statistical Machine Translation
  - Will show performance gains by matching decoder to the evaluation criterion

Minimum Bayes-Risk (MBR) Decoding Framework

- Decision processes optimized for specific loss functions
  - Automatic Speech Recognition (Goel and Byrne CSL ’00)
  - Bitext Word Alignment (Kumar and Byrne EMNLP ’02)
- MBR decoding for two translation scenarios
  - Loss functions derived from evaluation metrics
  - Design of specialized loss functions to incorporate desired characteristics such as syntactic structure

Outline

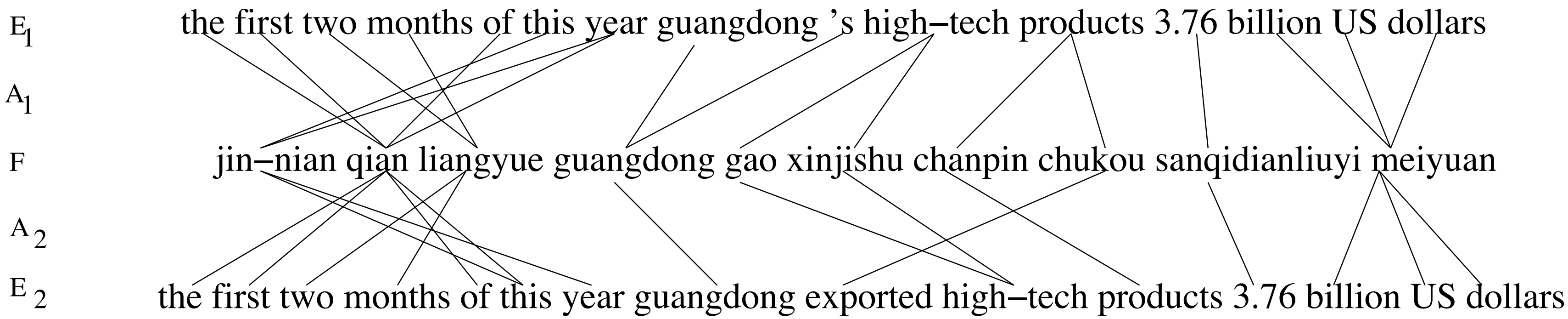
- Hierarchy of Translation Loss functions with different levels of lexical and syntactic Information
- MBR decoding framework
- Experiments
- Conclusions and Future Work

Translation Loss Functions

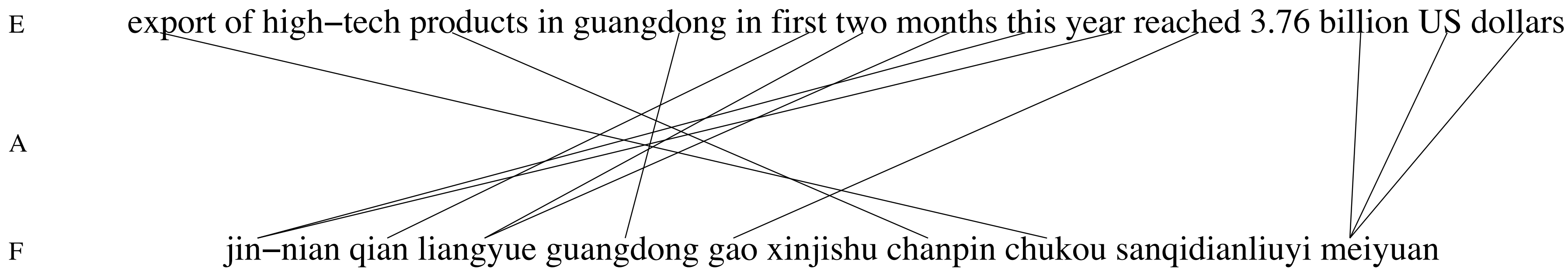
- For a sentence F in the foreign language with parse-tree T_F
  - Hypothesis Translation E' with word alignment A' and parse tree $T_{E'}$
  - Reference Translation E with word alignment A and parse tree T_E
  - Loss Function $L((E, A, T_E), (E', A', T_{E'}); F)$ measures quality of the hypothesis translation against the reference
- Hierarchy of Loss functions
  - Lexical Loss functions : $L(E, E')$
  - Target Language Parse Tree Loss functions : $L(T_E, T_{E'})$
  - Bilingual Parse Tree Loss functions: $L((T_E, A), (T_{E'}, A'); T_F)$

An example

Competing English translations for a Chinese sentence



Reference translation



Lexical Loss Functions

- $L((E, A, T_E), (E', A', T_{E'}); F)$ simplifies to $L(E, E')$
- Loss function depends only on word strings
- Examples
  - Sentence-level BLEU score
    $BLEU(E, E') = exp\left(\sum_{n=1}^{N} log\ \frac{p_n(E,E')}{N}\right) *$ Brev. Penalty(E, E')
    $L_{\text{BLEU}}(E, E') = 1 - BLEU(E, E')$
  - Word Error Rate (WER)
  - Position Independent Word Error Rate (PER)
    Minimum # of edit operations to transform E into any permutation of E'
  - Other examples: NIST score, Precision-Recall Measure (Melamed 2003)

Target Language Parse-Tree Loss Functions

- Information from parse-trees of the two translations
- $L((E, A, T_E), (E', A', T_{E'}); F)$ simplifies to $L(T_E, T_{E'})$
- Examples
  - Tree-edit distance between parse trees
  - String-edit distance between event representation of parse trees (Tang, Luo and Roukos '03)
  - Tree Kernel (Collins '02)
- No experiments involving these loss functions in this talk
- Problem can be simplified if we have a third tree in the foreign language with node-to-node alignments relative to T_E and $T_{E'}$

Bilingual Parse-Tree Loss Functions

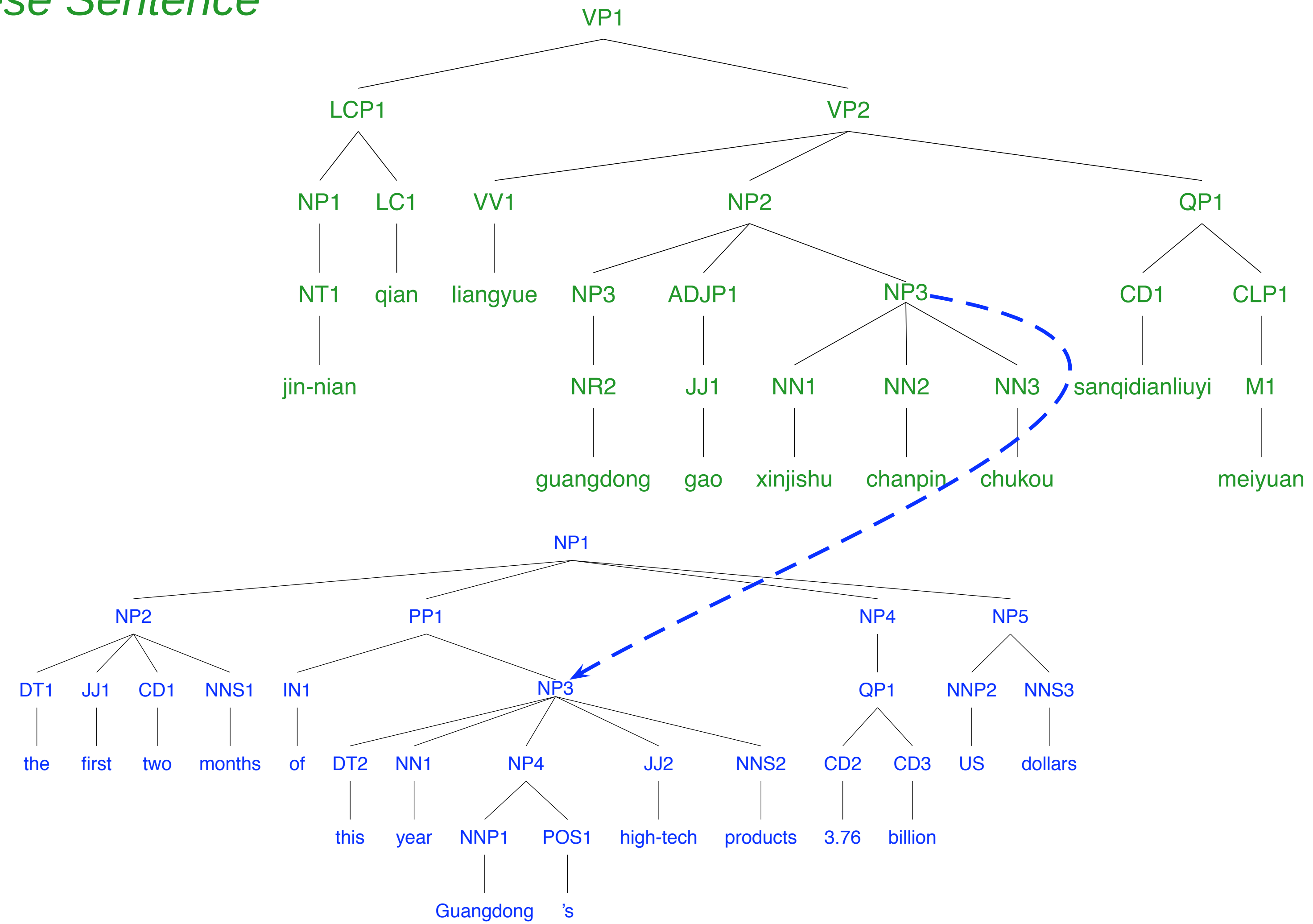
- Word alignments and parse-trees from English and foreign language strings
- $L((E, A, T_E), (E', A', T_{E'}); F)$ simplifies to $L((T_E, A), (T_{E'}, A'); T_F)$

Example BiTree Loss Function

- Alignment of Parse-Trees
  - Use MT word alignments to obtain node-to-node alignments between nodes $n \in T_F$ to nodes $m \in T_E$ and $m' \in T_{E'}$
  - Subtree t_n (in T_F) mapped to t_m (in T_E) and $t'_{m'}$ (in $T_{E'}$)
- Loss Computation between Aligned Parse-Trees
  - $\bar{N}_F$ is the subset of nodes in T_F which have corresponding nodes in both T_E and $T_{E'}$
  - BiTreeLoss$((T_E, A), (T_{E'}, A'); T_F) = \sum_{n \in \bar{N}_F} d(t_m, t'_{m'})$

Bitree Loss Function

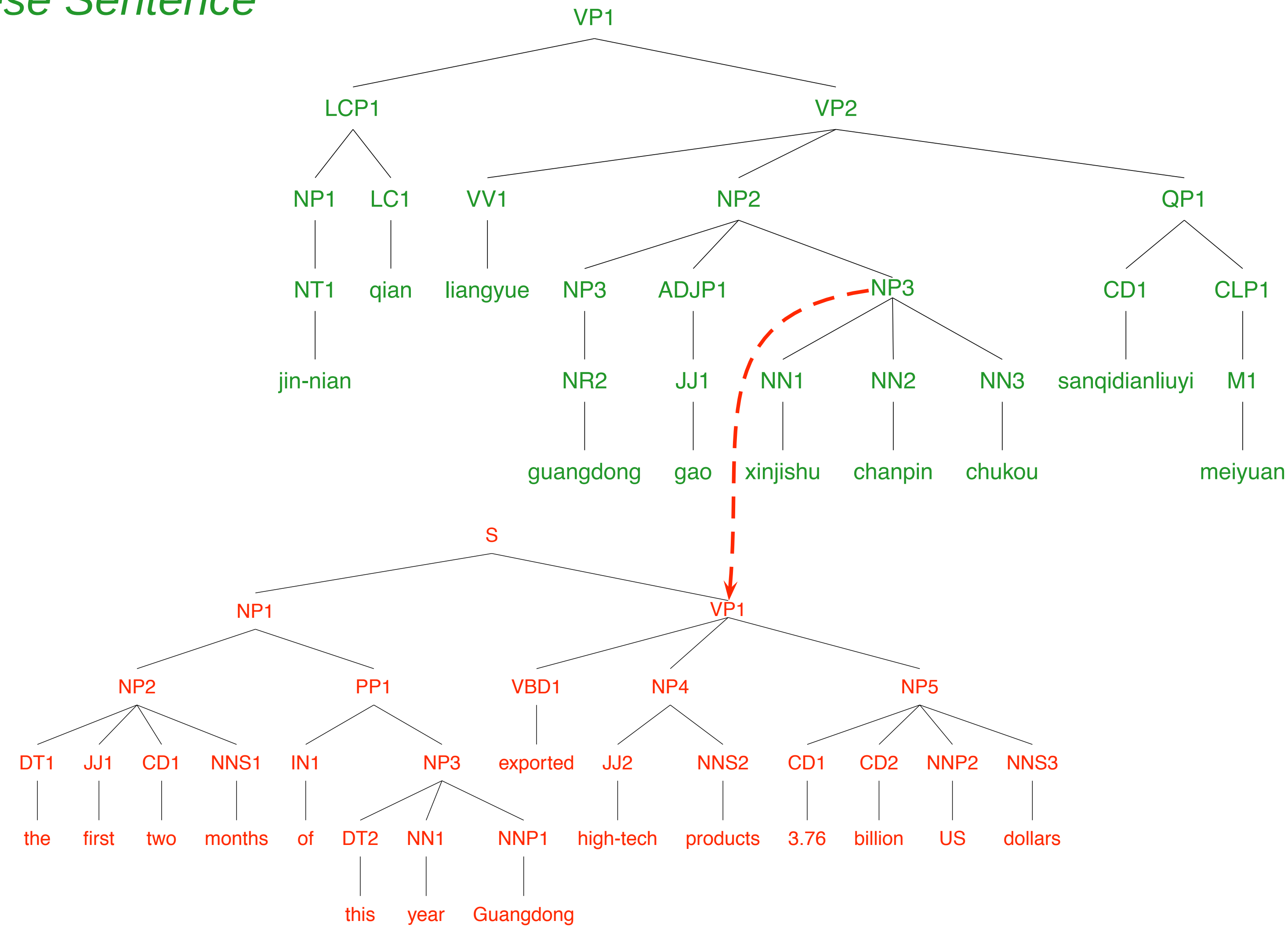
T_F : *Chinese Sentence*



T_1 : *English Hypothesis Translation 1*

Bitree Loss Function

T_F : *Chinese Sentence*



T_2 : *English Hypothesis Translation 2*

Comparison of Loss Functions

Loss Functions	$L(E, E_1)$	$L(E, E_2)$
BLEU (%)	26.4	26.4
WER (%)	70.6	70.6
PER (%)	23.5	23.5
BiTree Error Rate (%)	92.3	65.4

- BLEU, WER and PER are identical
- Parse-trees for the two translations differ substantially and BiTree Loss is quite different for the two translations
- Example of a loss function that can capture properties of translation that string based loss functions are unable to measure

Outline

- Hierarchy of Translation Loss functions with different levels of lexical and syntactic Information
- Minimum Bayes-Risk decoding framework
- Experiments
- Conclusions and Future Work

Minimum Bayes-Risk Decoding

Decoding in Statistical MT as a classification problem

- Map a foreign sentence F into its English translation E' with word alignment A'
  $\delta(F) = (E', A')$
- Given the reference translation (E, A), the Decoder Performance is measured by the Loss Function $L((E, A), \delta(F))$
- Goal: Find the decoder with the best performance over all translations
- Bayes-Risk: Expected Loss of a hypothesis under the true distribution $P(E, A, F)$: $E_{P(E,A,F)}[L((E, A), \delta(F))$
- Decision Rule that minimizes Bayes-Risk is : *MBR Decoder*

$$\delta(F) = \operatorname*{argmin}_{E',A'} \sum_{E,A} L((E, A), (E', A'); F)P(E, A|F)$$

MBR Decoders

- Consensus Translation: Select the hypothesis that is closest to other hypotheses
- MAP decoder is an MBR decoder under 0-1 loss function
- Implementation on an N-best List
  - N-Best List of Translations (E_i, A_i) from a baseline system
  - True distribution $P(E, A|F)$ is approximated using a baseline MT system (translation model and language model)
  - MBR Decision Rule via N-best Rescoring

$$\hat{i} = \underset{i \in \{1,2,\ldots,N\}}{\operatorname{argmin}} \sum_{j=1}^{N} L((E_j, A_j), (E_i, A_i)) P(E_j, A_j|F)$$

$$\delta(F) = (E_{\hat{i}}, A_{\hat{i}})$$

Experiments on Large Data Track of NIST Chinese-to-English MT

- Baseline MT system from JHU Summer Workshop WS '03 group on Syntax for Statistical Machine Translation
- Test Set : 993 sentences from Eval01 + 878 sentences from Eval02
- 1000-best lists for each Chinese sentence

	Performance Metrics			
Decoder	BLEU (%)	mWER(%)	mPER (%)	mBiTree Error Rate(%)
MAP(baseline)	31.2	64.9	41.3	69.0
MBR				
BLEU	31.5	65.1	41.1	68.9
WER	31.3	64.3	40.8	68.5
PER	31.3	64.6	40.4	68.6
BiTree Loss	30.7	64.1	41.1	68.0

Observations

- In most cases MBR decoder under a loss function performs best under the corresponding error metric
- MAP decoder is not optimal in any of the cases
- Performance under BLEU can be improved by using MBR relative to MAP
- Affinity among loss functions
- Useful to tune decoding procedures to the performance criterion of interest

Conclusions

- MBR decoding : Build special purpose decoders from general purpose MT models.
- Applicability to two MT scenarios
  - Given an MT evaluation metric (e.g. BLEU), MBR decoding can improve well trained statistical MT models by tuning translation to the particular evaluation metric
  - Suppose we have desiderata e.g. syntactic well-formedness to incorporate in the baseline MT system
    - Design a loss function to incorporate the desired criterion
    - Use MBR decoding to optimize performance under this loss function
    - Bitree loss function is an example of this type of loss function - we have not yet measured any correlations with human judgements

Outlook and Future Work

- MT evaluation is active area of research.
  - MBR decoding can be used to optimize existing MT systems for new metrics
  - Compensate mismatch between decoding criterion of MT systems and their evaluation criterion
- Loss functions can also incorporate task-based error criteria
  e.g. precision/recall for IR
- Extension of search space of MBR decoders to translation lattices.

Thank you!