Improving the Minimum Bayes' Risk Combination of Machine Translation Systems

Jesús González-Rubio, Francisco Casacuberta

{jegonzalez,fcn}@dsic.upv.es

Pattern Recognition and Human Language Technology Group

Universitat Politècnica de València (Spain)

Work supported by the EU 7th Framework program (FP/2007-2013) under the CasMaCat project (gran nº 287576)

Overview

- Introduction
- Minimum Bayes' Risk System Combination
- Dynamic Programming Decoding for MBRSC
- Evaluation
- Conclusions

Introduction

Motivation

- MT technology is still far from human translation quality
- Different MT approaches have complementary strengths and limitations
- Focus on Minimum Bayes' Risk System Combination (MBRSC)
  - Conceptually simple and provide competitive empirical results
- Our contributions:
  - New decoding algorithms based on Dynamic Programming (DP)
  - An MBRSC formulation based on linear BLEU [Tromble et al., 2008]

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

CASMACAT

Minimum Bayes' Risk System Combination

Model and Decision Function

- Weighted ensemble of K probability distributions (translation models)

$$\mathrm{P}(\mathbf{y} \mid \mathbf{x}) = \sum_{k=1}^{K} \alpha_k \cdot \mathrm{P}_k(\mathbf{y} \mid \mathbf{x})$$

- The minimum Bayes' risk classifier for BLEU is given by:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}} \sum_{k=1}^{K} \alpha_k \cdot \underbrace{\left(\sum_{\mathbf{y}' \in \mathcal{Y}} \mathrm{P}_k(\mathbf{y}' \mid \mathbf{x}) \cdot \mathrm{BLEU}(\mathbf{y}, \mathbf{y}') \right)}_{\mathrm{system-specific\ expected\ BLEU}}$$

- Complex decoding problem

Decoding

- Direct implementation has a temporal complexity in $O(\max(|\,\mathbf{y}\,|) \cdot |\,\mathcal{Y}\,|^2)$
- Practical approach: divide decoding into gain computation and search
- Expected BLEU is approximated by BLEU over expected n-gram counts

$$\underbrace{\sum_{\mathbf{y}' \in \mathcal{Y}} \mathrm{P}(\mathbf{y}' \mid \mathbf{x}) \cdot \mathrm{BLEU}(\mathbf{y}, \mathbf{y}')}_{\text{expected BLEU of } \mathbf{y}} \approx \underbrace{\widetilde{\mathrm{BLEU}}\left(\mathbf{y}, \overbrace{\sum_{\mathbf{y}' \in \mathcal{Y}} \mathrm{P}(\mathbf{y}' \mid \mathbf{x}) \cdot \#_{\mathbf{w}}(\mathbf{y}')}^{\text{expected count of } \mathbf{w}} \right)}_{\text{BLEU of } \mathbf{y} \text{ over expected counts}}$$

- Search is implemented as a gradient ascent algorithm
- Final temporal complexity in $O(\max(|\,\mathbf{y}\,|)^2 \cdot |\Sigma| \cdot S)$

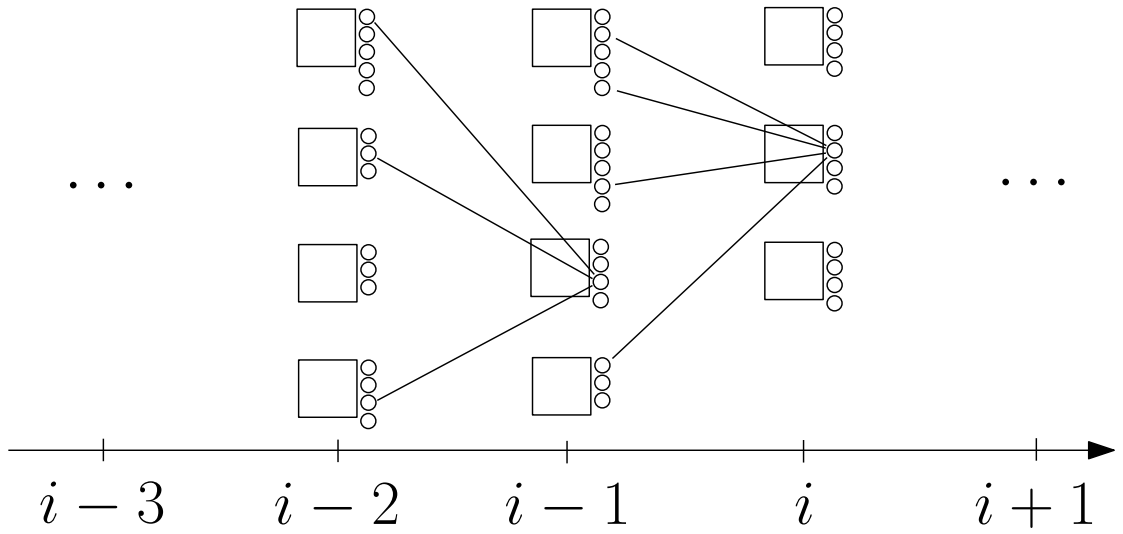
Dynamic Programming Decoding for MBRSC

Dynamic Programming Decoding

- Gradient ascent decoding is sensitive to an initial solution
  - Prone to get stuck in local optima
- Dynamic programming provides a more sophisticated solution
- Basic idea: iterative generation of new translation hypotheses
  - Start with an empty hypothesis
  - Repeatedly generate hypotheses of size $i+1$ by extending hypotheses of size i with one more target word

Dynamic Programming Decoding II

- Graph structure, nodes store hypotheses with the same n-grams



- Unfortunately, the number of nodes is exponential in $|\Sigma|$
- In practice, DP decoding is implemented as a beam search algorithm

Beam Search Implementation

- Key Idea: keep the M best-scoring hypotheses each step
- Breadth-first exploration to avoid repeated computations
- Upper bound (I) to the size of the consensus translations
- Rest score estimation to better compare the potential of each hypothesis
- Final complexity in $O(I^2 \cdot M \cdot D)$, where $D << |\Sigma|$

Dynamic Programming Decoding for Linear BLEU

Why Linear BLEU?

- Count clippings forbid the incremental computation of BLEU
  - We cannot exploit the full potential of the DP framework
- Linear BLEU approximates the logarithm of BLEU [Tromble et al., 2008]

$$\log(\mathrm{BLEU}(\mathbf{y}, \mathbf{y}')) \approx \lambda_0 \cdot |\,\mathbf{y}\,| + \sum_{\mathbf{w} \in \mathcal{W}(\mathbf{y})} \lambda_{\mathbf{w}} \cdot \#_{\mathbf{w}}(\mathbf{y}) \cdot \delta_{\mathbf{w}}(\mathbf{y}') \qquad (1)$$

- Expected linear BLEU gain can be computed incrementally

DP Decoding for Linear BLEU

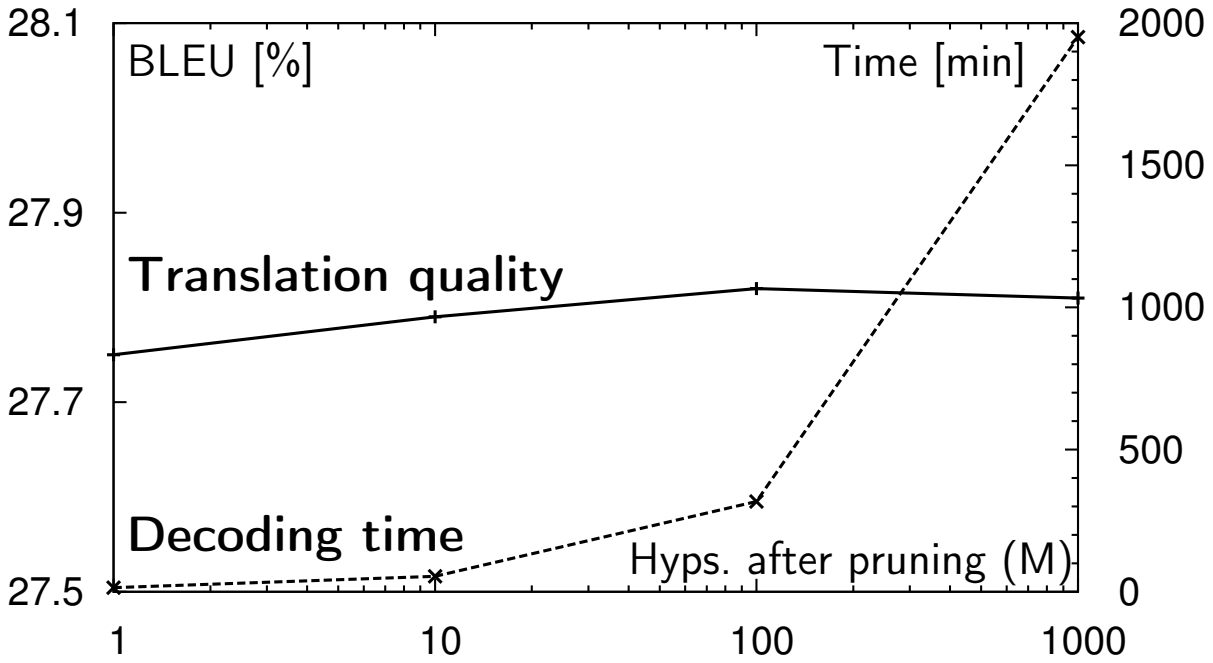
- Search nodes contain hypotheses that share their last three words
  - $|\Sigma|^3$ nodes in the search graph
  - DP decoding can be implemented exactly (no pruning)
- Breadth-first exploration and upper bound (I) for translation size
- No need for rest-score estimation
- Implementation has a complexity in $O(I \cdot |\Sigma|^3 \cdot D)$

Empirical Evaluation

Experimental Setup

- WMT 2009 French-English corpus
- Combine translations of the five systems that submitted N-best lists
  - 450 translations on average for each source sentence
- Maximum length (I) equal to the longest provided translation
- Uniform ensemble weights
  - Controlled environment to compare different setups
  - Initial results showed that weights did not deviate much from uniform

Preliminary Experiments



- We chose $M = 10$ as the pruning value for the next experiments

Translation Quality Results

System setup		BLEU[%]	TER[%]
worst single system		24.8	60.4
best single system		26.4	56.0
Gradient ascent	EC	27.7	55.4
	LB	26.3	59.6
Beam Search	EC	**27.8**	**55.1**
DP	LB	26.8	57.8

EC stands for BLEU over expected counts, and LC stands for linear BLEU

- Scarce quality improvements but better score for 53% of the sentences

Decoding Time Results

- Estimated by the number calls to compute the expected BLEU
  - Factors out potential effects of the particular implementations
- Beam search made ~ 15 million calls (~ 1.3 s. per sentence)
  - Gradient ascent made ~ 20 million calls
- DP-based decoding also improved the efficiency of MBRSC

Analysis of Linear BLEU Results

EC:	we have made great progress .
LB:	we have made great progress . *we have made*
EC:	it seems to be clear that it is better to buy only a phone .
LB:	*to be clear that* it seems to be clear that it is better to buy only a phone .
EC:	i am curious to know if i could see here .
LB:	*am curious to know if* i am curious to know if i could see here .

- The lack of count clippings results in repetitions of common n-grams
  - Explains the observed degradation in translation quality

Conclusions

Conclusions

- DP-based decoding outperformed previous gradient ascent search
  - Better-scoring translations with less temporal complexity
  - However, improvements in translation quality were scarce
- Linear BLEU boosts efficiency but penalizes translation quality
- An extended linear BLEU score may mitigate this effect
  - For example, by including a language model score

Thank you,

questions?

References

R. Tromble, S. Kumar, F. Och, and W. Macherey. Lattice minimum bayes-risk decoding for statistical machine translation. In *Proc. of the Empirical Methods in Natural Language Processing conference*, pages 620–629, 2008.