A Unified Approach to Minimum Risk Training and Decoding

Abhishek Arun, Barry Haddow and Philipp Koehn

University of Edinburgh

Fifth Workshop on Machine Translation, Uppsala, July 16th 2010
Outline

- Current Approaches to Minimum Risk Decoding
- A Unified Approach
- Markov Chain Monte Carlo for Phrase-based MT
- Minimum risk training
- Optimising corpus BLEU
- Experiments
- Conclusions and Future work
Minimum Risk Decoding in MT

Optimal Decision Rule?
- Find the target sentence which minimises expected risk
  - Equivalently: Maximises expected gain
- Summarised by the following equation

\[ e^* = \arg \max_e \sum_{e'} p(e'|f) \text{Gain}(e', e) \]

- We use \textbf{BLEU} as the gain function
- Referred to as \textbf{Minimum Bayes Risk (MBR) Decoding.}
Current Approaches to MBR Decoding

- First-pass decoder scores translations with linear model
- The scores must be scaled and normalised to give probabilities
  - Scaling requires hyper-parameter search
  - Normalisation requires intractable sum
- MBR Decoding implemented as a list re-ranker
- Feature weights in linear model trained with MERT
  - Non-probabilistic training algorithm
  - Aims to maximise 1-best (MAP) performance
Lattice-Based Approaches

- Represent many hypotheses compactly
- State-of-the-art performance from Lattice MBR
- But
  - Feature weights trained with MERT
  - Biased pruning - May be bad for sparse features
  - Need to approximate BLEU- more hyperparameters
A Unified Approach

| Training         | Decoding        |
|------------------|-----------------|
| Optimise Expected BLEU | Maximise Expected BLEU |

- Objective is differentiable
  - Can use gradient-based optimisation
- Use **Markov Chain Monte Carlo (MCMC)** to estimate:
  - Feature expectations during training - for gradient
  - Expected BLEU during decoding
Benefits of Our Approach

- Maintains a probabilistic formulation throughout
  - Theoretically sound
  - Unbiased estimates
- Avoids dynamic programming so non-local features easier
- Compared to MERT:
  - More stable
  - Generalises better
  - Gives better performance
MCMC Sampler for Phrase-based MT

- Used to draw samples \{ (e_i, a_i) \} from \( p(e, a | f) \)
  - Use the samples to estimate expectations

\[
E(h) \approx \frac{1}{N} \sum_{(e_i, a_i)} h(e_i, a_i, f)
\]

- Transitions \( T_i \) defined by Transition Operators
  - Make small local changes to hypothesis
  - Apply all operators in sequence before collecting sample
MCMC Operators

**RETRANS**
Retranslates one source-target phrase pair

**MERGE-SPLIT**
Operates at an inter-word position. May merge or split segments as appropriate, and retranslate.

**REORDER**
Swaps target position of two source-target phrase pairs
MCMC Example

(a) c’est un résultat remarquable
Initial
\[\text{it is some result remarkable}\]

(b) c’est un résultat remarquable
RETRANS
\[\text{but some result remarkable}\]

(c) c’est un résultat remarquable
MERGE
\[\text{it is a result remarkable}\]

(d) c’est un résultat remarquable
REORDER
\[\text{it is a remarkable result}\]
Minimum Risk Training

Our objective is the expected gain plus an entropic prior

\[ \hat{G} = \sum_{\langle \hat{e}, f \rangle \in D} \left[ \left( \sum_{e, a} p(e, a \mid f) \text{BLEU}_{\hat{e}}(e) \right) + T \cdot H(p) \right] \]

- The temperature \((T)\) starts off high and is gradually reduced.
- This moves from high entropy to low entropy, and helps avoid local maxima
- Known as **Deterministic Annealing (DA)**
- The gradient is calculated using the sampler, and optimisation is by stochastic gradient descent
But we’re optimising sentence BLEU
- And testing with corpus BLEU
To eradicate this mismatch, we propose Corpus Sampling
- Each sample is an aligned translation of the whole corpus
  - Sentence samples are collected for all sentences
  - These are resampled to give corpus samples
  - Now we can optimise corpus BLEU
Corpus Sampling Illustration

**SAMPLE FROM**

**P(e,a | f)**

| f1 | f2 | f3 |
|----|----|----|
| A  | D  | K  |
| B  | E  | L  |
| A  | F  | L  |
| C  | G  | L  |
| B  | H  | M  |

**SAMPLE FROM EMPIRICAL DISTRIBUTION**

| f1 | f2 | f3 |
|----|----|----|
| A  | F  | L  |
| B  | E  | L  |

**Extract Corpus Samples**

| Corpus Sample 1 |  {A, F, L}  |
|-----------------|-------------|
| Corpus Sample 2 |  {B, E, L}  |
Experimental Setup

| NIST Arabic-English | Europarl French-English | Europarl German-English |
|---------------------|-------------------------|-------------------------|
| 300k Sents Train In-Domain Test | 1.4M Sents Train In-Domain Test Out-of-domain Test | 1.4M Sents Train In-Domain Test Out-of-domain Test |

Moses Setup

- Standard phrase extraction pipeline
- Standard features (no lexicalised reordering)
- MERT/Moses for baselines
Effect of deterministic Annealing

- Graphs show heldout performance
- Converges much quicker without DA
- Maximum is lower
- At high entropy, MBR much better than max-derivation
- Advantage reduces with temperature
- We use early stopping to find best weights
## Corpus Sampling vs Sentence Sampling

| Test Set                | Sentence | Corpus |
|-------------------------|----------|--------|
| AR-EN MT05              | 44.6 (0.990) | 44.5 (0.989) |
| FR-EN In-domain         | 32.9 (1.003) | 33.2 (0.997) |
| FR-EN Out-domain        | 19.7 (1.049) | 19.8 (1.041) |
| DE-EN In-domain         | 26.9 (0.987) | 27.8 (0.993) |
| DE-EN Out-domain        | 16.6 (0.975) | 16.6 (0.980) |

- Expected **BLEU** training, MBR decoding
- Table shows **BLEU** and length penalty
- Corpus sampling slightly better
Comparison with Moses Baseline

| Test set      | MERT/Moses | Expected BLEU |
|---------------|------------|---------------|
|               | Best       | σ             | MBR | σ       |
| AR-EN MT05    | 44.5 (lMBR) | 0.12          | 44.5 | 0.14 |
| FR-EN In      | 33.4 (nMBR) | 0.12          | 33.2 | 0.06 |
| FR-EN Out     | 19.5 (nMBR) | 0.12          | 19.8 | 0.05 |
| DE-EN In      | 27.8 (MAP)  | 0.10          | 27.8 | 0.11 |
| DE-EN Out     | 16.0 (lMBR) | 0.30          | 16.6 | 0.12 |

- Compare corpus sampler with best MERT/moses result
  - For sampler, decode with n-best MBR
  - For Moses, best out of MAP, n-best MBR and lattice MBR
- Five runs of expected BLEU, ten runs of MERT, averaged.
Expected Bleu Training, Moses Decoding

| Test Set       | MAP  | nMBR | IMBR | Sampler MBR |
|----------------|------|------|------|-------------|
| AR-EN MT05     | 44.2 | 44.4 | 44.8 | 44.8        |
| FR-EN In       | 33.1 | 33.2 | 33.3 | 33.3        |
| FR-EN Out      | 19.6 | 19.8 | 19.9 | 19.9        |
| DE-EN In       | 27.7 | 27.9 | 28.0 | 28.0        |
| DE-EN Out      | 16.0 | 16.3 | 16.6 | 16.6        |

- We use the best expected BLEU trained weights
- Decoding with Moses (first three columns) or sampler
- Suggests that expected BLEU weights better for IMBR
Conclusions

- Unified Training and Decoding beats or equals MERT/Moses
- Deterministic Annealing (entropic prior) provides better performance
- Corpus sampling provides small gains over sentence sampling
- Expected bleu trained weights more suited to lattice MBR decoding, than MERT weights
- MBR and maximum-translation decoding better than maximum-derivation
Future Work

- Supplement dense features with many sparse features
  - eg. discriminative language models
- Incorporate non-local features
  - eg. long-distance agreement
- Metropolis-Hastings step to efficiently incorporate slow features
  - eg. higher-order language model
Thank you!

Questions?

Code:

https://mosesdecoder.svn.sourceforge.net/svnroot/mosesdecoder/branches/josiah