RWTH Aachen Machine Translation Systems for IWSLT 2013

Joern Wuebker, Stephan Peitz, Tamer Alkhouli, Jan-Thorsten Peter, Minwei Feng, Markus Freitag and Hermann Ney

surname@cs.rwth-aachen.de

International Workshop on Spoken Language Translation
Heidelberg
05.12.2013

Human Language Technology and Pattern Recognition
Chair of Computer Science 6
Computer Science Department
RWTH Aachen University, Germany

Overview

- **RWTH participated in the following tracks:**
  - **English ASR**
  - **German ASR**
  - **English→French MT**
  - **English↔German MT**
  - **Arabic→English MT**
  - **Chinese→English MT**
  - **English↔Slovenian MT**
  - **English→French SLT**
  - **English→German SLT**

Posters

The RWTH Aachen Machine Translation Systems for IWSLT 2013

Joern Wuebker, Stephan Peitz, Tamer Alkhouli, Jan-Thorsten Peter, Minwei Feng, Markus Freitag and Hermann Ney

The RWTH Aachen German and English LVCSR systems for IWSLT-2013

M. Ali Basha Shaik, Zoltan Tüske, Simon Wiesler, Markus Nußbaum-Thom, Stephan Peitz, Ralf Schlüter and Hermann Ney

EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project

Markus Freitag, Stephan Peitz, Joern Wuebker, Hermann Ney, Nadir Durrani, Matthias Huck, Philipp Koehn, Thanh-Le Ha, Jan Niehues, Mohammed Mediani, Teresa Herrmann, Alex Waibel, Nicola Bertoldi, Mauro Cettolo, Marcello Federico

Outline

1. **Tutorial:**

   How to quickly build a good, efficient and easy to maintain SMT system

   ... from my limited point of view

2. **Taking apart the RWTH MT/SLT systems:**

   The most effective/novel/interesting components

   ... plus an advertisement

3. **Appendix:**

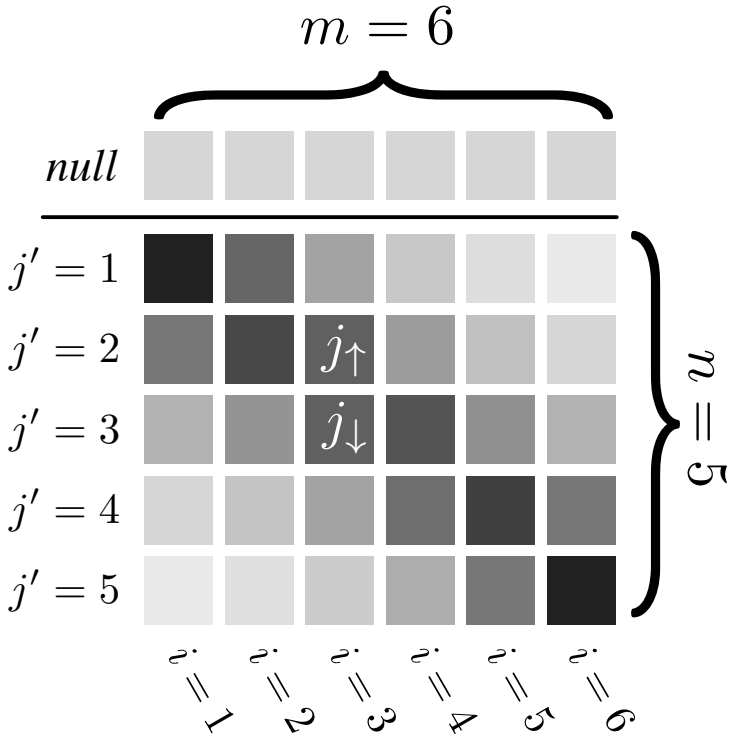
   Overview of the results

   ... basically, lots of tables with even more numbers

fast_align

	De-En		Sl-En		En-Sl	
	BLEU	runtime	BLEU	runtime	BLEU	runtime
GIZA++ (IBM-4)	30.4	60.5h	15.9	13 min	9.6	13 min
fast_align	30.3	4.4h	16.3	1 min	10.5	1 min

- [Dyer & Chahuneau+ 13]
- reparameterized IBM-2 to avoid overfitting
  - alignment model penalizes deviation from diagonal
- source code can be downloaded at `github.com/clab/fast_align`



Data selection

- [Moore & Lewis 10, Mansour & Wuebker$^+$ 11]
- usually no change in performance
- but: smaller and more efficient systems
- used throughout all tasks

- En$\rightarrow$Fr, De$\leftrightarrow$En, Sl$\rightarrow$En
  - monolingual: $\frac{1}{2}$ Shuffled News, $\frac{1}{4}$ Gigaword

- Ar$\rightarrow$En
  - monolingual: down to $\frac{1}{64}$, fine-tuned towards perplexity
  - bilingual: $\frac{1}{16}$ UN data

- hint: don't forget to de-duplicate!

Domain adaptation light (2TM)

	En-Fr	De-En
BLEU	+0.2	+0.9

- similar to fill-up [Bisazza & Ruiz+ 11]
- but even simpler!

- concatenate two phrase tables (in-domain data + all data)
- + indicator feature trained with MERT
- no meta-parameters $\Rightarrow$ no fine-tuning necessary
- little effort, easy to recompute if needed

Outline

1. Tutorial:

   How to quickly build a good, efficient and easy to maintain SMT system

   ... from my limited point of view

2. Taking apart the RWTH MT/SLT systems:

   The most effective/novel/interesting components

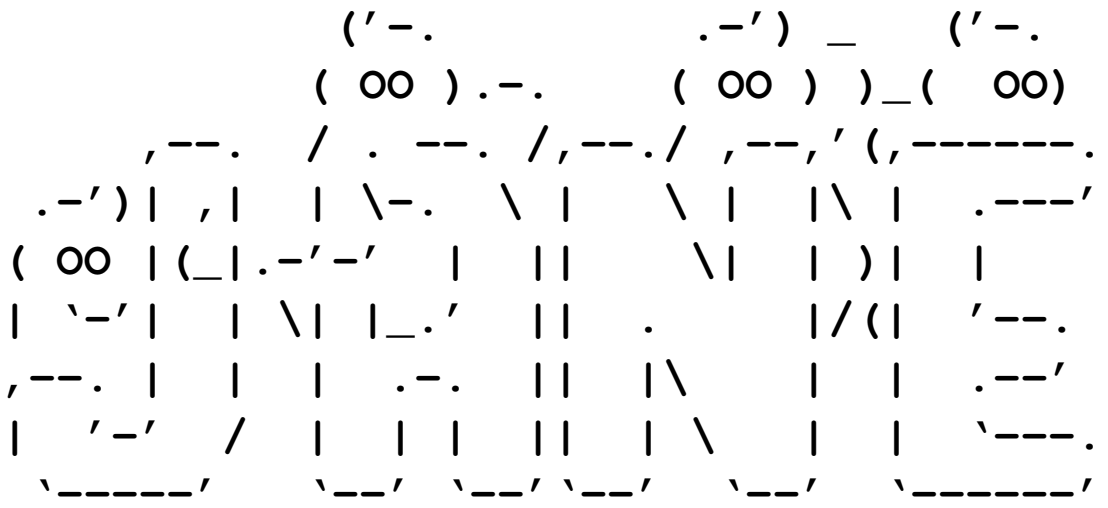
   ... plus an advertisement

3. Appendix:

   Overview of the results

   ... basically, lots of tables with even more numbers

Jane

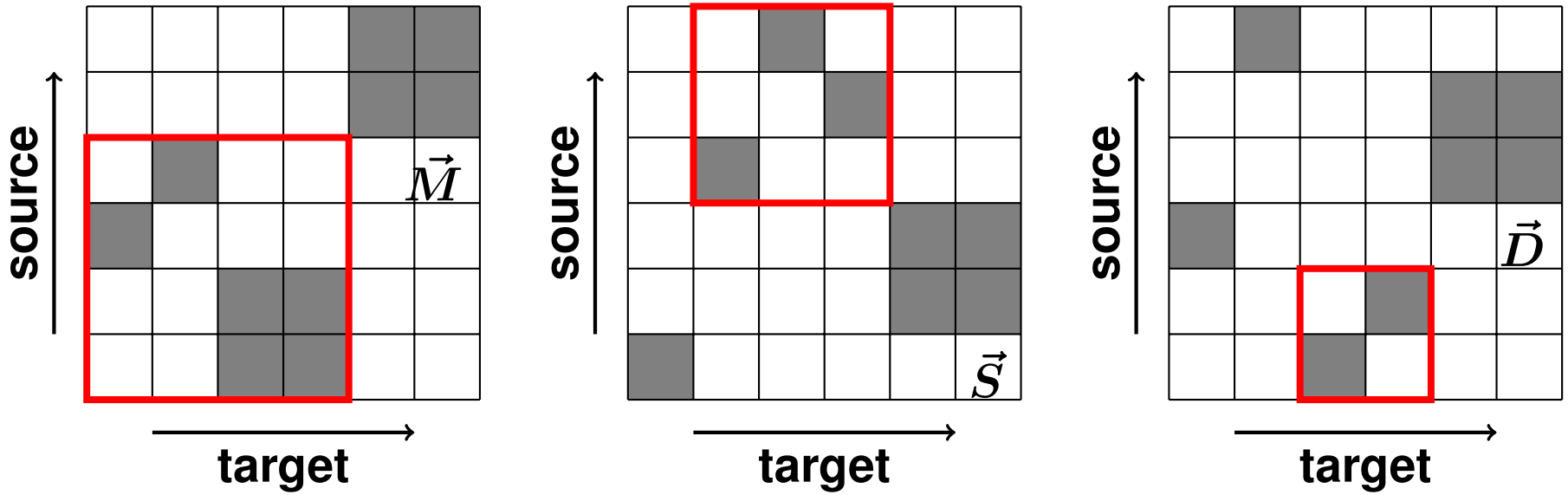


- RWTH's open-source translation toolkit
- includes hierarchical and phrase-based decoder [Wuebker & Huck+ 12]
- applied in all MT and SLT tasks
- `http://www.hltpr.rwth-aachen.de/jane`

Hierarchical Reordering Model (HRM)

	En-Fr	De-En	En-De	Ar-En
BLEU	+0.6	+0.1	+0.3	+1.1

- [Galley & Manning 08]
- phrases are merged into blocks
- orientation relative to largest block containing previous phrase



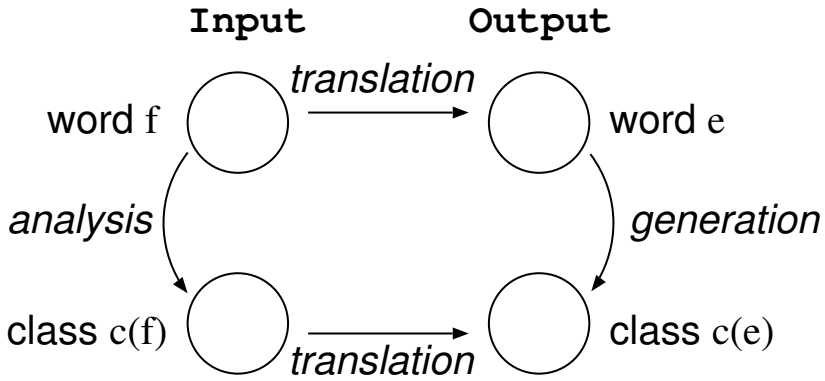
Monotone Swap Discontinuous

Word Class Language Model (wcLM)

	En-Fr	De-En	En-De
BLEU	+0.3	+0.5	+1.0

- [Wuebker & Peitz+ 13]
- use `mkcls` to train word classes
- train language model on word classes
- smaller vocabulary $\Rightarrow$ larger context possible (here: 7-gram)
- can be extended to translation / reordering model (wcTM):

  $\Rightarrow$ equivalent to factored translation model



Maximum Expected BLEU Training (discr.)

	De-En	En-De
BLEU	+0.5	+0.7

- [He & Deng 12]
- train $p(\tilde{e}|\tilde{f})$ and $p(\tilde{f}|\tilde{e})$ w.r.t. expected BLEU score $\langle \text{BLEU} \rangle_\Theta$

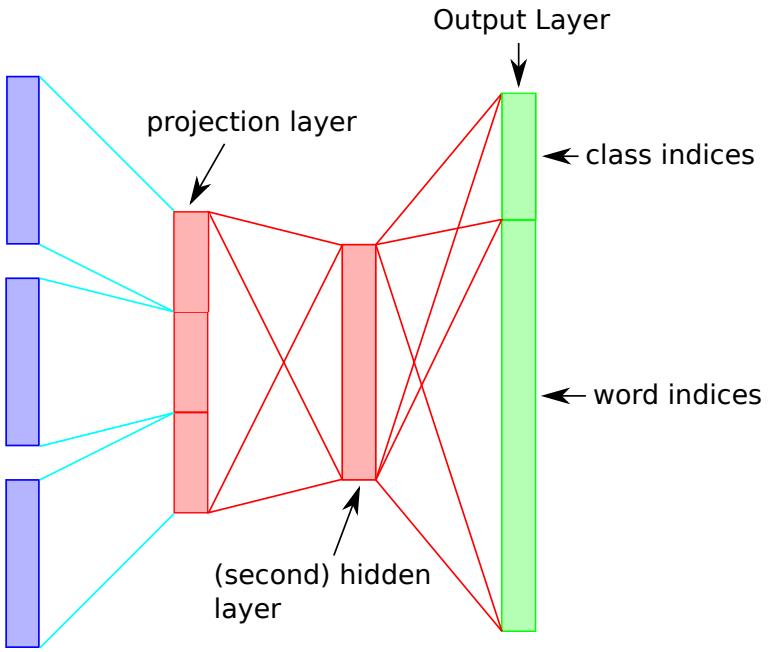
$$\langle \text{BLEU} \rangle_\Theta = \sum_{F \in \mathbb{F}} \sum_{E \in \mathbb{E}} p_\Theta(E, F) \text{BLEU}(E, R(F))$$

- approximate $(\mathbb{F}, \mathbb{E})$ with n-best lists on training data (here: TED portion)

Continuous Space Language Model (CSLM)

	En-Fr	En-De
BLEU	+0.2	+0.6

- similar to [Schwenk & Rousseau+ 12]
- clustered output layer [Goodman 01, Morin & Bengio 05]
- inputs: shortlist + word features
- rescoring on 200-best lists
- 7-gram language model
- implemented using Theano [Bergstra & Breuleux+ 10]



Continuous Space Language Model (CSLM)

- trained on TED data
- word features
  - most frequent words (shortlist)
  - most frequent prefixes of length n
  - most frequent suffixes of length n
  - most frequent substrings of length n
- number of features:

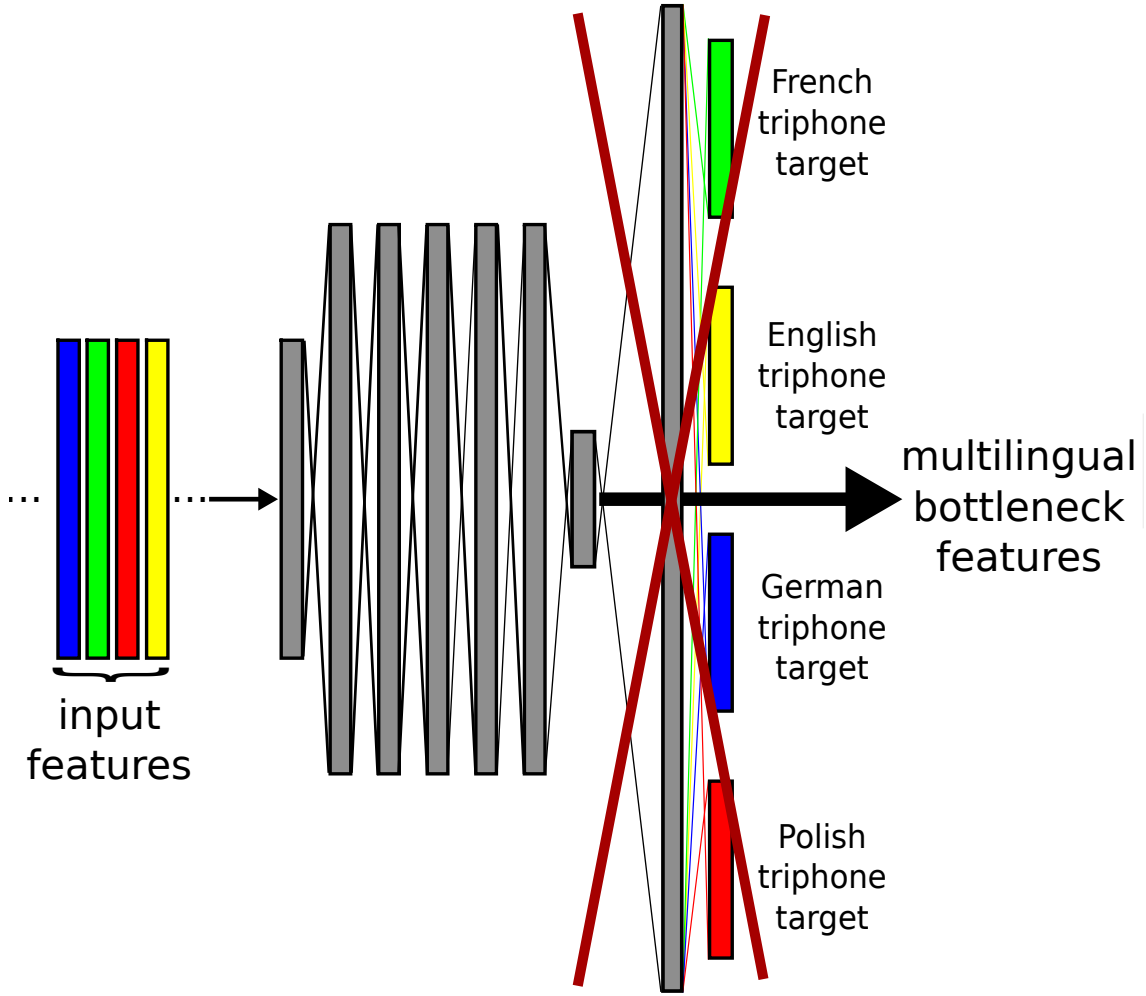
	shortlist	n-prefixes			n-suffixes			n-substrings		total
$n =$		1	2	3	1	2	3	2	3	
French	2000	51	130	435	48	84	256	249	1839	5092
German	2000	44	124	407	41	93	243	296	2542	5790

The English ASR system

- acoustic model
  - acoustic training data: 962 h
    (web, broadcast news, parliament speeches, TED talks)
  - 16 Mel-cepstral coefficients (MFCC) $\Rightarrow$ 45 features
  - multilingual bottleneck MRASTA features
    trained on German, English, French and Polish data
    [Tüske & Schlüter$^+$ 13a, Tüske & Schlüter$^+$ 13b]
  - speaker adaptation
- language model
  - training data: 6 Billion running words
- CN-based system combination of two systems

The English ASR system

Multilingual bottleneck MLP features



Enriching ASR output

- [Peitz & Freitag+ 11]
- enriching ASR output is modeled as machine translation
- reintroduction of case information and punctuation with MT system
- system is optimized with MERT on WER
- performed as preprocessing *before* translation
  - no modifications to the translation system
  - optional: rerun MERT on enriched ASR output (+0.3 BLEU on En-Fr)

ASR output	and you say doc what should i do
enriched ASR output	and you say , " Doc , what should I do ?
translation	et vous dites , " Docteur , que dois-je faire ?
reference translation	vous dites : " Que dois-je faire , docteur ? "

Outline

1. Tutorial:

   How to quickly build a good, efficient and easy to maintain SMT system

   ... from my limited point of view

2. Taking apart the RWTH MT/SLT systems:

   The most effective/novel/interesting components

   ... plus an advertisement

3. Appendix:

   Overview of the results

   ... basically, lots of tables with even more numbers

En→Fr

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
SCSS allData	28.3	55.7	31.9	49.8
+HRM	28.7	55.3	32.5	49.2
+2TM	29.2	54.7	32.7	48.9
+GW *	29.5	54.6	32.9	48.9
+DWL	29.8	54.3	33.2	48.5
+wcLM	29.7	54.2	33.5	48.3
+CSLM	30.0	53.8	33.7	48.0

* was used for the SLT track

- HRM: hierarchical reordering model
- 2TM: 2 translation models (domain adaptation light)
- GW: GigaWord
- DWL: discriminative word lexicon [Mauser & Hasan$^+$ 09]
- wcLM: word class language model
- CSLM: continuous space language model

En→De

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
SCSS TED	22.0	56.7	21.9	57.3
SCSS allData	22.7	56.1	22.3	57.2
+HRM	23.3	55.5	22.6	57.7
+wcLM	24.2	54.5	23.6	55.9
+discr. *	24.6	54.1	24.3	55.4
+CSLM	24.7	53.7	24.9	54.7

* was used for the SLT track

- HRM: hierarchical reordering model
- wcLM: word class language model
- discr.: maximum expected BLEU training
- 2TM: 2 translation models (domain adaptation light)

De→En

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
SCSS TED	31.5	47.6	30.0	49.2
SCSS allData	32.8	46.4	30.3	48.9
+HRM	33.0	46.1	30.4	48.9
+wcLM	33.5	45.8	30.9	48.4
+discr.	33.9	45.0	31.4	47.5
+2TM	34.2	45.2	32.3	47.4

- HRM: hierarchical reordering model
- wcLM: word class language model
- discr.: maximum expected BLEU training
- CSLM: continuous space language model

Ar→En

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
SCSS TED	27.4	52.0	25.7	55.1
+HRM	27.9	51.9	26.8	54.9
+UN	28.4	51.9	25.7	55.6
+UN interpolated	28.3	51.1	26.9	54.1
+sngLM	28.8	50.7	26.8	54.1
+x-entropy	28.6	51.8	26.7	55.0
+x-entropy+IBM-1	28.8	51.0	27.0	54.2

- HRM: hierarchical reordering model
- UN: UN data
- UN interpolated: phrase table interpolation
- sngLM: unpruned language model
- x-entropy: bilingual data selection based on LM cross-entropy
- x-entropy+IBM-1: bilingual data selection based on LM and IBM-1 cross-entropy

Zh→En

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
PBT-2012-standard	11.5	80.7	13.0	76.4
PBT-2012-reverse	11.7	80.9	13.6	75.5
HPBT-2012-standard	12.3	79.8	14.2	74.6
HPBT-2012-reverse	12.8	79.4	14.6	74.1
HPBT-2013-standard	12.4	79.5	14.5	74.1
HPBT-2013-reverse	12.6	79.4	14.4	74.3
system combination	13.5	78.5	15.1	73.6

- PBT: phrase-based decoder [Zens & Ney 08]
- HPBT: hierarchical phrase-based decoder
- reverse: reverse order models [Freitag & Feng+ 13]

Sl→En

system	dev1		dev2	
	BLEU	TER	BLEU	TER
SCSS GIZA++	17.6	65.7	15.9	67.6
SCSS *fast_align*	18.0	64.8	16.3	66.1
+wcLM	18.2	62.9	16.5	64.6
+wcTM +PRO	18.6	63.0	16.5	64.3
+discr.	18.8	62.6	16.9	63.9

- wcLM: word class language model
- wcTM: word class translation / reordering model
- PRO: [Hopkins & May 11]
- discr.: maximum expected BLEU training

En→Sl

system	dev1		dev2	
	BLEU	TER	BLEU	TER
SCSS GIZA++	11.3	70.5	9.6	71.4
SCSS *fast_align*	11.4	70.3	10.5	69.6
+wcLM	12.0	69.8	10.1	69.9
+wcTM	11.9	70.3	10.4	69.9
+discr.	11.9	70.2	10.7	69.7

- wcLM: word class language model
- wcTM: word class translation / reordering model
- discr.: maximum expected BLEU training

Thank you for your attention

Joern Wuebker

wuebker@cs.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/

References

[Bergstra & Breuleux+ 10] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, Y. Bengio: Theano: a CPU and GPU Math Expression Compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation. 13

[Bisazza & Ruiz+ 11] A. Bisazza, N. Ruiz, M. Federico: Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *International Workshop on Spoken Language Translation*, San Francisco, California, Dec. 2011. 7

[Dyer & Chahuneau+ 13] C. Dyer, V. Chahuneau, N.A. Smith: A Simple, Fast and Effective Reparameterization of IBM Model 2. In *Proceedings of NAACL-HLT*, pp. 644–648, Atlanta, Georgia, June 2013. 5

[Freitag & Feng+ 13] M. Freitag, M. Feng, M. Huck, S. Peitz, H. Ney: Reverse Word Order Models. In *Machine Translation Summit*, Nice, France, Sept. 2013. 23

[Galley & Manning 08] M. Galley, C.D. Manning: A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pp. 848–856, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. 10

[Goodman 01] J. Goodman: Classes for Fast Maximum Entropy Training. *CoRR*, Vol. cs.CL/0108006, 2001. 13

[He & Deng 12] X. He, L. Deng: Maximum Expected BLEU Training of Phrase and Lexicon Translation Models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 292–301, Jeju, Republic of Korea, Jul 2012. 12

[Hopkins & May 11] M. Hopkins, J. May: Tuning as ranking. In *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1352–1362, Edinburgh, Scotland, July 2011. 24, 25

[Mansour & Wuebker+ 11] S. Mansour, J. Wuebker, H. Ney: Combining Translation and Language Model Scoring for Domain-Specific Data Filtering. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, Dec. 2011. 6

[Mauser & Hasan+ 09] A. Mauser, S. Hasan, H. Ney: Extending Statistical Machine Translation with Discriminative and Trigger-Based Lexicon Models. In *Conference on Empirical Methods in Natural Language Processing*, pp. 210–217, Singapore, Aug. 2009. 19

[Moore & Lewis 10] R. Moore, W. Lewis: Intelligent Selection of Language Model Training Data. In *ACL (Short Papers)*, pp. 220–224, Uppsala, Sweden, July

2010. 6

[Morin & Bengio 05] F. Morin, Y. Bengio: Hierarchical Probabilistic Neural Network Language Model. In R.G. Cowell, Z. Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pp. 246–252. Society for Artificial Intelligence and Statistics, 2005. 13

[Peitz & Freitag+ 11] S. Peitz, M. Freitag, A. Mauser, H. Ney: Modeling Punctuation Prediction as Machine Translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, Dec. 2011. 17

[Schwenk & Rousseau+ 12] H. Schwenk, A. Rousseau, M. Attik: Large, Pruned or Continuous Space Language Models on a GPU for Statistical Machine Translation. In *NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pp. 11–19, Montréal, Canada, June 2012. 13

[Tüske & Schlüter+ 13a] Z. Tüske, R. Schlüter, H. Ney: Deep hierarchical bottleneck MRASTA features for LVCSR. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 6970–6974, Vancouver, Canada, May 2013. 15

[Tüske & Schlüter+ 13b] Z. Tüske, R. Schlüter, H. Ney: Multilingual Hierarchical MRASTA Features for ASR. In *Interspeech*, pp. 2222–2226, Lyon, France, Aug.

2013. 15

[Wuebker & Huck+ 12] J. Wuebker, M. Huck, S. Peitz, M. Nuhn, M. Freitag, J.T. Peter, S. Mansour, H. Ney: Jane 2: Open Source Phrase-based and Hierarchical Statistical Machine Translation. In *International Conference on Computational Linguistics*, Mumbai, India, Dec. 2012. To appear. 9

[Wuebker & Peitz+ 13] J. Wuebker, S. Peitz, F. Rietig, H. Ney: Improving Statistical Machine Translation with Word Class Models. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1377–1381, Seattle, USA, Oct. 2013. 11

[Zens & Ney 08] R. Zens, H. Ney: Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation. In *International Workshop on Spoken Language Translation*, pp. 195–205, Honolulu, Hawaii, Oct. 2008. 23