# Continuous Space Language Models for the IWSLT 2006 Task

Holger Schwenk
Marta R. Costa-jussà and José A. R. Fonollosa

LIMSI-CNRS, France
schwenk@limsi.fr

UPC, Spain
mruiz,adrian@gps.tsc.upc.edu

November, 27 2006

# Plan



1. **Context and motivation**
2. **Continuous space language model**
3. **Baseline SMT systems**
4. **Experimental evaluation on the IWSLT'06 tasks**
5. **Conclusion and perspectives**

# Introduction



## Context of this work

- BTEC task of IWSLT 2006
- Statistical MT systems rely on representative resources
- Resources to train SMT systems are very limited (40k sentences bitexts, 320k words for LM)

$\Rightarrow$ Need for techniques to take better advantage of the available resources

## Language modeling for SMT

- Most systems use *n*-gram word or class back-off LMs
- Language model adaptation [CMU, IWSLT'05]
- Factored LMs [Kirchoff, ACL wshop'05], syntax-based LMs [Charniak, MT Summit'03]

# Introduction



## Context of this work

- BTEC task of IWSLT 2006
- Statistical MT systems rely on representative resources
- Resources to train SMT systems are very limited (40k sentences bitexts, 320k words for LM)

$\Rightarrow$ Need for techniques to take better advantage of the available resources

## Language modeling for SMT

- Most systems use *n*-gram word or class back-off LMs
- Language model adaptation [CMU, IWSLT'05]
- Factored LMs [Kirchoff, ACL wshop'05], syntax-based LMs [Charniak, MT Summit'03]

## Theoretical Drawbacks of Back-off LM

- Words are represented in a high-dimensional discrete space
- Probability distributions are not smooth functions
- Any change of the word indices can result in an arbitrary change of LM probability

$\Rightarrow$ True generalization is difficult to obtain

## New Approach [Bengio, NIPS'01]:

- Project word indices onto a continuous space and use a probability estimator operating on this space
- Probability functions are smooth functions and better generalization can be expected

## Theoretical Drawbacks of Back-off LM

- Words are represented in a high-dimensional discrete space
- Probability distributions are not smooth functions
- Any change of the word indices can result in an arbitrary change of LM probability

$\Rightarrow$ True generalization is difficult to obtain

## New Approach [Bengio, NIPS'01]:

- Project word indices onto a continuous space and use a probability estimator operating on this space
- Probability functions are smooth functions and better generalization can be expected

# Continuous Space Language Models
## Introduction



### Application of Continuous Space Language Model

- Very successful in LVCSR
- Initial experiments with a word-based SMT system [Schwenk, ACL'06]

### Cooperation with UPC

- First application of the CSLM to a state-of-the-art SMT system
- *n*-best list rescoring of UPC's phrase and Ngram-based system
- All four languages are considered (translation of Mandarin, Japanese, Arabic and Italian to English)

# Continuous Space Language Models
## Introduction



### Application of Continuous Space Language Model

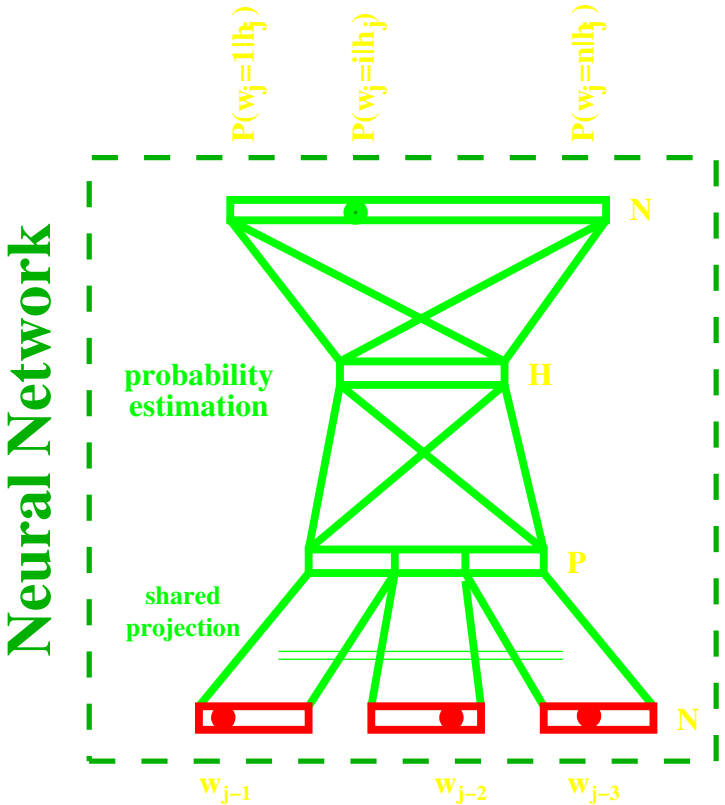
- Very successful in LVCSR
- Initial experiments with a word-based SMT system [Schwenk, ACL'06]

### Cooperation with UPC

- First application of the CSLM to a state-of-the-art SMT system
- *n*-best list rescoring of UPC's phrase and Ngram-based system
- All four languages are considered (translation of Mandarin, Japanese, Arabic and Italian to English)

# Continuous Space Language Models
## Architecture - Probability Calculation





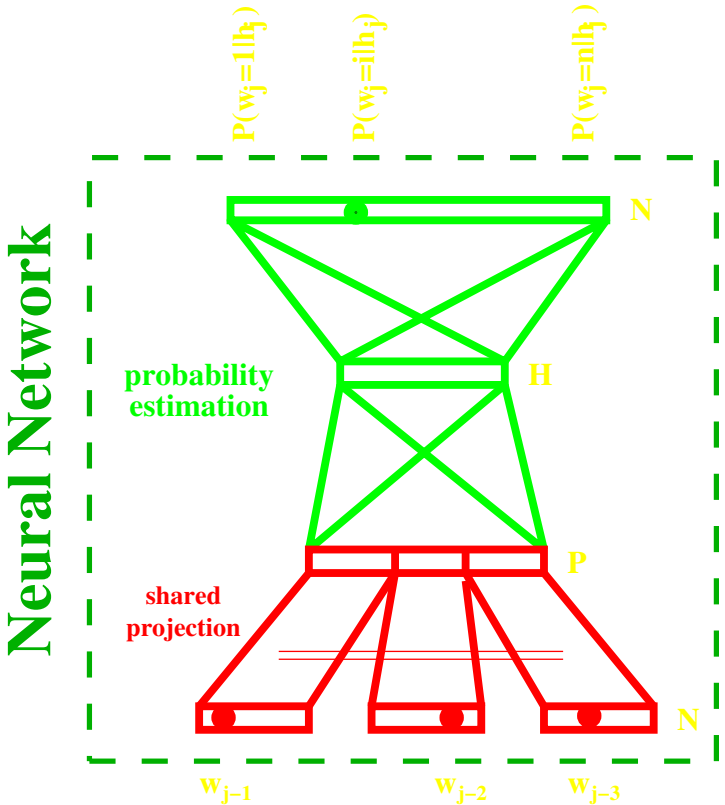

$$h_j = w_{j-n+1}, ..., w_{j-2}, w_{j-1}$$

### Probability Calculation

- Outputs = LM posterior probabilities of all words: $P(w_j = i|h_j) \quad \forall i \in [1, N]$
- Context $h_j$ = sequence of $n-1$ points in this space
- Word = point in the $P$ dimensional space
- Projection onto a continuous space
- Inputs = indices of the $n-1$ previous words

# Continuous Space Language Models
## Architecture - Probability Calculation





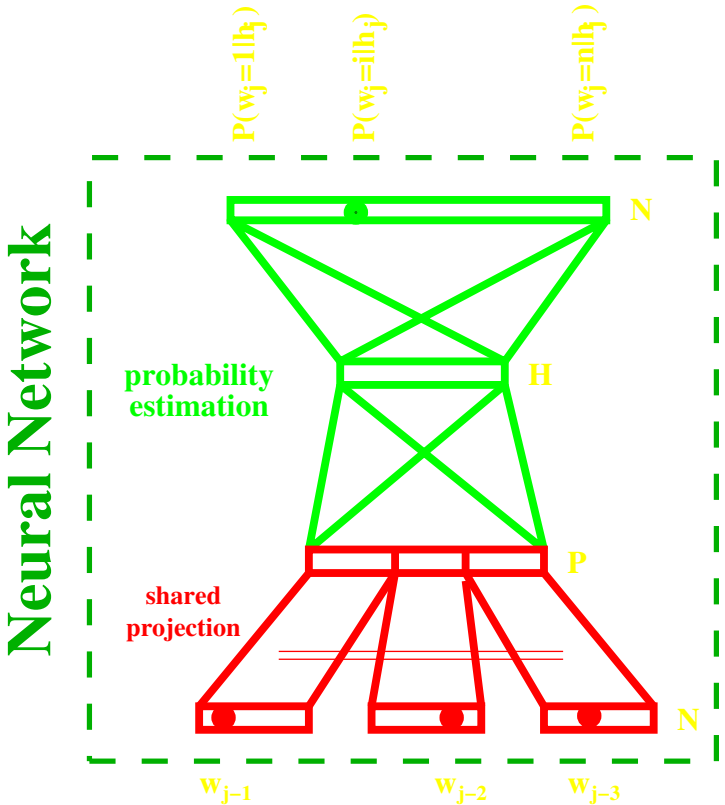

$$h_j = w_{j-n+1}, ..., w_{j-2}, w_{j-1}$$

### Probability Calculation

- Outputs = LM posterior probabilities of all words: $P(w_j = i|h_j) \quad \forall i \in [1, N]$
- Context $h_j$ = sequence of $n-1$ points in this space
- Word = point in the $P$ dimensional space
- Projection onto a continuous space
- Inputs = indices of the $n-1$ previous words

# Continuous Space Language Models
## Architecture - Probability Calculation





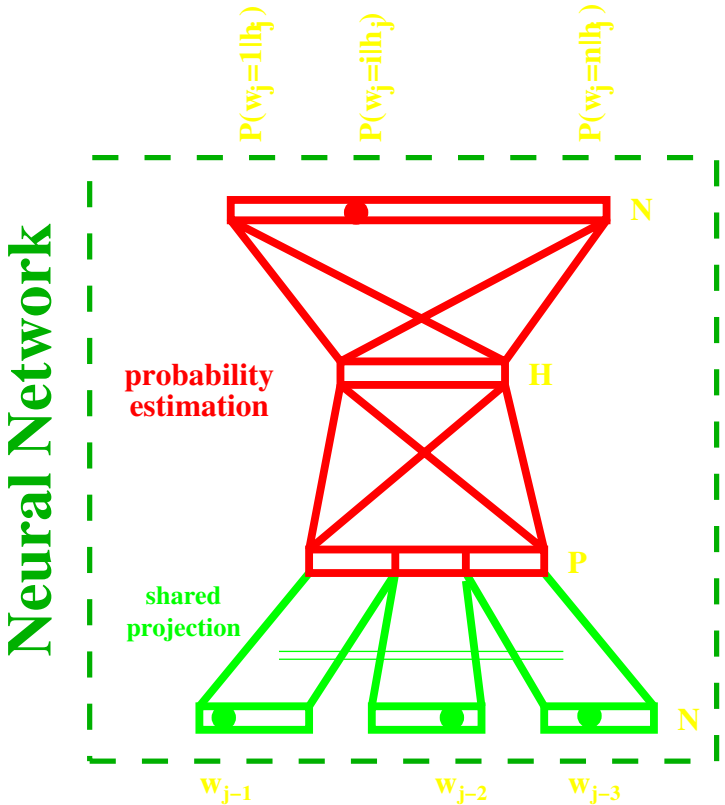

$$h_j = w_{j-n+1}, ..., w_{j-2}, w_{j-1}$$

### Probability Calculation

- Outputs = LM posterior probabilities of all words:
  $$P(w_j = i|h_j) \quad \forall i \in [1, N]$$
- Context $h_j$ = sequence of $n-1$ points in this space
- Word = point in the $P$ dimensional space
- Projection onto a continuous space
- Inputs = indices of the $n-1$ previous words

# Continuous Space Language Models
## Architecture - Probability Calculation





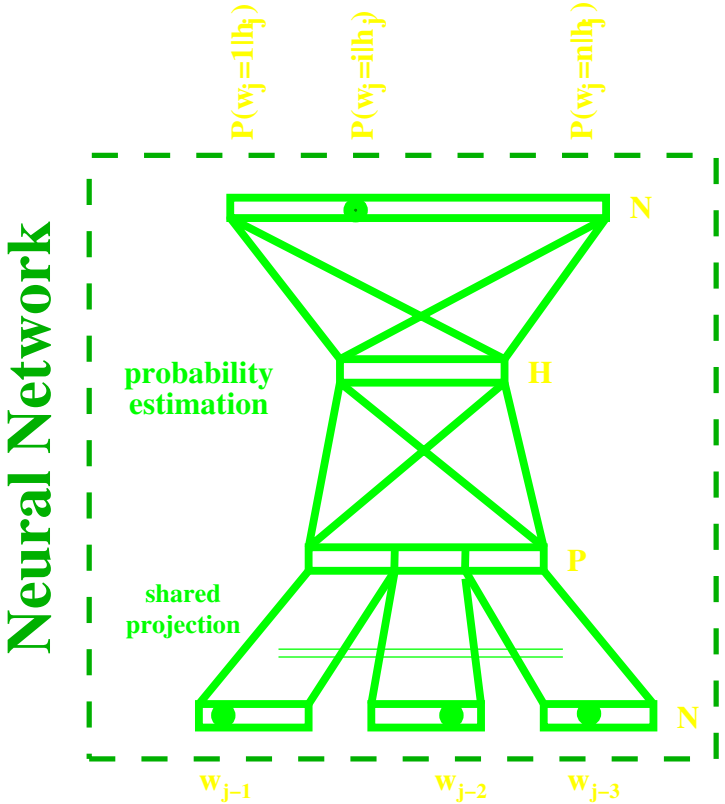

$$h_j = w_{j-n+1}, ..., w_{j-2}, w_{j-1}$$

### Probability Calculation

- Outputs = LM posterior probabilities of all words:
  $P(w_j = i|h_j) \quad \forall i \in [1, N]$
- Context $h_j$ = sequence of $n-1$ points in this space
- Word = point in the $P$ dimensional space
- Projection onto a continuous space
- Inputs = indices of the $n-1$ previous words

# Continuous Space Language Models
## Architecture - Training






### Training

- Backprop training, cross-entropy error

$$E = \sum_{i=1}^{N} d_i \log p_i$$

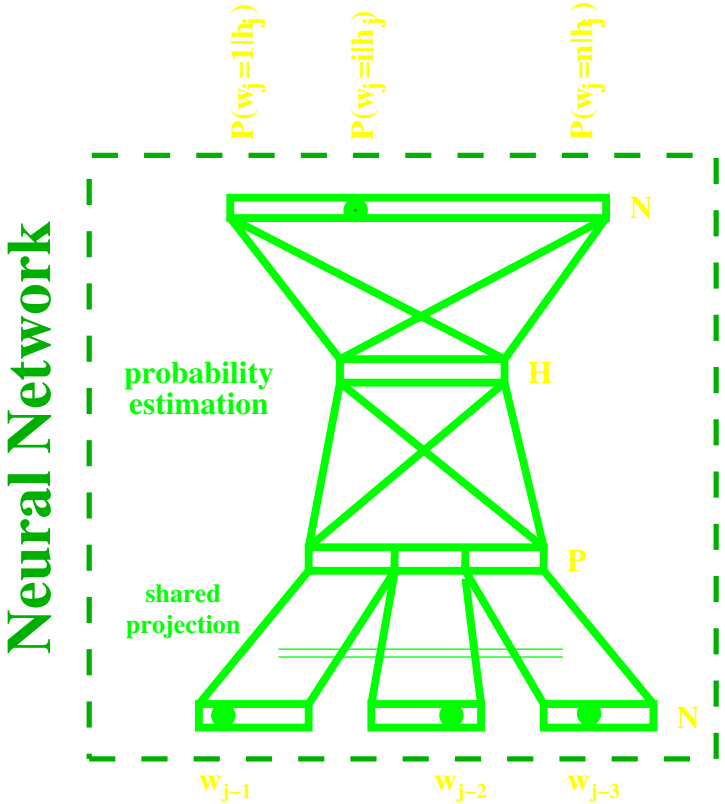
+ weight decay

$\Rightarrow$ NN minimizes perplexity on training data

- Continuous word codes are also learned (random initialization)

# Continuous Space Language Models
## Architecture - Training






### Training

- Backprop training, cross-entropy error

$$E = \sum_{i=1}^{N} d_i \log p_i$$

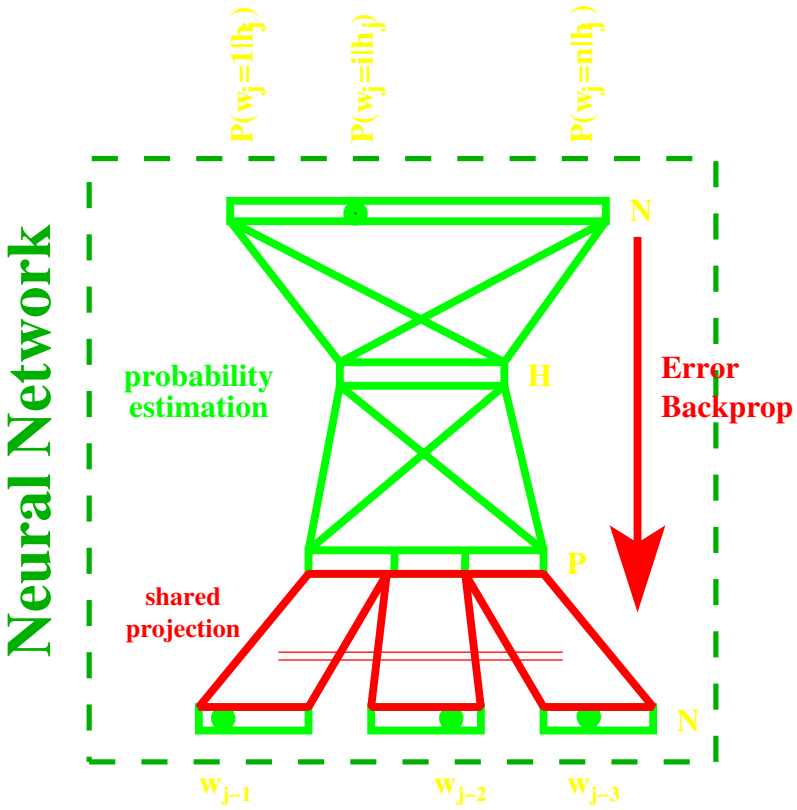
+ weight decay

$\Rightarrow$ NN minimizes perplexity on training data

- Continuous word codes are also learned (random initialization)

# Continuous Space Language Models
## Architecture - Training






### Training

- Backprop training, cross-entropy error

$$E = \sum_{i=1}^{N} d_i \log p_i$$

+ weight decay

$\Rightarrow$ NN minimizes perplexity on training data

- Continuous word codes are also learned (random initialization)

## Interpolation

- Back-off LM (modified Kneser-Ney smoothing, SRILM) and CSLM trained on 326k words,
- Both LM seem to be complementary
  $\rightarrow$ interpolated together
- Several neural networks are trained independently using different sizes of the continuous representation
- EM optimization of the interpolation coefficients: minimize perplexity on the Dev data (0.33 for LM)
- Replace the original LM scores with those of this interpolated LM
- Alternatively we could use several feature functions and tune the coefficients on the BLEU score

# Baseline SMT systems



## Incorporation into UPC's SMT systems

- Use of UPC's phrase-based and Ngram-based system
- Both systems were described in detail just before the break
- Slight difference with respect to official evaluation systems (most of them achieve better results)
- 1000-best list rescoring
  + re-optimization of feature function weights

## Phrase-based system

- Standard phrase extraction algorithm
- Translation model probabilities in both directions are estimated using relative frequencies

# Baseline SMT systems



## Incorporation into UPC's SMT systems

- Use of UPC's phrase-based and Ngram-based system
- Both systems were described in detail just before the break
- Slight difference with respect to official evaluation systems (most of them achieve better results)
- 1000-best list rescoring
  + re-optimization of feature function weights

## Phrase-based system

- Standard phrase extraction algorithm
- Translation model probabilities in both directions are estimated using relative frequencies

# Baseline SMT systems



## N-gram-based system

- Monotonic segmentation of each sentence pair
- Translation model probabilities are estimated as a bilingual LM

$$p(e,f) = Pr(t_1^K) = \prod_{k=1}^{K} p(t_k \mid t_{k-2}, t_{k-1})$$

- This translation model includes an implicit target language model

$\rightarrow$ Is an improved target LM still helpful ?

# Baseline SMT systems



## N-gram-based system

- Monotonic segmentation of each sentence pair
- Translation model probabilities are estimated as a bilingual LM

$$p(e,f) = Pr(t_1^K) = \prod_{k=1}^{K} p(t_k \mid t_{k-2}, t_{k-1})$$

- This translation model includes an implicit target language model

$\longrightarrow$ Is an improved target LM still helpful ?

# Baseline SMT systems
## Additional Features

### Log-linear combination of feature functions

$$\tilde{e}_1^I = \underset{e_1^I}{argmax} \left\{ \sum_{m=1}^{M} \lambda_m h_m(f_1^J, e_1^I) \right\} \quad (1)$$

- Phrase translation probabilities
  or Ngram translation language model
- Word bonus model (and phrase bonus model)
- Source $\rightarrow$ target lexicon model (IBM1 probabilities)
- Target $\rightarrow$ source lexicon model (IBM1 probabilities)
- Target language model
  (4-gram back-off or continuous space LM)

# Experimental Evaluation
## Data sets



### BTEC Open data track

- Open data track of the 2006 IWSLT evaluation

→ Only the supplied subset of the full BTEC corpus was used

- Results on the supplied Dev corpus of 489 sentences ($<$6k words) and the official test set (evaluation server)
- Scoring is case insensitive and punctuations are ignored

# Experimental Evaluation
## Results on Development Data (1)



### BLEU scores

| | Phrase-based system | | | *N*-gram-based system | | |
|---|---|---|---|---|---|---|
| | Oracle | Ref. | CSLM | Oracle | Ref. | CSLM |
| Mand. | 33.1 | 20.68 | 21.97 | 32.0 | 20.84 | 21.83 |
| Japan. | 26.9 | 17.29 | 18.27 | 28.6 | 18.34 | 19.77 |
| Arabic | 40.1 | 27.92 | 30.28 | 41.6 | 29.09 | 30.89 |
| Italian | 56.2 | 41.66 | 44.03 | 58.1 | 41.65 | 44.67 |

- Oracle scores calculated using cheating Dev-LM
- Improvements between 1 and 3 points BLEU
- Slightly better gains for Ngram-based systems
- Notable differences between the languages (also lower oracle BLEU scores)

# Experimental Evaluation
## Results on Development Data (1)



### BLEU scores

| | Phrase-based system | | | *N*-gram-based system | | |
|---|---|---|---|---|---|---|
| | Oracle | Ref. | CSLM | Oracle | Ref. | CSLM |
| Mand. | 33.1 | 20.68 | 21.97 | 32.0 | 20.84 | 21.83 |
| Japan. | 26.9 | 17.29 | 18.27 | 28.6 | 18.34 | 19.77 |
| Arabic | 40.1 | 27.92 | 30.28 | 41.6 | 29.09 | 30.89 |
| Italian | 56.2 | 41.66 | 44.03 | 58.1 | 41.65 | 44.67 |

- Oracle scores calculated using cheating Dev-LM
- Improvements between 1 and 3 points BLEU
- Slightly better gains for Ngram-based systems
- Notable differences between the languages (also lower oracle BLEU scores)

# Experimental Evaluation
## Results on Development Data (1)



### BLEU scores

| | Phrase-based system | | | *N*-gram-based system | | |
|---|---|---|---|---|---|---|
| | Oracle | Ref. | CSLM | Oracle | Ref. | CSLM |
| Mand. | 33.1 | 20.68 | 21.97 | 32.0 | 20.84 | 21.83 |
| Japan. | 26.9 | 17.29 | 18.27 | 28.6 | 18.34 | 19.77 |
| Arabic | 40.1 | 27.92 | 30.28 | 41.6 | 29.09 | 30.89 |
| Italian | 56.2 | 41.66 | 44.03 | 58.1 | 41.65 | 44.67 |

- Oracle scores calculated using cheating Dev-LM
- Improvements between 1 and 3 points BLEU
- Slightly better gains for Ngram-based systems
- Notable differences between the languages (also lower oracle BLEU scores)

# Experimental Evaluation
## Results on Development Data (1)



### BLEU scores

| | Phrase-based system | | | *N*-gram-based system | | |
|---|---|---|---|---|---|---|
| | Oracle | Ref. | CSLM | Oracle | Ref. | CSLM |
| Mand. | 33.1 | 20.68 | 21.97 | 32.0 | 20.84 | 21.83 |
| Japan. | 26.9 | 17.29 | 18.27 | 28.6 | 18.34 | 19.77 |
| Arabic | 40.1 | 27.92 | 30.28 | 41.6 | 29.09 | 30.89 |
| Italian | 56.2 | 41.66 | 44.03 | 58.1 | 41.65 | 44.67 |

- Oracle scores calculated using cheating Dev-LM
- Improvements between 1 and 3 points BLEU
- Slightly better gains for Ngram-based systems
- Notable differences between the languages (also lower oracle BLEU scores)

# Experimental Evaluation
## Results on Development Data (2)



### Word Error rates

| | Phrase-based | | | *N*-gram-based | | |
|---|---|---|---|---|---|---|
| | Oracle | Ref. | CSLM | Oracle | Ref. | CSLM |
| **Ma/En** mWER | 59.1 | 67.4 | 66.5 | 58.1 | 67.8 | 66.6 |
| mPER | 44.5 | 50.8 | 50.1 | 45.3 | 51.5 | 50.6 |
| **Ja/En** mWER | 70.8 | 74.6 | 77.0 | 63.5 | 73.0 | 71.3 |
| mPER | 48.5 | 52.2 | 54.6 | 46.1 | 53.4 | 51.8 |
| **Ar/En** mWER | 49.1 | 56.0 | 52.7 | 48.1 | 55.7 | 52.8 |
| mPER | 40.3 | 45.7 | 43.3 | 39.6 | 44.0 | 42.5 |
| **It/En** mWER | 34.1 | 42.3 | 40.7 | 33.1 | 42.8 | 40.8 |
| mPER | 26.6 | 31.6 | 30.5 | 26.0 | 31.9 | 30.7 |

- Nice gains for the Arabic/English system
- Problem with the phrase-based system for Japanese

# Experimental Evaluation
## Results on Development Data (2)



### Word Error rates

| | Phrase-based | | | *N*-gram-based | | |
|---|---|---|---|---|---|---|
| | Oracle | Ref. | CSLM | Oracle | Ref. | CSLM |
| **Ma/En** mWER | 59.1 | 67.4 | 66.5 | 58.1 | 67.8 | 66.6 |
| mPER | 44.5 | 50.8 | 50.1 | 45.3 | 51.5 | 50.6 |
| **Ja/En** mWER | 70.8 | 74.6 | 77.0 | 63.5 | 73.0 | 71.3 |
| mPER | 48.5 | 52.2 | 54.6 | 46.1 | 53.4 | 51.8 |
| **Ar/En** mWER | 49.1 | 56.0 | 52.7 | 48.1 | 55.7 | 52.8 |
| mPER | 40.3 | 45.7 | 43.3 | 39.6 | 44.0 | 42.5 |
| **It/En** mWER | 34.1 | 42.3 | 40.7 | 33.1 | 42.8 | 40.8 |
| mPER | 26.6 | 31.6 | 30.5 | 26.0 | 31.9 | 30.7 |

- Nice gains for the Arabic/English system
- Problem with the phrase-based system for Japanese

# Experimental Evaluation
## Example Translations



### Phrase-based system

Zh: *could you we arrive time is two thirty departure time is two five ten*
→ *you can the time we arrive at two thirty departure time is two fifty*
Ar: *information your will we arrive at two thirty and an appointment is two and the fifty minutes*
→ *information i'll arrive at two thirty and time is two and fifty minutes*
It: *for your information we'll be arriving at two o'clock and thirty and your departure time is at two o'clock and fifty*
→ *for your information we'll arrive at two thirty and your departure time is at two fifty*

### Ngram-based system

Ja: *we arrive at two thirty takeoff time is fifty two o'clock so you reference you please*
→ *we arrive at two thirty take off time is two o'clock in fifty so you your reference please*
Ar: *i'll information you arrive at two thirty time and is two and fifty minutes*
→ *i'll information you arrive at two thirty and time is two and fifty minutes*

# Experimental Evaluation
## Results on Evaluation Data (1)



| | Phrase-based | | *N*-gram-based | |
|---|---|---|---|---|
| | Ref. | CSLM | Ref. | CSLM |
| **Mandarin/English:** | | | | |
| BLEU | 19.74 | 21.01 | 20.34 | 21.16 |
| mWER | 67.95 | 68.16 | 68.30 | 67.63 |
| mPER | 52.46 | 51.87 | 52.81 | 52.31 |
| **Japanese/English:** | | | | |
| BLEU | 15.11 | 15.73 | 16.14 | 16.35 |
| mWER | 77.51 | 78.15 | 75.45 | 75.59 |
| mPER | 55.14 | 54.96 | 55.52 | 55.29 |

- Good generalization behavior for Mandarin (Dev +1.3/1.0)
- Small gain for Japanese
- mWER increases in mots cases (but not mPER)

# Experimental Evaluation
## Results on Evaluation Data (1)



| | Phrase-based | | *N*-gram-based | |
|---|---|---|---|---|
| | Ref. | CSLM | Ref. | CSLM |
| **Mandarin/English:** | | | | |
| BLEU | 19.74 | 21.01 | 20.34 | 21.16 |
| mWER | 67.95 | 68.16 | 68.30 | 67.63 |
| mPER | 52.46 | 51.87 | 52.81 | 52.31 |
| **Japanese/English:** | | | | |
| BLEU | 15.11 | 15.73 | 16.14 | 16.35 |
| mWER | 77.51 | 78.15 | 75.45 | 75.59 |
| mPER | 55.14 | 54.96 | 55.52 | 55.29 |

- Good generalization behavior for Mandarin (Dev +1.3/1.0)
- Small gain for Japanese
- mWER increases in mots cases (but not mPER)

# Experimental Evaluation
## Results on Evaluation Data (1)



| | Phrase-based | | *N*-gram-based | |
|---|---|---|---|---|
| | Ref. | CSLM | Ref. | CSLM |
| **Mandarin/English:** | | | | |
| BLEU | 19.74 | 21.01 | 20.34 | 21.16 |
| mWER | 67.95 | 68.16 | 68.30 | 67.63 |
| mPER | 52.46 | 51.87 | 52.81 | 52.31 |
| **Japanese/English:** | | | | |
| BLEU | 15.11 | 15.73 | 16.14 | 16.35 |
| mWER | 77.51 | 78.15 | 75.45 | 75.59 |
| mPER | 55.14 | 54.96 | 55.52 | 55.29 |

- Good generalization behavior for Mandarin (Dev +1.3/1.0)
- Small gain for Japanese
- mWER increases in mots cases (but not mPER)

# Experimental Evaluation
## Results on Evaluation Data (2)



| | Phrase-based | | *N*-gram-based | |
|---|---|---|---|---|
| | Ref. | CSLM | Ref. | CSLM |
| **Arabic/English:** | | | | |
| BLEU | 23.72 | 24.86 | 23.83 | 23.70 |
| mWER | 63.04 | 60.89 | 62.81 | 61.97 |
| mPER | 49.43 | 48.61 | 49.41 | 48.85 |
| **Italian/English:** | | | | |
| BLEU | 35.55 | 37.41 | 35.95 | 37.65 |
| mWER | 49.12 | 47.22 | 48.78 | 47.59 |
| mPER | 38.17 | 36.62 | 38.12 | 37.26 |

- No improvement in BLEU score with Ngram-system for Arabic (BLEU decreases despite gain in mWER and mPER)
- Improvements of 1.8 point BLEU for Italian

# Experimental Evaluation
## Results on Evaluation Data (2)



| | Phrase-based | | *N*-gram-based | |
|---|---|---|---|---|
| | Ref. | CSLM | Ref. | CSLM |
| **Arabic/English:** | | | | |
| BLEU | 23.72 | 24.86 | 23.83 | 23.70 |
| mWER | 63.04 | 60.89 | 62.81 | 61.97 |
| mPER | 49.43 | 48.61 | 49.41 | 48.85 |
| **Italian/English:** | | | | |
| BLEU | 35.55 | 37.41 | 35.95 | 37.65 |
| mWER | 49.12 | 47.22 | 48.78 | 47.59 |
| mPER | 38.17 | 36.62 | 38.12 | 37.26 |

- No improvement in BLEU score with Ngram-system for Arabic (BLEU decreases despite gain in mWER and mPER)
- Improvements of 1.8 point BLEU for Italian

# Discussion and Perspectives



## Summary

- Continuous space LM on top of UPC's evaluation systems
- Dev-data: gain between 1 and 3 points BLEU
- Eval data: up to 1.9 points BLEU

$\Rightarrow$ Promising approach for tasks with limited resources

## Ongoing Work

- Further analysis of the improvements
- Interaction with word reordering ?
- Usefulness of long span LMs
- Continuous space translation model (Ngram system)

# Discussion and Perspectives



## Summary

- Continuous space LM on top of UPC's evaluation systems
- Dev-data: gain between 1 and 3 points BLEU
- Eval data: up to 1.9 points BLEU

$\Rightarrow$ Promising approach for tasks with limited resources

## Ongoing Work

- Further analysis of the improvements
- Interaction with word reordering ?
- Usefulness of long span LMs
- Continuous space translation model (Ngram system)