# The University of Washington Machine Translation System for IWSLT 2006

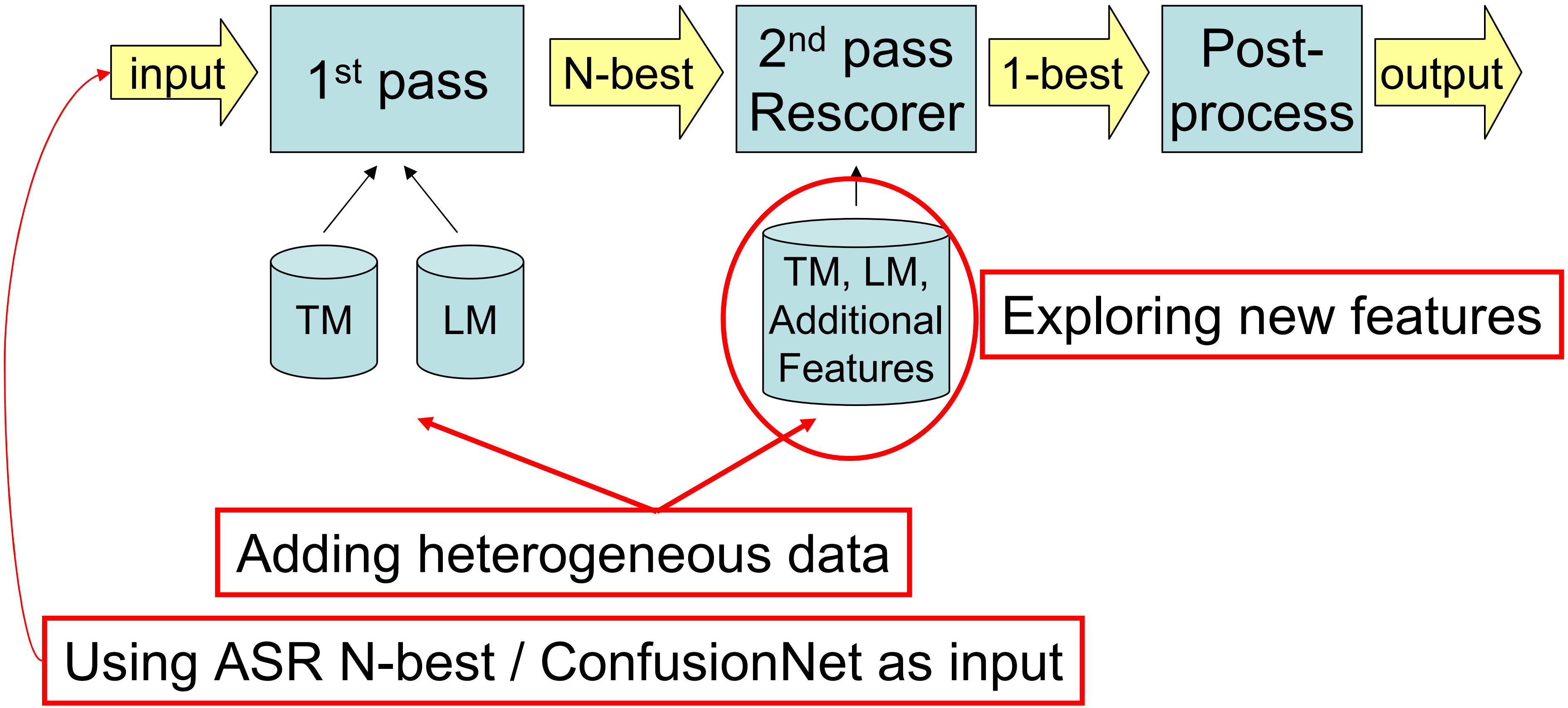
Katrin Kirchhoff, Kevin Duh, Chris Lim

{katrin,duh,chrislim}@ee.washington.edu

University of Washington, Seattle

# System Overview

- Multi-pass phrase-based statistical MT system



# Outline

1. Basic System & Data
   - Data
   - 1st-pass system & features
2. 2nd-pass Rescoring (novel features)
3. Adding heterogeneous data
4. Using ASR N-best / Confusion networks
5. Official results and conclusions

# Data

- **Task: Italian-English open-data track**
  - **Input conditions: ASR-Output & Corrected transcriptions**

- TRAIN SET:
  - BTEC training data + devset1,2,3 (190K words)
  - Europarl (European parliamentary proceedings)
    - (17M words) – for translation model
  - Fisher (Conversational telephone speech)
    - (2.3M words) – for 2nd pass language models

  Additional heterogeneous data

- DEV SET:
  - devset4 – 350 sentences (to optimize 2nd-pass rescorer)
- HELD-OUT SET:
  - devset4 – 139 sentences

# First-Pass Translation System

- Log-linear model:

$$e^{*} = \arg\max_{e} p(e \mid f) = \arg\max_{e} \{\sum_{k=1}^{K} \lambda_k \phi_k(e, f)\}$$

- Weights optimized on BLEU (minimum error rate training)
- Pharaoh decoder w/ monotone decoding
- 9 Features:
  - 2 phrase-based translation scores
  - 2 lexical translation scores
  - BTEC/Europarl data source indicator feature
  - word transition probability
  - phrase penalty
  - distortion penalty
  - language model score (3gram w/ KN smoothing, trained on BTEC)

# Translation models

- 2 separate BTEC & Europarl phrase tables
  - Run GIZA++ and obtain heuristic alignments separately for each corpus
  - Decoder uses both phrase tables, without re-normalization of probabilities

**Example:**

$$\left.\begin{array}{l} P(e1|f1) = 0.4 \\ P(e2|f1) = 0.6 \end{array}\right\} \text{From BTEC}$$

$$\left.\begin{array}{l} P(e1|f1) = 0.1 \\ P(e3|f1) = 0.9 \end{array}\right\} \text{From Europarl}$$

- An additional binary feature indicates the data source

# Outline

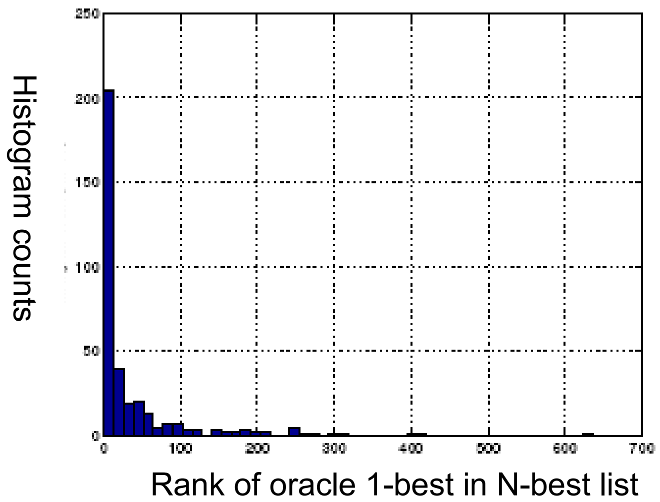
1. Basic System & Data
   - Data
   - 1st-pass system & features
   - Postprocessing
2. 2nd-pass Rescoring (novel features)
3. Adding heterogeneous data (Europarl, Fisher)
4. Using ASR N-best / Confusion networks
5. Official results and conclusions

# $2^{nd}$-pass Rescoring model

- Rescore N-best lists (N=2000max)
- Log-linear model, weights trained by downhill simplex to optimize BLEU
- 14 Features
  - 9 $1^{st}$-pass model scores
  - 4-gram language model score
  - POS 5-gram score [mxpost tagger]
  - Rank in N-best list
  - Factored language model score ratio
  - Focused language model score

# Rank in N-best list ($2^{nd}$-pass feature)

- Idea1: Leverage $1^{st}$-pass decoder rankings in N-best




- Idea2: Hypotheses with same surface string should be tied together

**Rank feature**
- indicates rank of hypothesis in N-best
- ties together identical surface strings

**Example N-best list**

1. The store is open today rank=1
2. The store is open today rank=1
3. The shop is open now rank=2
4. The store is open today rank=1
5. The store it is open rank=3

# Factored Language Model Ratio (2nd-pass feature)

- Factored LM: flexible framework for incorporating diverse information (e.g. morphology, POS) [Bilmes&Kirchhoff03]
  - We model $P(word_t|word_{t-1},pos_{t-1},cluster_{t-1})$
    & various backoffs e.g. $P(word_t|pos_{t-1},cluster_{t-1})$, $P(word_t|word_{t-1})$

- Data-driven FLM backoff selection [Duh&Kirchhoff04]
  - Use a Genetic Algorithm search
  - FLM1: optimize on N-best oracle 1-best sentences
  - FLM2: optimize on N-best oracle worst sentences

- Feature score:

$$\frac{\text{logprob}\{FLM_1(e)\}}{\text{logprob}\{FLM_2(e)\}}$$

  - Log-likelihood ratio: discriminate between good vs. bad sentences

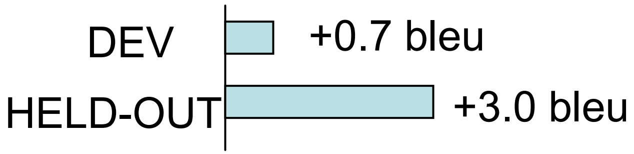
# Focused LM (2nd-pass feature)

- Motivation: LM trained on BTEC (BTEC+Fisher) *wastes probability mass* on words that never occur in the N-best list.
- Solution: train restricted-vocabulary n-grams

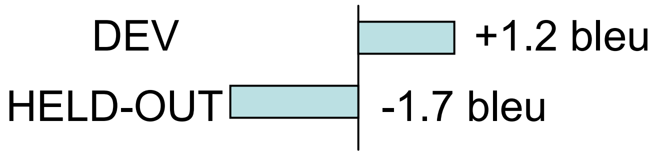
- During N-best optimization:
  1. Collect vocabulary from N-best lists (DEV set)
  2. Train n-gram on BTEC with restricted vocabulary
  3. Generate scores and optimize feature weight
- During evaluation:
  1. Collect vocabulary from N-best lists (EVAL set)
  2. Train *new* n-gram on BTEC with restricted vocabulary
  3. Generate scores for rescoring

- BIG Assumption: optimal feature weight in training is suitable in testing

LM vs. Focused LM (ASR-output)




LM vs. Focused LM (correct trans.)



# Rescoring Results on DEV set

| Correct transcription task | #f | BLEU | PER |
|---|---|---|---|
| Rescoring w/ 1st-pass features | 9 | 44.8 | 30.8 |
| +4gram | 10 | 44.9 | 31.0 |
| +FLM | 10 | 45.0 | 31.4 |
| +focus | 10 | 45.1 | 31.6 |
| +pos | 10 | 45.9 | 30.8 |
| +rank | 10 | 46.8 | 28.5 |

Observations:

-Rank is the strongest feature

-Combination of 14 features outperforms 1st-pass

| ASR-output task | #f | BLEU | PER |
|---|---|---|---|
| Rescoring w/ 1st-pass features | 9 | 34.6 | 39.6 |
| Rescoring w/ ALL FEATURES | 14 | 37.0 | 37.8 |

# Outline

1. Basic System & Data
   - Data
   - $1^{st}$-pass system & features
   - Postprocessing
2. $2^{nd}$-pass Rescoring (novel features)
3. Adding heterogeneous data (Europarl, Fisher)
4. Using ASR N-best / Confusion networks
5. Official results and conclusions

# Adding Europarl to 1st-pass Translation Model (1/2)

- *Does adding Europarl improve translation models, despite domain/style difference?*
- Answer:
  - Yes, for correct transcription task
  - No, for ASR-output task

# Adding Europarl to 1st-pass Translation Model (1/2)

- *Does adding Europarl improve translation models, despite domain/style difference?*
- Answer:
  - Yes, for correct transcription task
  - No, for ASR-output task

**Phrase coverage (%) on DEV**
**[correct transcription task]**

| | BTEC | Europarl | Both |
|---|---|---|---|
| 1 | 84.0 | 88.3 | 94.0 |
| 2 | 40.8 | 48.1 | 60.1 |
| 3 | 13.6 | 11.9 | 20.1 |
| 4 | 3.4 | 1.5 | 4.5 |
| 5 | 1.1 | 0.2 | 1.3 |

**1st-pass translation result on DEV**
**[correct transcription task]**

| | BLEU(%) | PER |
|---|---|---|
| BTEC | 44.5 | 29.9 |
| Both | **46.8** | **28.0** |

# Adding Europarl to 1st-pass Translation Model (2/2)

- *Does adding Europarl improve translation models, despite domain/style difference?*
- Answer:
  - Yes, for correct transcription task
  - No, for ASR-output task

**Phrase coverage (%) on DEV**
**[ASR-output task]**

| | BTEC | Europarl | Both |
|---|---|---|---|
| 1 | 84.0 | 87.7 | 94.6 |
| 2 | 38.9 | 43.0 | 54.7 |
| 3 | 13.6 | 9.9 | 19.1 |
| 4 | 4.2 | 1.0 | 4.9 |
| 5 | 1.4 | 0.2 | 1.6 |

**1st-pass translation result on DEV**
**[ASR-output task]**

| | BLEU(%) | PER |
|---|---|---|
| BTEC | **36.5** | 38.0 |
| Both | 35.4 | **37.3** |

# Adding Fisher to $2^{nd}$-pass Language Models

- *Does additional conversational-style Fisher data improve (1) 4gram LM, (2) POS LM, (3) Focus LM?*
- Answer:
  - No, in general
  - Yes, for Focus LM in correct transcription task (BLEU only)
  - Yes, for POS LM in ASR-output task

**$2^{nd}$-pass translation result on DEV**
**[correct transcription task]**

| | BLEU | PER |
|---|---|---|
| 4gram LM | 44.9 | 31.0 |
| + Fisher | 44.8 | 31.0 |
| POS LM | 45.8 | 30.8 |
| + Fisher | 45.9 | 30.8 |
| Focus LM | 44.4 | 31.3 |
| + Fisher | **45.1** | 31.6 |

**$2^{nd}$-pass translation result on DEV**
**[ASR-output task]**

| | BLEU | PER |
|---|---|---|
| 4gram LM | 34.3 | 39.2 |
| + Fisher | 34.1 | 39.6 |
| POS LM | 35.4 | 40.2 |
| + Fisher | **35.7** | **40.0** |
| Focus LM | 35.2 | 39.8 |
| + Fisher | 34.3 | 40.9 |

# Outline

1. Basic System & Data
    - Data
    - $1^{st}$-pass system & features
    - Postprocessing
2. $2^{nd}$-pass Rescoring (novel features)
3. Adding heterogeneous data (Europarl, Fisher)
4. Using ASR N-best / Confusion networks
5. Official results and conclusions

# ASR-outputs for machine translation

1. ASR 1-best → M-best translation hypotheses **Official submission**
2. ASR N-best → NxM-best translation hypotheses

3. Confusion Networks 1-best

ASR N-best → Confusion Networks → ConfNet 1-best → $1^{st}$-pass decoder → M-best Translations

- Idea: 1-best drawn from ConfusionNet may be more accurate than ASR 1-best
- [Post-evaluation] Significant DEV set improvement over ASR 1-best (37.0 vs. 38.0 BLEU)

# Outline

1. Basic System & Data
   - Data
   - $1^{st}$-pass system & features
   - Postprocessing
2. $2^{nd}$-pass Rescoring (novel features)
3. Adding heterogeneous data (Europarl, Fisher)
4. Using ASR N-best / Confusion networks
5. Official results and conclusions

# Official Results, (Rank)

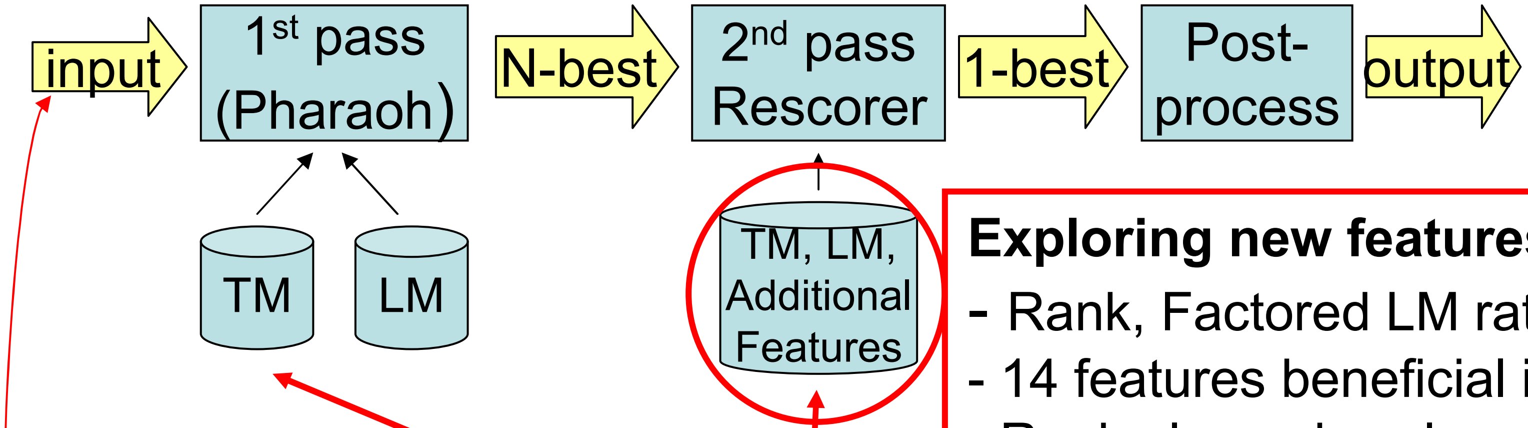
| | BLEU | NIST | METEOR | WER | PER |
|---|---|---|---|---|---|
| **Correct Transcription Task** | | | | | |
| Official | 35.43 (2nd) | 8.19 (1st) | 70.17 (1st) | 48.34 | 38.92 |
| No case/punc | 42.06 (1st) | 9.24 (1st) | 70.19 (1st) | 42.86 | 31.75 |
| **ASR-Output Task** | | | | | |
| Official | 27.87 (2nd) | 6.93 (1st) | 58.53 (1st) | 55.87 | 46.76 |
| No case/punc | 31.68 (2nd) | 7.69 (1st) | 58.53 (1st) | 53.17 | 42.11 |

Summary of submitted system:
1st pass Pharoah decoder
- Monotone decoding
- Translation table uses additional Europarl data

2nd pass Rescorer
- 14 features (incl. N-best rank, Factored LM, Focus LM)

Input for ASR-Output Task: 1-best ASR hypothesis

# Conclusions




**Exploring new features:**
- Rank, Factored LM ratio, Focus LM
- 14 features beneficial in combination
- Rank alone gives large improvements

**Adding heterogeneous data (Europarl, Fisher)**
- Europarl helps TM for correct transcription task
- Fisher did not help LM in general

**Using ASR N-best / ConfusionNet as input**
- Direct translation of N-best not useful
- Confusion network 1-best is promising

# THANKS!

Questions, suggestions, comments?

woof! ワン！bau!

**UW Husky**