N-best Reranking by Multitask Learning

Kevin Duh, Katsuhito Sudoh, Hajime Tsukada, Hideki Isozaki, Masaaki Nagata

NTT Communication Science Laboratories
Our Goal

Incorporate *millions of features* into MT without overfitting!
Main Ideas

1. Some features are just very sparse

2. Overfitting is inevitable for conventional training

3. But multitask learning can help by discovering lower-dimensional feature space
Outline

1. **WHY**: Motivations
   - The challenge of sparse features
2. **HOW**: Proposed training algorithm
3. **WHAT**: Reranking experiments
4. Conclusions
Background

- Goal: given $f$, score translations $e$ based on:

$$\hat{e} = \arg \max_{e \in N(f)} w^T \cdot h(e, f)$$

- We’re interested in systems employing millions of features

Note: Here we focus on N-best reranking but extension to 1st-pass training is possible
Sparse features for MT

- [Watanabe2007] proposed heavily-lexicalized features, e.g.

\[
h(e, f) = \begin{cases} 
1 & \text{if foreign word "Monsieur" and English word "Mr." co-occur in } e, f \\
0 & \text{otherwise}
\end{cases}
\]

\[
h(e, f) = \begin{cases} 
1 & \text{if English trigram "Mr. Smith said" occurs in } e \\
0 & \text{otherwise}
\end{cases}
\]

Never used if input sentence does not contain "Monsieur"

Many reordering possibilities → many potential features "said Smith Mr.", "Smith Mr. said", ..
Why does overfitting occur?

Because there exist very little feature overlap between any two N-best lists.
Visualizing feature overlap (or lack thereof)

**Feature Growth Rate**
Definition: ratio of new-feature to active feature
In the limit, 45% of active features are never seen before!

**Conditions for this long-tail behavior**
- Feature templates are heavily-lexicalized
- Input (f) has high variability
- Output (e) has high variability
Outline

1. WHY: Motivations
2. HOW: Proposed training algorithm
   - What is multitask learning
   - How N-best can be viewed as multitask problem
3. WHAT: Reranking experiments
4. Conclusions
What is Multitask Learning?

A set of machine learning techniques for exploiting heterogeneous training data

- Contrasts with i.i.d. assumption of traditional setup
- Instead assumes some underlying commonality
Examples of “Tasks”

- Multiple domains:
  - News
  - Europarl
  - Web

- Multiple related problems:
  - Hello (English)
  - God dag (Swedish)
  - Guten Tag (German)
  - Goddag (Danish)
N-bests with sparse features can be viewed as a Multitask problem

Train a single weight $w$ ?
Is data i.i.d. across N-bests?

NO!
Data is heterogeneous. Treat as multitask!
Our Meta-Algorithm

STEP 1: Train weights independently for each N-best
STEP 2: Find commonality among weights (and iterate)
STEP 3: Train conventional reranker on discovered common features

N-best Reranking by Multitask Learning
L1/L2 Joint Regularization
(one example multitask learning method)

\[
\arg \min_{\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^I} \sum_{i=1}^{I} \text{Loss}(\mathbf{w}^i, \text{nb}est^i) + \lambda \| \mathbf{W} \|_{1,2}
\]

\|\mathbf{W}\|_{1,2} \text{ computed by}
1. Stacking the weights into a matrix
2. Take L2 norm on columns, then L1 norm on result
Effect: encourage sharing of features

Exercise: which is the better solution?

\[
\mathbf{W}_a : \begin{bmatrix}
4 & 0 & 0 & 3 \\
0 & 4 & 3 & 0 \\
4 & 4 & 3 & 3 \rightarrow 14
\end{bmatrix} \quad \mathbf{W}_b : \begin{bmatrix}
4 & 3 & 0 & 0 \\
0 & 4 & 3 & 0 \\
4 & 5 & 3 & 0 \rightarrow 12
\end{bmatrix}
\]
Many multitask methods are available!

**Joint Regularization:**
- L1/L2 [Obozinski09, Argyriou08]
- L1/L-infinity [Quattoni09]

**Bayesian Prior:** [Daume09, Finkel09]
\[ \sum_i ||w^i - w^{avg}||_2 \]

**Shared Feature Subspace:**
- SVD-based [Ando05]
- Neural network [Caruana97]
- Deep Learning [Collobert08]
Outline

1. WHY: Motivations
2. HOW: Proposed training algorithm
3. WHAT: Reranking experiments
   - Data
   - Results
4. Conclusions
English→Japanese translation of PubMed abstracts

**Abstract**

BACKGROUND: Up to 80% of thyroid nodules with an indeterminate diagnosis on fine-needle aspiration (FNA) (e.g., "suspicious for follicular neoplasm") prove to be benign at the time of surgical resection. Analytic tests in current use are limited in their ability to improve the preoperative detection of malignant follicular thyroid nodules. Studies using paraffin-embedded tissue have indicated that high mobility group AT-hook 2 (HMGA2) overexpression is present in a high percentage of malignant thyroid neoplasms but not in benign thyroid neoplasms. In the current study, the ability of HMGA2 overexpression analysis to preoperatively distinguish benign from malignant thyroid nodules by reverse transcriptas...
Experiment comparison

What is best feature representation?

Baselines:
1. Original Feature Representation
2. Feature selection by L1 regularization

Features discovered by Multitask:
1. Joint Regularization (L1/L2)
2. Shared Subspace (SVD)

Specifics:
- Base reranker is RankSVM, similar to [Shen04]
- Original: 2.4 million features
- Tune multitask feature dimension: {250, 500, 1000}
## Results

<table>
<thead>
<tr>
<th>Feature Representation</th>
<th>No. of features</th>
<th>Train BLEU</th>
<th>Test BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>First pass system features</td>
<td>20</td>
<td>29.5</td>
<td>28.5</td>
</tr>
<tr>
<td>Baseline 1: Original Sparse Features</td>
<td>2.4M</td>
<td>36.9</td>
<td><strong>28.6</strong></td>
</tr>
<tr>
<td>Baseline 2: Original, with L1 regularization</td>
<td>1200</td>
<td>36.5</td>
<td>28.5</td>
</tr>
<tr>
<td>Oracle</td>
<td>--</td>
<td>36.9</td>
<td>36.9</td>
</tr>
<tr>
<td>Multitask 1: Joint Regularization (L1/L2)</td>
<td>250</td>
<td>31.8</td>
<td><strong>28.9</strong></td>
</tr>
<tr>
<td>Multitask 2: Shared Subspace (SVD)</td>
<td>1000</td>
<td>32.9</td>
<td><strong>29.1</strong></td>
</tr>
<tr>
<td>Feature Threshold (occurs in 10+ lists)</td>
<td>60k</td>
<td>35.8</td>
<td><strong>29.0</strong></td>
</tr>
<tr>
<td>+ Multitask 1: Joint Regularization</td>
<td>60.25k</td>
<td>36.1</td>
<td><strong>29.4</strong></td>
</tr>
<tr>
<td>+ Multitask 2: Shared Subspace</td>
<td>61k</td>
<td>36.2</td>
<td><strong>29.6</strong></td>
</tr>
</tbody>
</table>

Improvements in **red** are statistically significant by bootstrap test (p<0.05)
Outline

1. WHY: Motivations
2. HOW: Proposed training algorithm
3. WHAT: Reranking experiments
4. Conclusions (2 slides)
Contributions

- N-best Lists with sparse features may be cast as multitask problem
- Proposed meta-algorithm uses multitask methods to learn better features for reranking

![Diagram showing feature histograms and weights](image.png)

N-best Reranking by Multitask Learning
Final Words

MORE FEATURES IS THE WAY TO GO:
Translation is a delicate process requiring many fine-grained knowledge

But we must avoid overfitting:
1. Careful definition of features: e.g. [Chiang09,Marton08]
2. Feature mining
   [This work]
Thanks! Questions? Suggestions?

- **Citations:**
  - [Ando05]: A framework for learning predictive structures from multiple tasks, JMLR
  - [Argyriou08]: Convex multitask feature learning, MLJ
  - [Chiang09]: 11,001 new features for SMT, NAACL
  - [Collobert08]: A unified architecture for NLP: deep neural networks with multitask learning, ICML
  - [Daume09]: Bayesian multitask learning with latent hierarchies, UAI
  - [Marton08]: Soft syntactic constraints for hierarchical phrase based translation, ACL
  - [Finkel09]: Hierarchical Bayesian domain adaptation, NAACL
  - [Quattoni09]: An efficient projection for L1-Linf regularization, ICML
  - [Shen04]: Discriminative reranking for MT, NAACL
  - [Watanabe07]: Online large margin training for SMT, EMNLP

- **Acknowledgments:**
  - We thank Jun Suzuki, Shinji Watanabe, Albert Au Yeung and the three reviewers for their valuable comments!
<table>
<thead>
<tr>
<th>Feature Representation</th>
<th>#Feature</th>
<th>Train BLEU</th>
<th>Test BLEU</th>
<th>Test PER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(baselines)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First pass</td>
<td>20</td>
<td>29.5</td>
<td>28.5</td>
<td>38.3</td>
</tr>
<tr>
<td>All sparse features (Main baseline)</td>
<td>2.4M</td>
<td>36.9</td>
<td>28.6</td>
<td>38.2</td>
</tr>
<tr>
<td>All sparse features w/ $\ell_1$ regularization</td>
<td>1200</td>
<td>36.5</td>
<td>28.5</td>
<td>38.6</td>
</tr>
<tr>
<td>Random hash representation</td>
<td>4000</td>
<td>33.0</td>
<td>28.5</td>
<td>38.2</td>
</tr>
<tr>
<td><strong>(multitask learning)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsupervised FeatureSelect</td>
<td>500</td>
<td>32.0</td>
<td><strong>28.8</strong></td>
<td>37.7</td>
</tr>
<tr>
<td>Joint Regularization</td>
<td>250</td>
<td>31.8</td>
<td><strong>28.9</strong></td>
<td>37.5</td>
</tr>
<tr>
<td>Shared Subspace</td>
<td>1000</td>
<td>32.9</td>
<td><strong>29.1</strong></td>
<td>37.3</td>
</tr>
<tr>
<td><strong>(combination w/ high-frequency features)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) Feature threshold $x &gt; 100$</td>
<td>3k</td>
<td>31.7</td>
<td>27.9</td>
<td>38.2</td>
</tr>
<tr>
<td>(b) Feature threshold $x &gt; 10$</td>
<td>60k</td>
<td>35.8</td>
<td>29.0</td>
<td>37.9</td>
</tr>
<tr>
<td>Unsupervised FeatureSelect + (b)</td>
<td>60.5k</td>
<td>36.2</td>
<td><strong>29.3</strong></td>
<td>37.6</td>
</tr>
<tr>
<td>Joint Regularization + (b)</td>
<td>60.25k</td>
<td>36.1</td>
<td><strong>29.4</strong></td>
<td>37.5</td>
</tr>
<tr>
<td>Shared Subspace + (b)</td>
<td>61k</td>
<td>36.2</td>
<td><strong>29.6</strong></td>
<td>37.3</td>
</tr>
<tr>
<td>Oracle (best possible)</td>
<td>–</td>
<td>36.9</td>
<td>36.9</td>
<td>33.1</td>
</tr>
</tbody>
</table>
Open Questions

- Interactive feature engineering?
- Different partition of tasks?
- Multitask on lattices or larger N-bests?
- Comparison to online learning?
A Bayesian perspective

1\textsuperscript{st} Pass Decoder $P(e|f)$ generates data conditioned on $f$

- $f$ is task-specific “parameter”
- $P(e|f)$ is common across tasks

\[
P(e|f) \rightarrow P(e|f=f^1) \rightarrow \text{Nbest List 1}
\]

\[
P(e|f) \rightarrow P(e|f=f^2) \rightarrow \text{Nbest List 2}
\]

\[
P(e|f) \rightarrow P(e|f=f^3) \rightarrow \text{Nbest List 3}
\]