The UEDIN Systems for the IWSLT 2012 Evaluation

Eva Hasler, Peter Bell, Arnab Ghoshal, Barry Haddow, Philipp Koehn, Fergus McInnes, Steve Renals, Pawel Swietojanski

School of Informatics
University of Edinburgh

December 6th

Overview

- UEDIN participated in ASR (English),
  MT (English-French, German-English), SLT (English-French)
- This presentation focuses on experiments carried out for the SLT and MT tasks

Spoken Language Translation

Problem

- ASR output has recognition errors and no punctuation

Approach: Punctuation insertion as machine translation

- Best-performing SLT system of [Wuebker et al., 2011] used this approach (PPMT before translation)
- Advantage: can reuse best MT system for translation into French
- Compare different training data, pre-/postprocessing and tuning setups

Spoken Language Translation

Problem

- ASR output has recognition errors and no punctuation

Approach: Punctuation insertion as machine translation

- Best-performing SLT system of [Wuebker et al., 2011] used this approach (PPMT before translation)
- Advantage: can reuse best MT system for translation into French
- Compare different training data, pre-/postprocessing and tuning setups

Spoken Language Translation

Problem

- ASR output has recognition errors and no punctuation

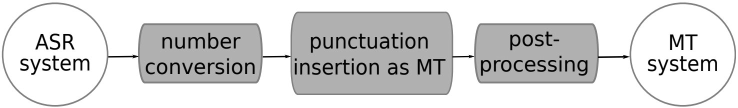
Approach: Punctuation insertion as machine translation

- Best-performing SLT system of [Wuebker et al., 2011] used this approach (PPMT before translation)
- Advantage: can reuse best MT system for translation into French
- Compare different training data, pre-/postprocessing and tuning setups

Spoken Language Translation

SLT pipeline

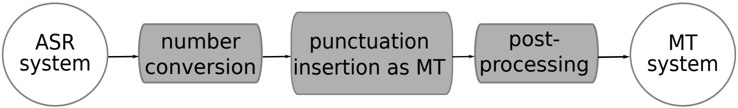
1. Preprocessing of ASR output: number conversion
2. Punctuation insertion by translation from English w/o punctuation to English with punctuation
3. Postprocessing: fix sentence initial/final punctuation, single quotation marks
4. Translation from English to French



Spoken Language Translation

SLT pipeline

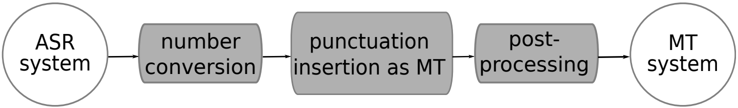
1. Preprocessing of ASR output: number conversion
2. Punctuation insertion by translation from English w/o punctuation to English with punctuation
3. Postprocessing: fix sentence initial/final punctuation, single quotation marks
4. Translation from English to French



Spoken Language Translation

SLT pipeline

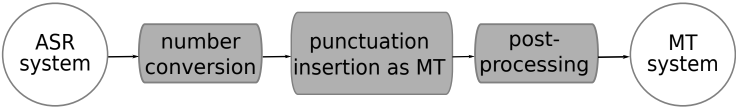
1. Preprocessing of ASR output: number conversion
2. Punctuation insertion by translation from English w/o punctuation to English with punctuation
3. Postprocessing: fix sentence initial/final punctuation, single quotation marks
4. Translation from English to French



Spoken Language Translation

SLT pipeline

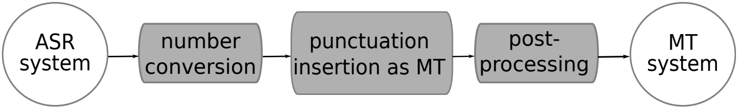
1. Preprocessing of ASR output: number conversion
2. Punctuation insertion by translation from English w/o punctuation to English with punctuation
3. Postprocessing: fix sentence initial/final punctuation, single quotation marks
4. Translation from English to French



Spoken Language Translation

SLT pipeline

1. Preprocessing of ASR output: number conversion
2. Punctuation insertion by translation from English w/o punctuation to English with punctuation
3. Postprocessing: fix sentence initial/final punctuation, single quotation marks
4. Translation from English to French



Spoken Language Translation

Training data for punctuation insertion system

- 141K parallel sentences from the TED corpus
- **Source** side: ASR transcripts of TED talks (w/o punctuation, cased)
- **Target** side: source side of MT data (w/ punctuation, cased)
- Source and target TED talks mapped according to talkids, then sentence-aligned
- Differences between ASR transcripts and MT source: (punctuation,) representation of numbers, spellings
  - Doctor $\rightarrow$ Dr.
  - MP three $\rightarrow$ MP3
- Implicit conversion of spellings

Spoken Language Translation

Training data for punctuation insertion system

- 141K parallel sentences from the TED corpus
- **Source** side: ASR transcripts of TED talks (w/o punctuation, cased)
- **Target** side: source side of MT data (w/ punctuation, cased)
- Source and target TED talks mapped according to talkids, then sentence-aligned
- Differences between ASR transcripts and MT source: (punctuation,) representation of numbers, spellings
  - Doctor → Dr.
  - MP three → MP3
- Implicit conversion of spellings

Spoken Language Translation

Training data for punctuation insertion system

- 141K parallel sentences from the TED corpus
- **Source** side: ASR transcripts of TED talks (w/o punctuation, cased)
- **Target** side: source side of MT data (w/ punctuation, cased)
- Source and target TED talks mapped according to talkids, then sentence-aligned
- Differences between ASR transcripts and MT source: (punctuation,) representation of numbers, spellings
  - Doctor → Dr.
  - MP three → MP3
- Implicit conversion of spellings

Spoken Language Translation

Training data for punctuation insertion system

- 141K parallel sentences from the TED corpus
- **Source** side: ASR transcripts of TED talks (w/o punctuation, cased)
- **Target** side: source side of MT data (w/ punctuation, cased)
- Source and target TED talks mapped according to talkids, then sentence-aligned
- Differences between ASR transcripts and MT source: (punctuation,) representation of numbers, spellings
  - Doctor $\rightarrow$ Dr.
  - MP three $\rightarrow$ MP3
- Implicit conversion of spellings

Spoken Language Translation

Training data for punctuation insertion system

- 141K parallel sentences from the TED corpus
- **Source** side: ASR transcripts of TED talks (w/o punctuation, cased)
- **Target** side: source side of MT data (w/ punctuation, cased)
- Source and target TED talks mapped according to talkids, then sentence-aligned
- Differences between ASR transcripts and MT source: (punctuation,) representation of numbers, spellings
  - Doctor → Dr.
  - MP three → MP3
- Implicit conversion of spellings

Spoken Language Translation

Number conversion

- Explicit conversion as preprocessing step
- Year numbers: mostly consistent in MT data
  - *nineteen thirty two* → *1932*
  - *two thousand and nine* → *2009*
  - *nineteen nineties* → *1990s*
- Other numbers: not always constistent in MT data, but conversion still helps
  - *ten thousand* → *10 thousand* or *10,000* (more frequent)
  - *one hundred seventy four* → *174*
  - *a hundred and twenty* → *120*
  - *twenty sixth* → *26th*

Spoken Language Translation

Number conversion

- Explicit conversion as preprocessing step
- Year numbers: mostly consistent in MT data
  - *nineteen thirty two* → *1932*
  - *two thousand and nine* → *2009*
  - *nineteen nineties* → *1990s*
- Other numbers: not always constistent in MT data, but conversion still helps
  - *ten thousand* → *10 thousand* or *10,000* (more frequent)
  - *one hundred seventy four* → *174*
  - *a hundred and twenty* → *120*
  - *twenty sixth* → *26th*

Spoken Language Translation

Number conversion

- Explicit conversion as preprocessing step
- Year numbers: mostly consistent in MT data
    - *nineteen thirty two* $\rightarrow$ *1932*
    - *two thousand and nine* $\rightarrow$ *2009*
    - *nineteen nineties* $\rightarrow$ *1990s*
- Other numbers: not always constistent in MT data, but conversion still helps
    - *ten thousand* $\rightarrow$ *10 thousand* or *10,000* (more frequent)
    - *one hundred seventy four* $\rightarrow$ *174*
    - *a hundred and twenty* $\rightarrow$ *120*
    - *twenty sixth* $\rightarrow$ *26th*

Spoken Language Translation

Number conversion

- Explicit conversion as preprocessing step
- Year numbers: mostly consistent in MT data
  - *nineteen thirty two* → *1932*
  - *two thousand and nine* → *2009*
  - *nineteen nineties* → *1990s*
- Other numbers: not always constistent in MT data, but conversion still helps
  - *ten thousand* → *10 thousand* or *10,000* (more frequent)
  - *one hundred seventy four* → *174*
  - *a hundred and twenty* → *120*
  - *twenty sixth* → *26th*

Spoken Language Translation

Number conversion

- Explicit conversion as preprocessing step
- Year numbers: mostly consistent in MT data
  - *nineteen thirty two* → *1932*
  - *two thousand and nine* → *2009*
  - *nineteen nineties* → *1990s*
- Other numbers: not always constistent in MT data, but conversion still helps
  - *ten thousand* → *10 thousand* or *10,000* (more frequent)
  - *one hundred seventy four* → *174*
  - *a hundred and twenty* → *120*
  - *twenty sixth* → *26th*

Spoken Language Translation

Punctuation insertion system

- Phrasebased Moses, monotone decoding
- Avoid excessive punctuation insertion
  - Only using cased instead of truecased data improved performance
- Tuning sets (target: MT input)
  - dev2010 transcripts, dev2010+test2010 transcripts, dev2010+test2010 ASR outputs (all number-converted)
- Evaluate different systems in terms of BLEU on MT source

Spoken Language Translation

Punctuation insertion system

- Phrasebased Moses, monotone decoding
- Avoid excessive punctuation insertion
  - Only using cased instead of truecased data improved performance
- Tuning sets (target: MT input)
  - dev2010 transcripts, dev2010+test2010 transcripts, dev2010+test2010 ASR outputs (all number-converted)
- Evaluate different systems in terms of BLEU on MT source

Spoken Language Translation

Punctuation insertion system

- Phrasebased Moses, monotone decoding
- Avoid excessive punctuation insertion
  - Only using cased instead of truecased data improved performance
- Tuning sets (target: MT input)
  - dev2010 transcripts, dev2010+test2010 transcripts, dev2010+test2010 ASR outputs (all number-converted)
- Evaluate different systems in terms of BLEU on MT source

Spoken Language Translation

Punctuation insertion system

- Phrasebased Moses, monotone decoding
- Avoid excessive punctuation insertion
  - Only using cased instead of truecased data improved performance
- Tuning sets (target: MT input)
  - dev2010 transcripts, dev2010+test2010 transcripts, dev2010+test2010 ASR outputs (all number-converted)
- Evaluate different systems in terms of BLEU on MT source

Spoken Language Translation

Punctuation insertion system

- Phrasebased Moses, monotone decoding
- Avoid excessive punctuation insertion
  - Only using cased instead of truecased data improved performance
- Tuning sets (target: MT input)
  - dev2010 transcripts, dev2010+test2010 transcripts, dev2010+test2010 ASR outputs (all number-converted)
- Evaluate different systems in terms of BLEU on MT source

Spoken Language Translation

SLT pipeline	BLEU(MT source)
test2010 ASR transcript	70.79
+ number conversion	71.37
+ punctuation insertion	84.80
+ postprocessing	85.17
test2010 ASR out + SLT pipeline	*61.82*

Punctuation Insertion System	BLEU(MT source)
Tune: dev2010 ASR transcript	
test2011 ASR output + SLT pipeline	62.39
Tune: dev2010+test2010 ASR transcripts	
test2011 ASR output + SLT pipeline	**63.03**
Tune: dev2010+test2010 ASR outputs	
test2011 ASR output + SLT pipeline	**63.35**

Spoken Language Translation

SLT pipeline	BLEU(MT source)
test2010 ASR transcript	70.79
+ number conversion	71.37
+ punctuation insertion	84.80
+ postprocessing	85.17
test2010 ASR out + SLT pipeline	*61.82*

Punctuation Insertion System	BLEU(MT source)
Tune: dev2010 ASR transcript	
test2011 ASR output + SLT pipeline	62.39
Tune: dev2010+test2010 ASR transcripts	
test2011 ASR output + SLT pipeline	**63.03**
Tune: dev2010+test2010 ASR outputs	
test2011 ASR output + SLT pipeline	**63.35**

Spoken Language Translation

SLT pipeline	BLEU(MT source)
test2010 ASR transcript	70.79
+ number conversion	71.37
+ punctuation insertion	84.80
+ postprocessing	85.17
test2010 ASR out + SLT pipeline	*61.82*

Punctuation Insertion System	BLEU(MT source)
Tune: dev2010 ASR transcript	
test2011 ASR output + SLT pipeline	62.39
Tune: dev2010+test2010 ASR transcripts	
test2011 ASR output + SLT pipeline	**63.03**
Tune: dev2010+test2010 ASR outputs	
test2011 ASR output + SLT pipeline	**63.35**

Spoken Language Translation

SLT pipeline	BLEU(MT source)
test2010 ASR transcript	70.79
+ number conversion	71.37
+ punctuation insertion	84.80
+ postprocessing	85.17
test2010 ASR out + SLT pipeline	*61.82*

Punctuation Insertion System	BLEU(MT source)
Tune: dev2010 ASR transcript	
test2011 ASR output + SLT pipeline	62.39
Tune: dev2010+test2010 ASR transcripts	
test2011 ASR output + SLT pipeline	**63.03**
Tune: dev2010+test2010 ASR outputs	
test2011 ASR output + SLT pipeline	**63.35**

Spoken Language Translation

SLT pipeline	BLEU(MT source)
test2010 ASR transcript	70.79
+ number conversion	71.37
+ punctuation insertion	84.80
+ postprocessing	85.17
test2010 ASR out + SLT pipeline	*61.82*

Punctuation Insertion System	BLEU(MT source)
Tune: dev2010 ASR transcript	
test2011 ASR output + SLT pipeline	62.39
Tune: dev2010+test2010 ASR transcripts	
test2011 ASR output + SLT pipeline	**63.03**
Tune: dev2010+test2010 ASR outputs	
test2011 ASR output + SLT pipeline	**63.35**

Spoken Language Translation

SLT pipeline	BLEU(MT source)
test2010 ASR transcript	70.79
+ number conversion	71.37
+ punctuation insertion	84.80
+ postprocessing	85.17
test2010 ASR out + SLT pipeline	*61.82*

Punctuation Insertion System	BLEU(MT source)
Tune: dev2010 ASR transcript	
test2011 ASR output + SLT pipeline	62.39
Tune: dev2010+test2010 ASR transcripts	
test2011 ASR output + SLT pipeline	**63.03**
Tune: dev2010+test2010 ASR outputs	
test2011 ASR output + SLT pipeline	**63.35**

Spoken Language Translation

SLT pipeline	BLEU(MT source)
test2010 ASR transcript	70.79
+ number conversion	71.37
+ punctuation insertion	84.80
+ postprocessing	85.17
test2010 ASR out + SLT pipeline	*61.82*

Punctuation Insertion System	BLEU(MT source)
Tune: dev2010 ASR transcript	
test2011 ASR output + SLT pipeline	62.39
Tune: dev2010+test2010 ASR transcripts	
test2011 ASR output + SLT pipeline	**63.03**
Tune: dev2010+test2010 ASR outputs	
test2011 ASR output + SLT pipeline	**63.35**

Spoken Language Translation

SLT pipeline	BLEU(MT source)
test2010 ASR transcript	70.79
+ number conversion	71.37
+ punctuation insertion	84.80
+ postprocessing	85.17
test2010 ASR out + SLT pipeline	*61.82*

Punctuation Insertion System	BLEU(MT source)
Tune: dev2010 ASR transcript	
test2011 ASR output + SLT pipeline	62.39
Tune: dev2010+test2010 ASR transcripts	
test2011 ASR output + SLT pipeline	**63.03**
Tune: dev2010+test2010 ASR outputs	
test2011 ASR output + SLT pipeline	**63.35**

Spoken Language Translation

SLT pipeline	BLEU(MT source)
test2010 ASR transcript	70.79
+ number conversion	71.37
+ punctuation insertion	84.80
+ postprocessing	85.17
test2010 ASR out + SLT pipeline	*61.82*

Punctuation Insertion System	BLEU(MT source)
Tune: dev2010 ASR transcript	
test2011 ASR output + SLT pipeline	62.39
Tune: dev2010+test2010 ASR transcripts	
test2011 ASR output + SLT pipeline	**63.03**
Tune: dev2010+test2010 ASR outputs	
test2011 ASR output + SLT pipeline	**63.35**

Spoken Language Translation

SLT pipeline	BLEU(MT source)
test2010 ASR transcript	70.79
+ number conversion	71.37
+ punctuation insertion	84.80
+ postprocessing	85.17
test2010 ASR out + SLT pipeline	*61.82*

Punctuation Insertion System	BLEU(MT source)
Tune: dev2010 ASR transcript	
test2011 ASR output + SLT pipeline	62.39
Tune: dev2010+test2010 ASR transcripts	
test2011 ASR output + SLT pipeline	**63.03**
Tune: dev2010+test2010 ASR outputs	
test2011 ASR output + SLT pipeline	**63.35**

Spoken Language Translation

SLT pipeline + MT System	MT src	MT tgt	Oracle
test2010 ASR transcript	85.17	30.54	33.98
test2010 ASR out UEDIN	61.82	22.89	33.98
test2011 ASR out system0	**67.40**	27.37	40.44
test2011 ASR out system1	65.73	**27.47**	40.44
test2011 ASR out system2	65.82	**27.48**	40.44
test2011 ASR out UEDIN	63.35	26.83	40.44

Table: *SLT end-to-end results (BLEU)*

Spoken Language Translation

SLT pipeline + MT System	MT src	MT tgt	Oracle
test2010 ASR transcript	85.17	30.54	33.98
test2010 ASR out UEDIN	61.82	22.89	33.98
test2011 ASR out system0	**67.40**	27.37	40.44
test2011 ASR out system1	65.73	**27.47**	40.44
test2011 ASR out system2	65.82	**27.48**	40.44
test2011 ASR out UEDIN	63.35	26.83	40.44

Table: *SLT end-to-end results (BLEU)*

Spoken Language Translation

SLT pipeline + MT System	MT src	MT tgt	Oracle
test2010 ASR transcript	85.17	30.54	33.98
test2010 ASR out UEDIN	61.82	22.89	33.98
test2011 ASR out system0	**67.40**	27.37	40.44
test2011 ASR out system1	65.73	**27.47**	40.44
test2011 ASR out system2	65.82	**27.48**	40.44
test2011 ASR out UEDIN	63.35	26.83	40.44

Table: *SLT end-to-end results (BLEU)*

Spoken Language Translation

SLT pipeline + MT System	MT src	MT tgt	Oracle
test2010 ASR transcript	85.17	30.54	33.98
test2010 ASR out UEDIN	61.82	22.89	33.98
test2011 ASR out system0	**67.40**	27.37	40.44
test2011 ASR out system1	65.73	**27.47**	40.44
test2011 ASR out system2	65.82	**27.48**	40.44
test2011 ASR out UEDIN	63.35	26.83	40.44

Table: *SLT end-to-end results (BLEU)*

Spoken Language Translation

SLT pipeline + MT System	MT src	MT tgt	Oracle
test2010 ASR transcript	85.17	30.54	33.98
test2010 ASR out UEDIN	61.82	22.89	33.98
test2011 ASR out system0	**67.40**	27.37	40.44
test2011 ASR out system1	65.73	**27.47**	40.44
test2011 ASR out system2	65.82	**27.48**	40.44
test2011 ASR out UEDIN	63.35	26.83	40.44

Table: *SLT end-to-end results (BLEU)*

Spoken Language Translation

SLT pipeline + MT System	MT src	MT tgt	Oracle
test2010 ASR transcript	85.17	30.54	33.98
test2010 ASR out UEDIN	61.82	22.89	33.98
test2011 ASR out system0	**67.40**	27.37	40.44
test2011 ASR out system1	65.73	**27.47**	40.44
test2011 ASR out system2	65.82	**27.48**	40.44
test2011 ASR out UEDIN	63.35	26.83	40.44

Table: *SLT end-to-end results (BLEU)*

Spoken Language Translation

SLT pipeline + MT System	MT src	MT tgt	Oracle
test2010 ASR transcript	85.17	30.54	33.98
test2010 ASR out UEDIN	61.82	22.89	33.98
test2011 ASR out system0	**67.40**	27.37	40.44
test2011 ASR out system1	65.73	**27.47**	40.44
test2011 ASR out system2	65.82	**27.48**	40.44
test2011 ASR out UEDIN	63.35	26.83	40.44

Table: *SLT end-to-end results (BLEU)*

Machine Translation

Problem

- Limited amount of TED talks data, larger amounts of out-of-domain data
- Need to make best use of both kinds of data

English-French, German-English

- Compare approaches to data filtering and PT adaptation (previous work)
- Adaptation to TED talks by adding sparse lexicalised features
- Explore different tuning setups on in-domain and mixed-domain systems

Machine Translation

Problem

- Limited amount of TED talks data, larger amounts of out-of-domain data
- Need to make best use of both kinds of data

English-French, German-English

- Compare approaches to data filtering and PT adaptation (previous work)
- Adaptation to TED talks by adding sparse lexicalised features
- Explore different tuning setups on in-domain and mixed-domain systems

Machine Translation

Problem

- Limited amount of TED talks data, larger amounts of out-of-domain data
- Need to make best use of both kinds of data

English-French, German-English

- Compare approaches to data filtering and PT adaptation (previous work)
- Adaptation to TED talks by adding sparse lexicalised features
- Explore different tuning setups on in-domain and mixed-domain systems

Machine Translation

Problem

- Limited amount of TED talks data, larger amounts of out-of-domain data
- Need to make best use of both kinds of data

English-French, German-English

- Compare approaches to data filtering and PT adaptation (previous work)
- Adaptation to TED talks by adding sparse lexicalised features
- Explore different tuning setups on in-domain and mixed-domain systems

Machine Translation

Problem

- Limited amount of TED talks data, larger amounts of out-of-domain data
- Need to make best use of both kinds of data

English-French, German-English

- Compare approaches to data filtering and PT adaptation (previous work)
- Adaptation to TED talks by adding sparse lexicalised features
- Explore different tuning setups on in-domain and mixed-domain systems

Machine Translation

Baseline systems in-domain, mixed domain

- Phrase-based/hierarchical Moses
- 5gram LMs with modified Kneser-Ney smoothing
- German-English:
  compound splitting [Koehn and Knight, 2003] and syntactic preordering on source side [Collins et al., 2005]

Data

- Parallel in-domain data: 140K/130K TED talks
- Parallel out-of-domain data:
  Europarl, News Commentary, MultiUN, (10^9)
- Additional LM data: Gigaword, Newscrawl
  (fr: 1.3G words, en: 6.4G words)
- Dev set: dev2010, Devtest set: test2010, Test set: test2011

Machine Translation

Baseline systems

System	de-en (test2010)
IN-PB (CS)	28.26
IN-PB (PRE)	28.04
IN-PB (CS + PRE)	**28.54**

	test2010	
System	en-fr	de-en
IN hierarchical	28.94	27.88
IN phrasebased	**29.58**	**28.54**
IN+OUT phrasebased	31.67	28.39
+ only in-domain LM	30.97	28.61
+ gigaword + newscrawl	31.96	30.26

Machine Translation

Baseline systems

System	de-en (test2010)
IN-PB (CS)	28.26
IN-PB (PRE)	28.04
IN-PB (CS + PRE)	**28.54**

System	test2010 en-fr	test2010 de-en
IN hierarchical	28.94	27.88
IN phrasebased	**29.58**	**28.54**
IN+OUT phrasebased	31.67	28.39
+ only in-domain LM	30.97	28.61
+ gigaword + newscrawl	31.96	30.26

Machine Translation

Baseline systems

System	de-en (test2010)
IN-PB (CS)	28.26
IN-PB (PRE)	28.04
IN-PB (CS + PRE)	**28.54**

	test2010	
System	en-fr	de-en
IN hierarchical	28.94	27.88
IN phrasebased	**29.58**	**28.54**
IN+OUT phrasebased	31.67	28.39
+ only in-domain LM	30.97	28.61
+ gigaword + newscrawl	31.96	30.26

Machine Translation

Baseline systems

System	de-en (test2010)
IN-PB (CS)	28.26
IN-PB (PRE)	28.04
IN-PB (CS + PRE)	**28.54**

System	test2010 en-fr	test2010 de-en
IN hierarchical	28.94	27.88
IN phrasebased	**29.58**	**28.54**
IN+OUT phrasebased	31.67	28.39
+ only in-domain LM	30.97	28.61
+ gigaword + newscrawl	31.96	30.26

Machine Translation

Baseline systems

System	de-en (test2010)
IN-PB (CS)	28.26
IN-PB (PRE)	28.04
IN-PB (CS + PRE)	**28.54**

System	test2010 en-fr	test2010 de-en
IN hierarchical	28.94	27.88
IN phrasebased	**29.58**	**28.54**
IN+OUT phrasebased	31.67	28.39
+ only in-domain LM	30.97	28.61
+ gigaword + newscrawl	31.96	30.26

Machine Translation

Baseline systems

System	de-en (test2010)
IN-PB (CS)	28.26
IN-PB (PRE)	28.04
IN-PB (CS + PRE)	**28.54**

	test2010	
System	en-fr	de-en
IN hierarchical	28.94	27.88
IN phrasebased	**29.58**	**28.54**
IN+OUT phrasebased	31.67	28.39
+ only in-domain LM	30.97	28.61
+ gigaword + newscrawl	31.96	30.26

Data selection and PT adaptation

Bilingual cross-entropy difference [Axelrod et al., 2011]

- Select out-of-domain sentences that are similar to in-domain and dissimilar from out-of-domain data
- Select 10%, 20%, 50% of OUT data (incl. LM data)

In-domain PT + fill-up OUT
[Bisazza et al., 2011], [Haddow and Koehn, 2012]

- Train phrase-table on both IN and OUT data
- Replace all scores of phrase pairs found in IN table with the scores from that table

Data selection and PT adaptation

Bilingual cross-entropy difference [Axelrod et al., 2011]

- Select out-of-domain sentences that are similar to in-domain and dissimilar from out-of-domain data
- Select 10%, 20%, 50% of OUT data (incl. LM data)

In-domain PT + fill-up OUT
[Bisazza et al., 2011], [Haddow and Koehn, 2012]

- Train phrase-table on both IN and OUT data
- Replace all scores of phrase pairs found in IN table with the scores from that table

Data selection and PT adaptation

	test2010	
System	en-fr	de-en
IN+OUT	31.67	28.39
IN		
+ 10% OUT	32.30	**29.29**
+ 20% OUT	**32.45**	29.11
+ 50% OUT	32.32	28.68
best + gigaword + newscrawl	32.93	31.06
IN + fill-up OUT	**32.19**	**29.59**
+ gigaword + newscrawl	32.72	31.30

Data selection and PT adaptation

System	test2010 en-fr	de-en
IN+OUT	31.67	28.39
IN		
+ 10% OUT	32.30	**29.29**
+ 20% OUT	**32.45**	29.11
+ 50% OUT	32.32	28.68
best + gigaword + newscrawl	32.93	31.06
IN + fill-up OUT	**32.19**	**29.59**
+ gigaword + newscrawl	32.72	31.30

Data selection and PT adaptation

System	test2010 en-fr	de-en
IN+OUT	31.67	28.39
IN		
+ 10% OUT	32.30	**29.29**
+ 20% OUT	**32.45**	29.11
+ 50% OUT	32.32	28.68
best + gigaword + newscrawl	32.93	31.06
IN + fill-up OUT	**32.19**	**29.59**
+ gigaword + newscrawl	32.72	31.30

Data selection and PT adaptation

	test2010	
System	en-fr	de-en
IN+OUT	31.67	28.39
IN		
+ 10% OUT	32.30	**29.29**
+ 20% OUT	**32.45**	29.11
+ 50% OUT	32.32	28.68
best + gigaword + newscrawl	32.93	31.06
IN + fill-up OUT	**32.19**	**29.59**
+ gigaword + newscrawl	32.72	31.30

Sparse feature tuning

Adapt to style and vocabulary of TED talks

- Add sparse word pair and phrase pair features to in-domain system, tune with online MIRA
- Word pairs: indicators of aligned words in source and target
- Phrase pairs: depend on phrase segmentation of decoder
- Bias translation model towards in-domain style and vocabulary

Sparse feature tuning

Adapt to style and vocabulary of TED talks

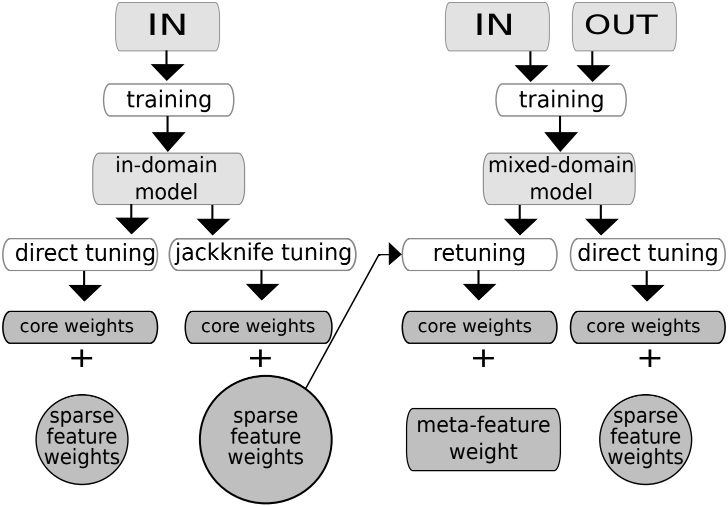
- Add sparse word pair and phrase pair features to in-domain system, tune with online MIRA
- Word pairs: indicators of aligned words in source and target
- Phrase pairs: depend on phrase segmentation of decoder
- Bias translation model towards in-domain style and vocabulary

Sparse feature tuning

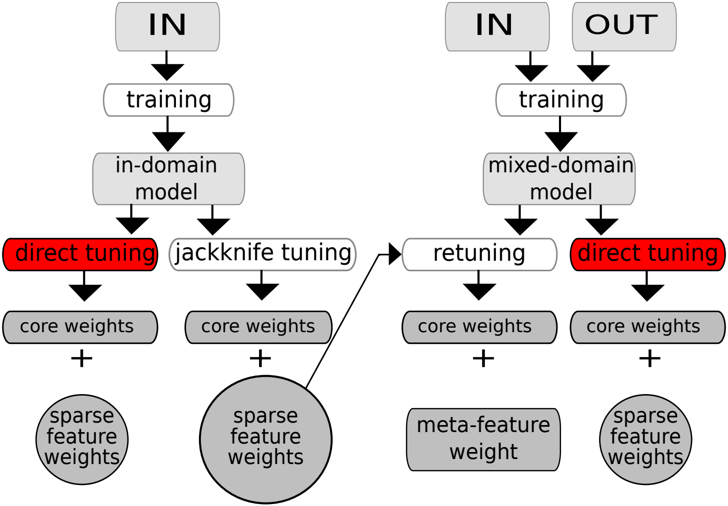
Adapt to style and vocabulary of TED talks

- Add sparse word pair and phrase pair features to in-domain system, tune with online MIRA
- Word pairs: indicators of aligned words in source and target
- Phrase pairs: depend on phrase segmentation of decoder
- Bias translation model towards in-domain style and vocabulary

Sparse feature tuning schemes



Sparse feature tuning schemes



Direct tuning with MIRA

- Tune on development set
- Online MIRA: Select hope/fear translations from a 30best list
- Sentence-level BLEU scores
- Separate learning rate for core features to reduce fluctuation and keep MIRA training more stable
- Learning rate set to 0.1 for core features
(1.0 for sparse features)

Direct tuning with MIRA

- Tune on development set
- Online MIRA: Select hope/fear translations from a 30best list
- Sentence-level BLEU scores
- Separate learning rate for core features to reduce fluctuation and keep MIRA training more stable
- Learning rate set to 0.1 for core features (1.0 for sparse features)

Direct tuning with MIRA

- Tune on development set
- Online MIRA: Select hope/fear translations from a 30best list
- Sentence-level BLEU scores
- Separate learning rate for core features to reduce fluctuation and keep MIRA training more stable
- Learning rate set to 0.1 for core features
  (1.0 for sparse features)

Direct tuning with MIRA

- Tune on development set
- Online MIRA: Select hope/fear translations from a 30best list
- Sentence-level BLEU scores
- Separate learning rate for core features to reduce fluctuation and keep MIRA training more stable
- Learning rate set to 0.1 for core features (1.0 for sparse features)

Direct tuning with MIRA

Sparse feature sets

Source sentence:

[a language] [is a] [flash of] [the human spirit] [.]

Hypothesis translation:

[une langue] [est une] [flash de] [l' esprit humain] [.]

Word pair features	**Phrase pair features**
wp_a∼une=2	pp_a,language∼une,langue=1
wp_language∼langue=1	pp_is,a∼est,une=1
wp_is∼est=1	pp_flash,of∼flash,de=1
wp_flash∼ flash=1	...
wp_of∼de=1	
...	

Direct tuning with MIRA

Sparse feature sets

Source sentence:

[a language] [is a] [flash of] [the human spirit] [.]

Hypothesis translation:

[une langue] [est une] [flash de] [l' esprit humain] [.]

Word pair features

wp_a∼une=2
wp_language∼langue=1
wp_is∼est=1
wp_flash∼ flash=1
wp_of∼de=1
...

Phrase pair features

pp_a,language∼une,langue=1
pp_is,a∼est,une=1
pp_flash,of∼flash,de=1
...

Direct tuning with MIRA

Sparse feature sets

Source sentence:

[a language] [is a] [flash of] [the human spirit] [.]

Hypothesis translation:

[une langue] [est une] [flash de] [l' esprit humain] [.]

Word pair features	**Phrase pair features**
wp_a∼une=2	pp_a,language∼une,langue=1
wp_language∼langue=1	pp_is,a∼est,une=1
wp_is∼est=1	pp_flash,of∼flash,de=1
wp_flash∼ flash=1	...
wp_of∼de=1	
...	

Direct tuning with MIRA

Sparse feature sets

Source sentence:

[a language] [is a] [flash of] [the human spirit] [.]

Hypothesis translation:

[une langue] [est une] [flash de] [l' esprit humain] [.]

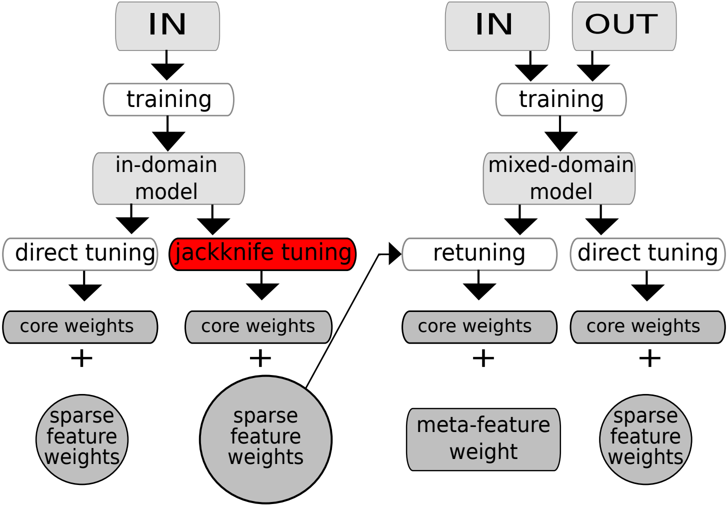
Word pair features

wp_a∼une=2
wp_language∼langue=1
wp_is∼est=1
wp_flash∼ flash=1
wp_of∼de=1
...

Phrase pair features

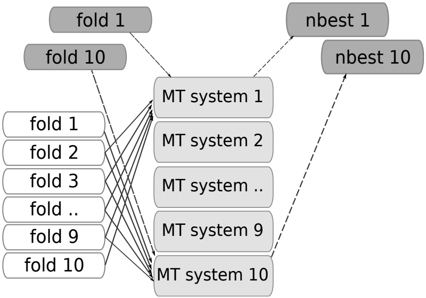
pp_a,language∼une,langue=1
pp_is,a∼est,une=1
pp_flash,of∼flash,de=1
...

Sparse feature tuning schemes



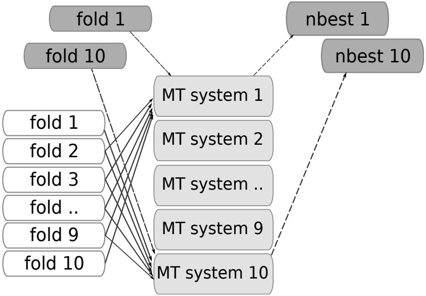
Jackknife tuning with MIRA

- To avoid overfitting to tuning set, train lexicalised features on all in-domain training data
- Train 10 systems on in-domain data, leaving out one fold at a time
- Then translate each fold with respective system
- Iterative parameter mixing by running MIRA on all 10 systems in parallel



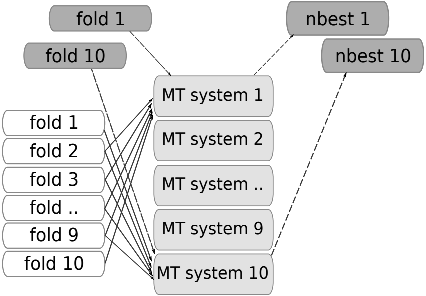
Jackknife tuning with MIRA

- To avoid overfitting to tuning set, train lexicalised features on all in-domain training data
- Train 10 systems on in-domain data, leaving out one fold at a time
- Then translate each fold with respective system
- Iterative parameter mixing by running MIRA on all 10 systems in parallel

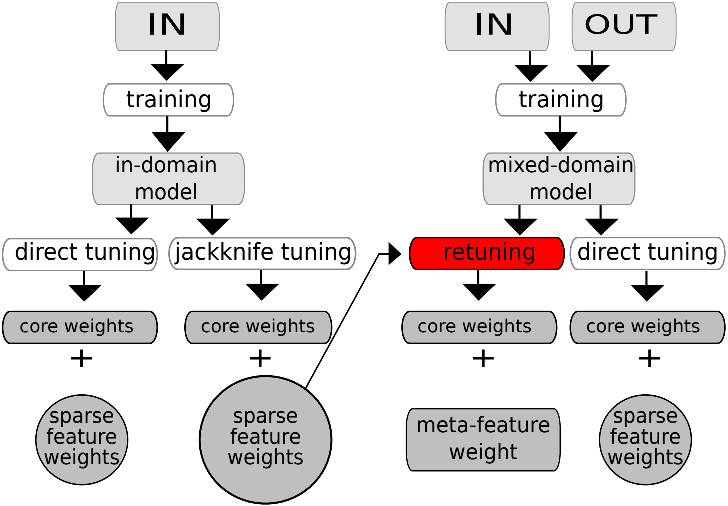


Jackknife tuning with MIRA

- To avoid overfitting to tuning set, train lexicalised features on all in-domain training data
- Train 10 systems on in-domain data, leaving out one fold at a time
- Then translate each fold with respective system
- Iterative parameter mixing by running MIRA on all 10 systems in parallel



Sparse feature tuning schemes



Retuning with MIRA

Motivation

- Tuning sparse features for large translation models is time/memory-consuming
- Avoid overhead of jackknife tuning on larger data sets
- Port tuned features from in-domain to mixed-domain models

Feature integration

- Rescale jackknife-tuned features to integrate into mixed-domain model
- Combine into aggregated meta-feature with a single weight
- During decoding, meta-feature weight is applied to all sparse features of the same class
- Retuning step: core weights of mixed-domain model tuned together with meta-feature weight

Retuning with MIRA

Motivation

- Tuning sparse features for large translation models is time/memory-consuming
- Avoid overhead of jackknife tuning on larger data sets
- Port tuned features from in-domain to mixed-domain models

Feature integration

- Rescale jackknife-tuned features to integrate into mixed-domain model
- Combine into aggregated meta-feature with a single weight
- During decoding, meta-feature weight is applied to all sparse features of the same class
- Retuning step: core weights of mixed-domain model tuned together with meta-feature weight

Retuning with MIRA

Motivation

- Tuning sparse features for large translation models is time/memory-consuming
- Avoid overhead of jackknife tuning on larger data sets
- Port tuned features from in-domain to mixed-domain models

Feature integration

- Rescale jackknife-tuned features to integrate into mixed-domain model
- Combine into aggregated meta-feature with a single weight
- During decoding, meta-feature weight is applied to all sparse features of the same class
- Retuning step: core weights of mixed-domain model tuned together with meta-feature weight

Retuning with MIRA

Motivation

- Tuning sparse features for large translation models is time/memory-consuming
- Avoid overhead of jackknife tuning on larger data sets
- Port tuned features from in-domain to mixed-domain models

Feature integration

- Rescale jackknife-tuned features to integrate into mixed-domain model
- Combine into aggregated meta-feature with a single weight
- During decoding, meta-feature weight is applied to all sparse features of the same class
- Retuning step: core weights of mixed-domain model tuned together with meta-feature weight

Retuning with MIRA

Motivation

- Tuning sparse features for large translation models is time/memory-consuming
- Avoid overhead of jackknife tuning on larger data sets
- Port tuned features from in-domain to mixed-domain models

Feature integration

- Rescale jackknife-tuned features to integrate into mixed-domain model
- Combine into aggregated meta-feature with a single weight
- During decoding, meta-feature weight is applied to all sparse features of the same class
- Retuning step: core weights of mixed-domain model tuned together with meta-feature weight

Retuning with MIRA

Motivation

- Tuning sparse features for large translation models is time/memory-consuming
- Avoid overhead of jackknife tuning on larger data sets
- Port tuned features from in-domain to mixed-domain models

Feature integration

- Rescale jackknife-tuned features to integrate into mixed-domain model
- Combine into aggregated meta-feature with a single weight
- During decoding, meta-feature weight is applied to all sparse features of the same class
- Retuning step: core weights of mixed-domain model tuned together with meta-feature weight

Retuning with MIRA

Motivation

- Tuning sparse features for large translation models is time/memory-consuming
- Avoid overhead of jackknife tuning on larger data sets
- Port tuned features from in-domain to mixed-domain models

Feature integration

- Rescale jackknife-tuned features to integrate into mixed-domain model
- Combine into aggregated meta-feature with a single weight
- During decoding, meta-feature weight is applied to all sparse features of the same class
- Retuning step: core weights of mixed-domain model tuned together with meta-feature weight

Results with sparse features

	test2010	
System	en-fr	de-en
IN, MERT	29.58	28.54
IN, MIRA	30.28	28.31
+ word pairs	**30.36**	**28.45**
+ phrase pairs	**30.62**	**28.40**
+ word pairs (JK)	**30.80**	**28.78**
+ phrase pairs (JK)	**30.77**	**28.61**

Table: *Direct tuning and jackknife tuning on in-domain data*

- en-fr: +0.34/+0.52 BLEU with direct/jackknife tuning
- de-en: +0.14/+0.47 BLEU with direct/jackknife tuning

Results with sparse features

	test2010	
System	en-fr	de-en
IN, MERT	29.58	28.54
IN, MIRA	30.28	28.31
+ word pairs	**30.36**	**28.45**
+ phrase pairs	**30.62**	**28.40**
+ word pairs (JK)	**30.80**	**28.78**
+ phrase pairs (JK)	**30.77**	**28.61**

Table: *Direct tuning and jackknife tuning on in-domain data*

- en-fr: +0.34/+0.52 BLEU with direct/jackknife tuning
- de-en: +0.14/+0.47 BLEU with direct/jackknife tuning

Results with sparse features

	test2010	
System	en-fr	de-en
IN, MERT	29.58	28.54
IN, MIRA	30.28	28.31
+ word pairs	**30.36**	**28.45**
+ phrase pairs	**30.62**	**28.40**
+ word pairs (JK)	**30.80**	**28.78**
+ phrase pairs (JK)	**30.77**	**28.61**

Table: *Direct tuning and jackknife tuning on in-domain data*

- en-fr: +0.34/+0.52 BLEU with direct/jackknife tuning
- de-en: +0.14/+0.47 BLEU with direct/jackknife tuning

MT Results

System	en-fr test2010	en-fr test2011	de-en test2010	de-en test2011
IN + %OUT, MIRA	33.22	40.02	28.90	34.03
+ word pairs	**33.59**	39.95	28.93	33.88
+ phrase pairs	33.44	40.02	**29.13**	33.99
IN + %OUT, MERT	32.32	39.36	29.13	33.29
+ retune(word pair JK)	**32.90**	40.31	**29.58**	33.31
+ retune(phrase pairs JK)	32.69	39.32	29.38	33.23
Submission system (grey) + gigaword + newscrawl	33.98	40.44	31.28	36.03

Table: *(Data selection + Sparse features (direct/retuning)) + large LMs*

MT Results

System	en-fr test2010	en-fr test2011	de-en test2010	de-en test2011
IN + %OUT, MIRA	33.22	40.02	28.90	34.03
+ word pairs	**33.59**	39.95	28.93	33.88
+ phrase pairs	33.44	40.02	**29.13**	33.99
IN + %OUT, MERT	32.32	39.36	29.13	33.29
+ retune(word pair JK)	**32.90**	40.31	**29.58**	33.31
+ retune(phrase pairs JK)	32.69	39.32	29.38	33.23
Submission system (grey) + gigaword + newscrawl	33.98	40.44	31.28	36.03

Table: *(Data selection + Sparse features (direct/retuning)) + large LMs*

MT Results

System	en-fr test2010	en-fr test2011	de-en test2010	de-en test2011
IN + %OUT, MIRA	33.22	40.02	28.90	34.03
+ word pairs	**33.59**	39.95	28.93	33.88
+ phrase pairs	33.44	40.02	29.13	33.99
IN + %OUT, MERT	32.32	39.36	29.13	33.29
+ retune(word pair JK)	32.90	40.31	**29.58**	33.31
+ retune(phrase pairs JK)	32.69	39.32	29.38	33.23
Submission system (grey) + gigaword + newscrawl	33.98	40.44	31.28	36.03

Table: *(Data selection + Sparse features (direct/retuning)) + large LMs*

MT Results

System	en-fr test2010	en-fr test2011	de-en test2010	de-en test2011
IN + %OUT, MIRA	33.22	40.02	28.90	34.03
+ word pairs	**33.59**	39.95	28.93	33.88
+ phrase pairs	33.44	40.02	29.13	33.99
IN + %OUT, MERT	32.32	39.36	29.13	33.29
+ retune(word pair JK)	32.90	40.31	**29.58**	33.31
+ retune(phrase pairs JK)	32.69	39.32	29.38	33.23
Submission system (grey) + gigaword + newscrawl	33.98	40.44	31.28	36.03

Table: *(Data selection + Sparse features (direct/retuning)) + large LMs*

MT Results

System	en-fr test2010	en-fr test2011	de-en test2010	de-en test2011
IN + %OUT, MIRA	33.22	40.02	28.90	**34.03**
+ word pairs	**33.59**	39.95	28.93	33.88
+ phrase pairs	33.44	40.02	29.13	33.99
IN + %OUT, MERT	32.32	39.36	29.13	33.29
+ retune(word pair JK)	32.90	**40.31**	**29.58**	33.31
+ retune(phrase pairs JK)	32.69	39.32	29.38	33.23
Submission system (grey) + gigaword + newscrawl	33.98	40.44	31.28	36.03

Table: *(Data selection + Sparse features (direct/retuning)) + large LMs*

MT Results

System	en-fr test2010	en-fr test2011	de-en test2010	de-en test2011
IN + %OUT, MIRA	33.22	40.02	28.90	**34.03**
+ word pairs	**33.59**	39.95	28.93	33.88
+ phrase pairs	33.44	40.02	29.13	33.99
IN + %OUT, MERT	32.32	39.36	29.13	33.29
+ retune(word pairs JK)	32.90	**40.31**	**29.58**	33.31
+ retune(phrase pairs JK)	32.69	39.32	29.38	33.23
Submission system (grey)				
+ gigaword + newscrawl	**33.98**	**40.44**	**31.28**	**36.03**

Table: *(Data selection + Sparse features (direct/retuning)) + large LMs*

Summary MT

- Used data selection for final systems (IN+OUT)
- Sparse lexicalised features to adapt to style and vocabulary of TED talks, larger gains with jackknife tuning
- Compared three tuning setups for sparse features

- On test2010, all systems with sparse features improved over baselines, less systematic differences on test2011
- Best system for de-en:
  test2010: IN+10%OUT, MERT+retune(wp JK)
  test2011: IN+10%OUT, MIRA
- Best systems for en-fr:
  test2010: IN+20%OUT, MIRA+wp
  test2011: IN+20%OUT, MERT+retune(wp JK)

Summary MT

- Used data selection for final systems (IN+OUT)
- Sparse lexicalised features to adapt to style and vocabulary of TED talks, larger gains with jackknife tuning
- Compared three tuning setups for sparse features

- On test2010, all systems with sparse features improved over baselines, less systematic differences on test2011
- Best system for de-en:
  test2010: IN+10%OUT, MERT+retune(wp JK)
  test2011: IN+10%OUT, MIRA
- Best systems for en-fr:
  test2010: IN+20%OUT, MIRA+wp
  test2011: IN+20%OUT, MERT+retune(wp JK)

Summary MT

- Used data selection for final systems (IN+OUT)
- Sparse lexicalised features to adapt to style and vocabulary of TED talks, larger gains with jackknife tuning
- Compared three tuning setups for sparse features

- On test2010, all systems with sparse features improved over baselines, less systematic differences on test2011
- Best system for de-en:
  test2010: IN+10%OUT, MERT+retune(wp JK)
  test2011: IN+10%OUT, MIRA
- Best systems for en-fr:
  test2010: IN+20%OUT, MIRA+wp
  test2011: IN+20%OUT, MERT+retune(wp JK)

Summary MT

- Used data selection for final systems (IN+OUT)
- Sparse lexicalised features to adapt to style and vocabulary of TED talks, larger gains with jackknife tuning
- Compared three tuning setups for sparse features

- On test2010, all systems with sparse features improved over baselines, less systematic differences on test2011
- Best system for de-en:
  test2010: IN+10%OUT, MERT+retune(wp JK)
  test2011: IN+10%OUT, MIRA
- Best systems for en-fr:
  test2010: IN+20%OUT, MIRA+wp
  test2011: IN+20%OUT, MERT+retune(wp JK)

Summary MT

- Used data selection for final systems (IN+OUT)
- Sparse lexicalised features to adapt to style and vocabulary of TED talks, larger gains with jackknife tuning
- Compared three tuning setups for sparse features

- On test2010, all systems with sparse features improved over baselines, less systematic differences on test2011
- Best system for de-en:
  test2010: IN+10%OUT, MERT+retune(wp JK)
  test2011: IN+10%OUT, MIRA
- Best systems for en-fr:
  test2010: IN+20%OUT, MIRA+wp
  test2011: IN+20%OUT, MERT+retune(wp JK)

Summary MT

- Used data selection for final systems (IN+OUT)
- Sparse lexicalised features to adapt to style and vocabulary of TED talks, larger gains with jackknife tuning
- Compared three tuning setups for sparse features

- On test2010, all systems with sparse features improved over baselines, less systematic differences on test2011
- Best system for de-en:
  test2010: **IN+10%OUT, MERT+retune(wp JK)**
  test2011: **IN+10%OUT, MIRA**
- Best systems for en-fr:
  test2010: **IN+20%OUT, MIRA+wp**
  test2011: **IN+20%OUT, MERT+retune(wp JK)**

Thank you!

Axelrod, A., He, X., and Gao, J. (2011).
Domain adaptation via pseudo in-domain data selection.
In *Proceedings of EMNLP 2011*, Stroudsburg, PA, USA. ACL.

Bisazza, A., Ruiz, N., and Federico, M. (2011).
Fill-up versus interpolation methods for phrase-based SMT adaptation.
In *Proceedings of IWSLT*, California, USA.

Collins, M., Koehn, P., and Kučerová, I. (2005).
Clause restructuring for statistical machine translation.
In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 531–540, Stroudsburg, PA, USA. Association for Computational Linguistics.

Haddow, B. and Koehn, P. (2012).
Analysing the effect of Out-of-Domain data on SMT systems.
In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montréal, Canada. ACL.

Koehn, P. and Knight, K. (2003).
Empirical methods for compound splitting.
In *In Proceedings of EACL*, pages 187–193.

Wuebker, J., Huck, M., Mansour, S., Freitag, M., Feng, M., Peitz, S., Schmidt, C., and Ney, H. (2011).
The RWTH Aachen machine translation system for IWSLT 2011.
In *Proceedings of IWSLT*, California, USA.