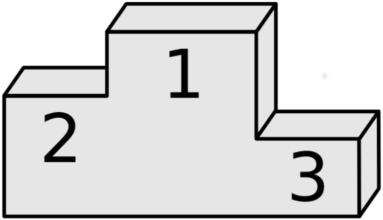
ListNet-based MT Rescoring

Jan Niehues, Quoc Khanh Do, Alexandre Allauzen and Alex Waibel

KIT - Institute for Anthropomatics and Robotics and LIMSI-CNRS



Motivation

- Log-linear model is widely used in SMT
  - Use during decoding
  - Use in MT rescoring
- MT Rescoring
  - Easy and efficient way to integrate of complex models
- Machine learning view
  - Ranking problem
  - Promising approach: ListNet algorithm
- Apply ListNet algorithm to SMT

Related Work

Optimization in Machine translation

- Minimum Error Rate Training (MERT) (Och, 2003)
  - Standard in most machine translation systems
- MIRA (Watanabe et al., 2007; Chiang et al., 2008)
- PRO (Hopkins and May, 2011)
- Expected BLEU (Rosti et al, 2011; He and Deng, 2012)

Ranking in machine learning

- ListNet algorithm (Cao et al., 2007)
- Overview over different ranking algorithms (Chen et al., 2009)

Overview

- Motivation
- ListNet Algorithm
- MT Rescoring
  - MT specific problems
- Evaluation
  - WMT
  - IWSLT - TED

ListNet -Ranking

- Input:
  - List
  - Model score
  - Metric for reference ranking

Hypothesis	Model	Metric
A	7.4	24.4
B	7.8	24.2
C	7.2	24.5
D	7.1	24.1

ListNet -Ranking

- Input:
  - List
  - Model score
  - Metric for reference ranking

Hypothesis	Model	Metric
B	7.8	24.2
A	7.4	24.4
C	7.2	24.5
D	7.1	24.1

According to the model

ListNet -Ranking

- Input:
  - List
  - Model score
  - Metric for reference ranking

Hypothesis	Model	Metric
C	7.2	24.5
A	7.4	24.4
B	7.8	24.2
D	7.1	24.1

According to the metric

ListNet -Ranking

- Input:
  - List
  - Model score
  - Metric for reference ranking

Hypothesis	Model	Metric
B	7.8	24.2
A	7.4	24.4
C	7.2	24.5
D	7.1	24.1

$\Rightarrow$

Hypothesis	Model	Metric
C	7.2	24.5
A	7.4	24.4
B	7.8	24.2
D	7.1	24.1

Aim:
Learn a model to rank like the metric

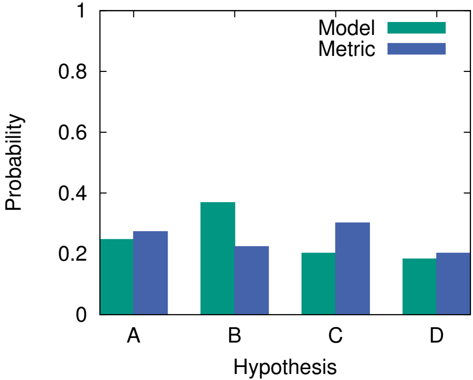
ListNet - Idea

Karlsruhe Institute of Technology

- Define a probability distribution over possible rankings
- Learn model that produces a distribution similar to the one defined by the metric
- Problem: large number of possible rankings
- Define a probability distribution associated to the model ranking based on first ranked object

$$P_s(j) = \frac{\exp(s_j)}{\sum_{k=1}^{n} \exp(s_k)} \tag{1}$$

ListNet - Distribution



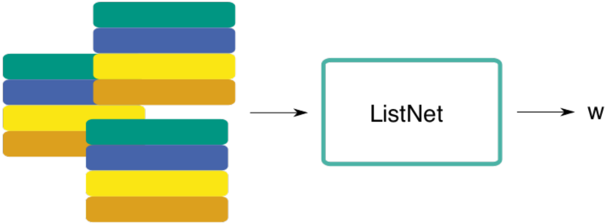
- Minimize cross-entropy difference

Overview

- Motivation
- ListNet Algorithm
- MT Rescoring
    - MT specific problems
- Evaluation
    - WMT
    - IWSLT - TED

MT Rescoring

- Use ListNet to rescore N-Best list
  - Train log-linear model
- Input:
  - N-Best list
  - Additional features
- Learn new weights for log-linear model



Model

- Define probability distribution associated to the model ranking

$$P_s(j) = \frac{\exp(s_j)}{\sum_{k=1}^{n} \exp(s_k)} \qquad (2)$$

- Problem:
  - Many scores are small probabilities
  - Log-probabilities are very small negative values
  - $exp(s)$ calculation may be erroneous
- Feature normalization:
  - Linear transform all features to the range $[-1, 1]$
- Score normalization:
  - Linear transform the final score of the model to the range $[-r, r]$

Metric

- Define probability distribution associated to the reference ranking
- Reference ranking for every sentence needed
- Ranking induced by MT metric
- Sentence-wise MT metric
  - Metric: BLEU+1 (Liang et al. 2006)
  - Smoothed version of BLEU score

$$P_{y^{(i)}}(x_j^{(i)}) = \frac{\exp(\mathrm{BLEU}(x_j^{(i)}))}{\sum_{j'=1}^{n^i} \exp(\mathrm{BLEU}(x_{j'}^{(i)}))} \quad (3)$$

Training

- Minimize cross-entropy difference between model-based and BLEU+1-based probability distribution
  - Use ListNet algorithm to calculate derivation
- Stochastic gradient descent
  - 100,000 batches
  - Batch size of 10

Overview

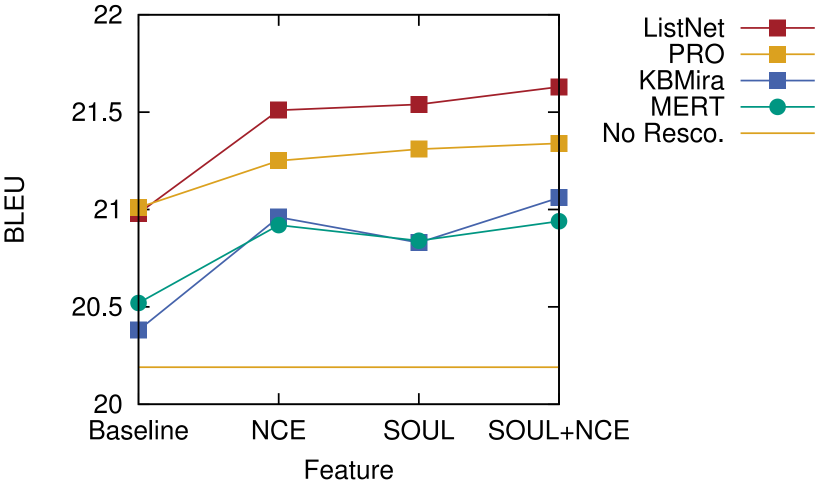
Karlsruhe Institute of Technology

- Motivation
- ListNet Algorithm
- MT Rescoring
    - MT specific problems
- Evaluation
    - WMT
    - IWSLT - TED

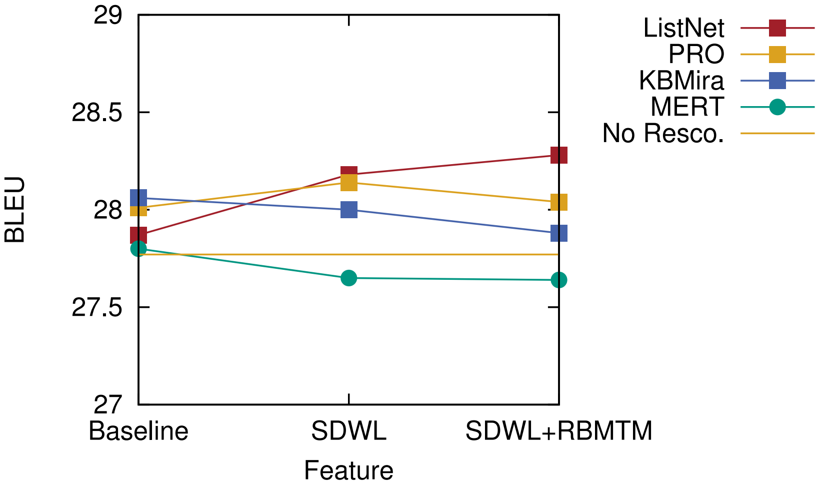
Evaluation

- WMT 2015 EN-DE
  - PBMT System
  - Additional features based on neural network translation models
- WMT 2015 DE-EN
  - PBMT System
  - Additional features using RBM-based translation models and source DWL
- TED 2014 EN-DE
  - Translation of TED talks

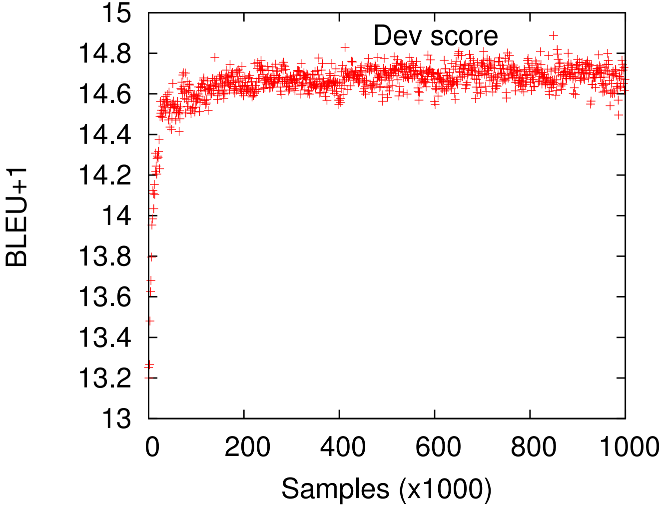
WMT – English to German



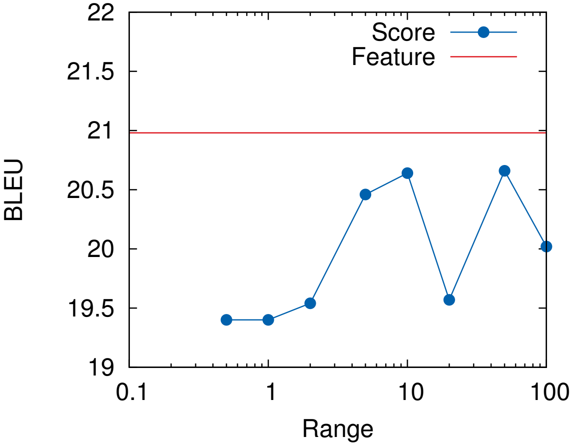
WMT – German to English



Convergence

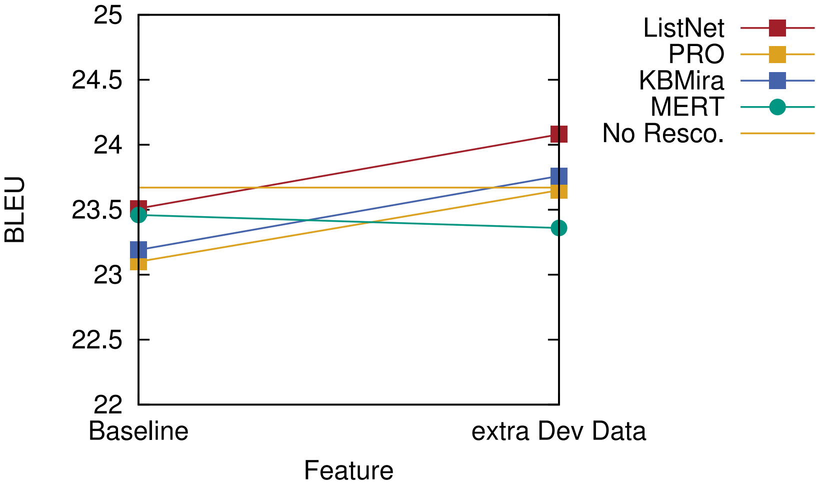


Score normalization



KIT - Institute for Anthropomatics and Robotics and LIMSI-CNRS

TED – English to German



Conclusion

- Presented a new technique to train log-linear model
  - Scale to many features
  - Consider whole list
  - Technique can also be applied to more complex models
- Evaluated using different tasks and languages
  - WMT English – German
  - WMT German – English
  - IWSLT –TED English – German
- Translation quality improvements measured in BLEU score
  - Outperform MERT in all configurations
  - Less prone to overfitting

WMT – English to German

	Baseline		NCE		SOUL		SOUL+NCE	
System	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Baseline		20.19						
MERT	20.63	20.52	21.24	20.92	21.36	20.84	21.36	20.94
KB-MIRA	20.64	20.38	21.51	20.96	21.65	20.83	21.71	21.06
PRO	20.17	**21.01**	21.04	21.25	21.18	21.31	21.14	21.34
ListNet	19.95	20.98	21.00	**21.51**	21.02	**21.54**	21.14	**21.63**

WMT – German to English

	Baseline		SDWL		SDWL+RBMTM	
System	Dev	Test	Dev	Test	Dev	Test
Baseline		27.77				
MERT	28.18	27.80	28.24	27.65	28.23	27.64
KB-MIRA	28.23	**28.06**	28.18	28.00	28.00	27.88
PRO	27.38	28.01	27.56	28.14	28.68	28.04
ListNet	28.00	27.87	27.89	**28.18**	27.94	**28.28**

TED – English to German

System	Baseline Dev	Baseline Test	extra Dev Data Dev	extra Dev Data Test
Baseline		23.67		
MERT	27.69	23.46	25.63	23.36
KB-MIRA	27.47	23.19	25.65	23.76
PRO	26.67	23.10	25.00	23.65
ListNet	27.37	**23.51**	25.49	**24.08**