The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation

Arne Mauser, Richard Zens, Evgeny Matusov, Saša Hasan, Hermann Ney

{lastname}@i6.informatik.rwth-aachen.de

International Workshop on Spoken Language Translation (IWSLT) 2006
November 27, 2006

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Outline

RWTH System 2006

- overview
- improvements since 2005
- setup
- results

System Combination (TC-Star project partners)

- method
- setup and results

Main Facts

Standard phrase-based statistical MT system: 2 pass-approach

1. translation (N-best lists)

- **log-linear model**
- **DP-based search**
- **standard phrase extraction from GIZA++ alignments**
- **features: phrase lexica, word lexica, language model, distance-based reordering, word and phrase penalty**
- **(up to) 10k-best lists**

2. rescoring/reranking

- **sentence length models**
- **lexicon models**
- **additional language models**

minimum error training (for BLEU) in both passes

From 2005 to 2006: Improvements

- translation: phrase count features
  - smoothing of phrase probabilities
- rescoring: sentences mixture language model
  - reflect topic depencedies in the language model
- rescoring: sentence length posterior probability
  - explicitly model sentence length.

Phrase count features

Motivation:

- rare phrases are overestimated
- estimated probabilities not reliable

Idea:

- adjust probabilities of rare phrases
- "mark" phrases with a occurrence count below a given threshold

$$h_{\mathbf{C},\tau}(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^{K} [N(\tilde{f}_k, \tilde{e}_k) \leq \tau]$$

- include these marker as a binary feature in the log-linear translation model
- using positive weights in log-linear combination: penalty for "infrequent" phrases

τ: threshold, $N(\tilde{f}_k, \tilde{e}_k)$: bilingual phrase count,
s_1^K: segmentation of the source sentence

Settings: 3 features for $\tau = 0.9, 1.9, 2.9$

Sentence-level Mixture Language Model

Motivation:

- represent topic dependencies in the language model [Iyer & Ostendorf 99]

Idea:

- combine M different language models with a global one
- training sentences are grouped automatically (ML-clustering)
- model weights λ_m are trained on a development set

$$p(e_1^I) = \sum_{m=0}^{M} \lambda_m \left[\prod_{i=1}^{I} p_m(e_i|e_{i-1}, e_{i-2}) \right]$$

Settings: 5gram-LMs, 10 clusters

Examples for found clusters: what/how-questions, do/can-questions, request (I would like to/I want to), ...

Sentence Length Posterior

Motivation:

- **sentence length is important for MT quality**
- **usually not modelled explicitly**

compute posterior probability of sentence length [Zens & Ney 06]

$$h_{\mathrm{SL}}(f_1^J, e_1^I) = \log \sum_{\tilde{e}_1^I} p(\tilde{e}_1^I | f_1^J)$$

the sum is carried out only over those target hypotheses that have length I

approximation of sum over N-best hypotheses

Special Issues 2006: Punctuation and Case

source language input data did not contain punctuation marks

output should contain punctuation and case

punctuation

- **Idea: let the translation system do the work**
- **remove punctuation from source language training data**
- **method found to be superior to insertion of punctuation on the source or target side**
- **see talk by Matusov et al. later today for more details**

case restoration

- **standard approach using disambig tool**
- **trained only on provided data**
- **casing error: 2% of the words in the DEV4 set**

Summary Procedure

training and tuning:

1. remove punctuation from source and preprocess target
2. train alignments, extract phrases
3. optimize first pass and generate N-best list
4. select, train and optimize models for rescoring
5. add additional data (dev-corpora)

translation:

1. translate test data (N-best list generation and rescoring)
2. postprocessing and truecasing

systems for Japanese-English and Chinese-English almost identical (less reordering for J-E)

tuning for TEXT, no changes for ASR output

Results

Evaluation submissions

Translation Direction	Input	Accuracy Measures			Error Rates	
		BLEU [%]	NIST	Meteor [%]	WER [%]	PER [%]
Chinese-English	Correct	24.2	6.10	50.3	66.7	50.9
	Read	21.1	5.40	44.3	69.5	55.3
	Spont	19.0	5.05	42.0	71.2	57.1
Japanese-English	Correct	23.7	5.92	48.9	68.5	51.5
	Read	21.4	5.65	45.7	70.7	53.8

same system for all input types (optimized on text)

questions to be answered by contrastive submissions:

- What is the effect of rescoring?
- What is the effect of adding the dev corpora?

Results - Contrastive Submissions

Translation Direction	System	Accuracy Measures			Error Rates	
		BLEU [%]	NIST	Meteor [%]	WER [%]	PER [%]
Chinese-English	final	24.2	6.10	50.3	66.7	50.9
	no rescoring	22.9	6.02	50.3	67.4	51.0
	no dev	22.3	5.90	49.4	68.5	51.5
	without both	20.9	5.72	48.6	68.1	52.4
Japanese-English	final	23.7	5.92	48.9	68.5	51.5
	no rescoring	23.3	5.84	47.9	68.4	51.9
	no dev	21.0	5.67	47.5	68.7	52.4
	without both	21.5	5.64	46.7	69.3	52.4

rescoring mostly optimized for Chinese-English

small task: dev-corpora helpful as additional training data

Models in Detail

effect of successively adding models for the Chinese-English IWSLT 2006 development set (DEV4)

System	BLEU [%]	NIST	WER [%]	PER [%]
Baseline	21.2	6.18	69.2	54.5
+Countfeatures	21.9	6.31	66.4	50.8
+Clustered Language Model	22.5	6.09	63.7	49.7
+Length Models	23.0	6.36	66.7	51.3
+Sentence Mixtures	23.2	6.30	65.6	50.4
+Deletion Model	23.4	6.37	66.1	50.4
+IBM1 lexicon model	23.5	6.33	64.8	49.4

consistent improvements on DEV4 (+1.6 BLEU%) and TEST (+1.3 BLEU%)

Progress over time

RWTH in the IWSLT evaluations 2004-2006 on the IWSLT 2005 evaluation set

Translation Direction	System	BLEU [%]	NIST	WER [%]	PER [%]
Chinese-English	2004	40.4	8.59	52.4	42.2
	2005	46.3	8.73	47.4	39.7
	2006	48.8	8.56	47.3	39.2
	2006 (40k)	51.4	9.00	40.0	33.2
Japanese-English	2004	44.8	9.41	50.0	37.7
	2005	49.8	9.52	46.5	36.8
	2006	56.5	8.72	41.9	32.8
	2006 (40k)	57.1	8.69	41.8	33.6

separating the effect of additional data

improved English preprocessing for Japanese-English

TC-Star System Combination

TC-Star System Combination

TC-Star: Translation and Corpora for Speech-to-Speech Translation
European project aimed at translating the Eurpean Parliament plenary speeches

several partners of the TC-Star Project participated in IWSLT

creating “TC-Star system” by combining the output of all partners.

Participants: ITC-irst, RWTH, UKA, TALP (UPC)

Tracks: Chinese-English correct transcription and read speech only

System Combination

Introduction

- different MT systems make different errors
- consensus translation:
  - align outputs of multiple systems
  - majority voting over aligned words
- a possibly new translation can be generated
- consider reordering of words/phrases

TC-Star System Combination

Idea of the algorithm [Matusov & Ueffing+ 06]:

- **align different MT system outputs for each source sentence:**
  - **alignment as in SMT training (GIZA++)**
  - **between sentences of the same language**
  - **trained over the whole test document**
- **construct a confusion of hypotheses from alignment**
- **select best consensus translations using**
  - **global system probabilities**
  - **other statistical models (rescoring)**

Building Confusion Networks

- consider alignments of hypotheses E_n to a primary hypothesis E_m $(n, m \in \{1, ..., M\}, n \neq m)$
- reorder each E_n based on the alignment with E_m
- merge monotone aligments into one confusion network
- M confusion networks are created in total, by letting each hypothesis be the primary one

Building Confusion Networks: Example

original hypotheses	1. would you like coffee or tea 2. would you have tea or coffee 3. would you like your coffee or 4. I have some coffee tea would you like
alignment and reordering	would|would you|you have|like coffee|coffee or|or tea|tea would|would you|you like|like your|$ coffee|coffee or|or $\|tea I\|$ would|would you|you like|like have|$ some|$ coffee|coffee $|or tea|tea
confusion network	$ would you like $ $ coffee or tea $ would you have $ $ coffee or tea $ would you like your $ coffee or $ I would you like have some coffee $ tea

Extracting Consensus Translation

- introduce global system probabilities
  - tuned manually based on the performance of the individual systems on a development set
- perform "voting" on each of the M confusion networks:

0.25	$	would	you	like	$	$	coffee	or	tea
0.35	$	would	you	have	$	$	coffee	or	tea
0.1	$	would	you	like	your	$	coffee	or	$
0.3	I	would	you	like	have	some	coffee	$	tea
voting	$/0.7 I/0.3	would/1.0	you/1.0	have/0.35 like/0.65	$/0.6 your/0.1 have/0.3	$/0.7 some/0.3	coffee/1.0	or/0.7 $/0.3	tea/0.9 $/0.1

- unite M confusion networks into one automaton
- extract the single-best path as the consensus translation
- or extract N best paths for further processing (e.g. rescoring)

Results for the TC-Star System

Input	System	Accuracy Measures			Error Rates	
		BLEU [%]	NIST	Meteor [%]	WER [%]	PER [%]
Correct	TC-Star	24.1	6.40	51.8	65.4	49.8
	RWTH	24.2	6.10	50.3	66.7	50.9
	UKA/CMU	20.0	5.76	47.3	68.4	54.6
	TALP (UPC)	19.2	5.40	47.5	66.7	54.6
	ITC-irst	18.4	5.83	48.5	68.6	53.2
Read	TC-Star	20.0	5.59	46.0	69.1	54.7
	RWTH	21.1	5.40	44.3	69.5	55.3
	UKA/CMU	17.1	5.08	42.3	70.7	57.9
	TALP/UPC	16.5	4.89	42.7	69.8	58.5
	ITC-irst	15.6	5.22	43.7	72.0	57.9

system combination leads to improvements over the best system in all measures except BLEU

problem: heterogeneous performance of combined systems

Conclusions

RWTH System

- standard statistical phrase-based translation model
- 2-pass approach with rescoring
- new features and models: phrase count, sentence mixture, length model
- best performing system for Japanese-English and Chinese-English

TC-Star System

- consensus translation of several mt system outputs by alignment and voting
- leads to improvement over the best single system
- dependent on the individual system scores

The systems were the two best performing in Chinese-English

Thank you for your attention

Arne Mauser

mauser@i6.informatik.rwth-aachen.de

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

This work was in part funded by the European Union under the integrated project TC-STAR – Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738).

References

[Banerjee & Lavie 05] S. Banerjee, A. Lavie: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, MI, June 2005.

[Bender & Zens+ 04] O. Bender, R. Zens, E. Matusov, H. Ney: Alignment Templates: the RWTH SMT System. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pp. 79–84, Kyoto, Japan, September 2004.

[Brown & Cocke+ 90] P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, P.S. Roossin: A Statistical Approach to Machine Translation. *Computational Linguistics*, Vol. 16, No. 2, pp. 79–85, June 1990.

[Doddington 02] G. Doddington: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. ARPA Workshop on Human Language Technology*, 2002.

[Fiscus 97] J. Fiscus: A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

[Graham & Knuth+ 94] R.L. Graham, D.E. Knuth, O. Patashnik: *Concrete Mathematics*. Addison-Wesley Publishing Company, Reading, Mass., 2 edition, 1994.

[Hasan & Ney 05] S. Hasan, H. Ney: Clustered Language Models based on Regular Expressions for SMT. In *Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT)*, Budapest, Hungary, May 2005.

[Iyer & Ostendorf 99] R.M. Iyer, M. Ostendorf: Modeling Long Distance Dependence in Language: Topic Mixtures Versus Dynamic Cache Models. *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 1, pp. 30–39, 1999. 6

[Kanthak & Vilar+ 05] S. Kanthak, D. Vilar, E. Matusov, R. Zens, H. Ney: Novel Reordering Approaches in Phrase-Based Statistical Machine Translation. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pp. 167–174, Ann Arbor, MI, June 2005.

[Matusov & Ney 05] E. Matusov, H. Ney: Phrase-based Translation of Speech Recognizer Word Lattices using Loglinear Model Combination. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Cancun, Mexiko, Nov/Dec 2005. To appear.

[Matusov & Ueffing+ 06] E. Matusov, N. Ueffing, H. Ney: Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. In *Proceedings of EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*, pp. 33–40, Trento, Italy, April 2006. 17

[Och 03] F.J. Och: Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 160–167, Sapporo, Japan, July 2003.

[Och & Gildea+ 03] F.J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, D. Radev: Syntax for Statistical Machine Translation. Technical report, Johns Hopkins University 2003 Summer Workshop on Language Engineering, Center for Language and Speech Processing, Baltimore, MD, August 2003.

[Och & Ney 02] F.J. Och, H. Ney: Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 295–302, Philadelphia, PA, July 2002.

[Och & Ney 03] F.J. Och, H. Ney: A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, March 2003.

[Och & Tillmann+ 99] F.J. Och, C. Tillmann, H. Ney: Improved Alignment Models for Statistical Machine Translation. In *Proc. Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 20–28, University of Maryland, College Park, MD, June 1999.

[Papineni & Roukos+ 02] K. Papineni, S. Roukos, T. Ward, W.J. Zhu: Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, Philadelphia, PA, July 2002.

[Press & Teukolsky+ 02] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery: *Numerical Recipes in C++*. Cambridge University Press, Cambridge, UK, 2002.

[Stolcke 02] A. Stolcke: SRILM – An Extensible Language Modeling Toolkit. In *Proc. Int. Conf. on Spoken Language Processing*, Vol. 2, pp. 901–904, Denver, CO, 2002.

[Takezawa & Sumita+ 02] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, S. Yamamoto: Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of the Third Int. Conf. on Language Resources and Evaluation (LREC)*, pp. 147–152, Las Palmas, Spain, May 2002.

[Tillmann & Ney 03] C. Tillmann, H. Ney: Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation. *Computational Linguistics*, Vol. 29, No. 1, pp. 97–133, March 2003.

[Ueffing & Och+ 02] N. Ueffing, F.J. Och, H. Ney: Generation of Word Graphs in Statistical Machine Translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pp. 156–163, Philadelphia, PA, July 2002.

[Zens & Bender+ 05] R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, H. Ney: The RWTH Phrase-based Statistical Machine Translation System. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pp. 155–162, Pittsburgh, PA, October 2005.

[Zens & Ney+ 04] R. Zens, H. Ney, T. Watanabe, E. Sumita: Reordering Constraints for Phrase-Based Statistical Machine Translation. In *COLING '04: The 20th Int. Conf. on Computational Linguistics*, pp. 205–211, Geneva, Switzerland, August 2004.

[Zens & Ney 05] R. Zens, H. Ney: Word Graphs for Statistical Machine Translation. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pp. 191–198, Ann Arbor, MI, June 2005.

[Zens & Ney 06] R. Zens, H. Ney: N-gram Posterior Probabilities for Statistical Machine Translation. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Proceedings of the Workshop on Statistical Machine Translation*, pp. 72–77, New York City, NY, June 2006. 7

[Zens & Och+ 02] R. Zens, F.J. Och, H. Ney: Phrase-Based Statistical Machine Translation. In M. Jarke, J. Koehler, G. Lakemeyer, editors, *25th German Conf. on Artificial Intelligence (KI2002)*, Vol. 2479 of *Lecture Notes in Artificial Intelligence (LNAI)*, pp. 18–32, Aachen, Germany, September 2002. Springer Verlag.