International Workshop on Spoken Language Translation
Kyoto, Japan
September 30 - October 1, 2004

Alignment Templates: the RWTH SMT System

Oliver Bender, Richard Zens, Evgeny Matusov, and Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik VI
Computer Science Department
RWTH Aachen University
D-52056 Aachen

Content

1. overview: statistical machine translation
2. loglinear models
3. alignment templates
4. feature functions
5. minimum error training
6. n-best lists and rescoring
7. experimental results
8. summary

Related work

- F. J. Och and H. Ney. 2002. 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.
- 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, and D. Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 161–168, Boston, MA, May.
- A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. Intl. Conf. Spoken Language Processing*, pp. 901–904, Denver, CO, September.
- D. Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, vol. 23, no. 3, pp. 377–403, September.

Overview: Statistical Machine Translation

- **source string $f_1^J = f_1 ... f_j ... f_J$ to be translated into a target string $e_1^I = e_1 ... e_i ... e_I$.**
- **classical source-channel approach:**

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I | f_1^J) \right\}$$
$$= \underset{e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I) \cdot Pr(f_1^J | e_1^I) \right\}$$

- $Pr(f_1^J | e_1^I)$**: translation model (usually can be further decomposed into alignment and lexicon model)**
- $Pr(e_1^I)$**: language model**

Loglinear models

- alternative: direct modeling of the posterior probability $Pr(e_1^I|f_1^J)$
- use a loglinear model (Och and Ney 2002):

$$Pr(e_1^I|f_1^J) = p_{\lambda_1^M}(e_1^I|f_1^J) = \frac{\exp\left[\sum\limits_{m=1}^{M}\lambda_m h_m(e_1^I, f_1^J)\right]}{\sum\limits_{{e'}_1^I}\exp\left[\sum\limits_{m=1}^{M}\lambda_m h_m({e'}_1^I, f_1^J)\right]}$$

- decision rule:

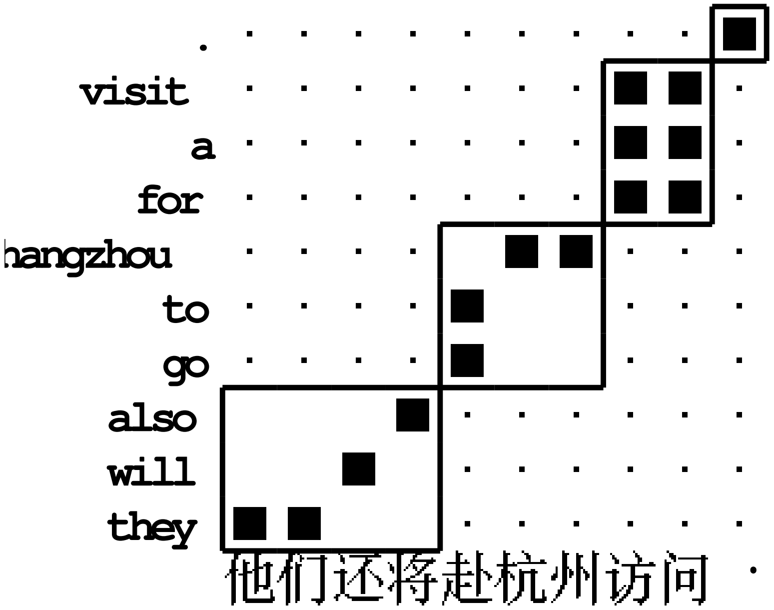
$$\hat{e}_1^I = \operatorname*{argmax}_{e_1^I}\left\{\sum_{m=1}^{M}\lambda_m h_m(e_1^I, f_1^J)\right\}$$

- advantages:
  - easy integration of additional models/feature functions h_m
  - minimum error training of model scaling factors λ_m

Alignment Templates

- primary translation model: alignment templates
- describes the alignment between sequences of source and target words
- automatically trained word classes are used instead of words for better generalization
- translation model incorporates:
  - phrase alignment probability
  - probability to apply an alignment template
  - phrase translation probability
- alignment templates extracted automatically from automatic word alignments

Alignment Templates: Example



- alignment A is a mapping from source sentence positions to target sentence positions $a_1 ... a_J$, $a_j \in \{0, \ldots, I\}$.
- alignment may contain connections $a_j = 0$ with the 'empty' word e_0
- alignments are created automatically with GIZA++ using IBM-1, HMM, and IBM-4 models

Alignment Combination Heuristics

- word alignments A_1 and A_2 are trained in source-to-target and target-to-source direction, respectively
- such alignments contain many-to-one mappings in one direction only
- alignment combination depends on the particular language pair
- best translation results achieved:
  - Chinese-English: using alignments which only allow many-to-one mappings of English words
  - Japanese-English: using "refined" alignments
    - extend intersection $A_1 \cap A_2$ by additional points
    - add a new point if either a horizontal or a vertical direct neighbor point exists

Base Models Used in Search

- **alignment templates**
- **single-word translation model** $p(e|f)$
- **word-based trigram language model**
- **class-based five-gram language model**
- **word penalty model**
- **phrase penalty model**
- **penalty for alignment template reorderings**

Minimum Error Training

- optimize the model scaling factors λ_1^M
- training criterion: minimal number of errors on a development corpus
- optimization with respect to a certain automatic translation score ($100 -$ NIST, $1 -$ BLEU, WER)
- use the downhill simplex optimization algorithm
- translate the whole development corpus in each iteration of the algorithm
- algorithm converges after about 200 iterations

Search

- search characteristics:
  - reordering: within alignment templates: fixed in training
  - reordering of alignment templates: unconstrained or ITG (Japanese-English)
  - search organization along target string positions
  - beam search to handle the huge search space
- generation of n-best lists:
  - during search, generate word graphs
  - using the A^* search algorithm,
    compute n-best lists from the word graphs

Additional n-best List Features

- **(inverse) IBM-1 lexicon model $p(f|e)$ (as trained with GIZA++)**

  $+$ **captures lexical co-occurrences, helpful for translation adequacy**

- **deletion model**

  $+$ **penalizes too short translation hypotheses**

- **high-order n-gram language models ($n = 4, 5, ..., 9$)**

  $+$ **enrich the system with knowledge about longer target language phrases**

Deletion Model

- **the produced translations are often shorter than the reference translations**
- **longer hypotheses are to be favored**
- **deletion model feature (Och et al. 2004): for a given threshold α:**
  - **count the number of source words, for which the IBM-1 translation probability given any of the target words in the hypothesis is below α.**
  - **use several features with different values of α ($0.1, 0.01$, etc.)**
- **threshold α tuned on a development corpus**

Experimental results

- IWSLT 2004 Evaluation
- rescoring improvements

Evaluation Methodology

- subjective evaluation as specified by the IWSLT 2004 consortium
  - translation fluency: from 1 (“incomprehensible”) to 5 (“flawless English”)
  - translation adequacy: how much information from a gold standard translation is contained in the hypothesis, from 1 (“none”) to 5 (“all”)
- objective evaluation: different automatic metrics computed using multiple references
  - Word Error Rate (mWER)
  - Position-Independent Word Error Rate (mPER)
  - BLEU score
  - NIST score
  - GTM score

BTEC Chinese-English Supplied Corpus Statistics

		Chinese	English
train	sentences	20 000	
	words	182 904	160 523
	singletons	3 525	2948
	vocabulary	7 643	6 982
dev	sentences	506	
	words	3 515	3 595
test	sentences	500	
	words	3 794	–

BTEC Japanese-English Supplied Corpus Statistics

		Japanese	English
train	sentences	20 000	
	words	209 012	160 427
	singletons	4 108	2 956
	vocabulary	9 277	6 932
dev	sentences	506	
	words	4 374	3 595
test	sentences	500	
	words	4 370	–

BTEC Japanese-English Unrestricted Data Track Corpus Statistics

- additional resources:
  - full BTEC 1 Japanese-English corpus
  - Spoken Language Database (dialogs, hotel reservation domain)
- kindly provided by ATR

		Japanese	English
train	sentences	240 672	
	words	1 974 407	1 770 190
	singletons	8 975	3 658
	vocabulary	26 037	14 301
dev	sentences	506	
	words	3 515	3 595
test	sentences	500	
	words	3 794	–

Official Evaluation Results

Language Pair	Data Track	Automatic Evaluation mWER [%]	mPER [%]	BLEU [%]	NIST	GTM [%]	Subj. Evaluation Fluency	Adequacy
CE	Small	45.6	39.0	40.9	8.55	72.1	3.36	3.34
JE	Small	41.9	33.8	45.3	9.49	76.4	3.48	3.41
	Unrestricted	30.6	24.9	61.9	10.72	79.7	4.04	4.07

- balanced fluency/adequacy scores
- NIST score has the highest correlation with subjective ratings

Rescoring Improvements - Chinese-English

- error rates and scores on the development corpus (CSTAR 2003 test set)
- best overall performance achieved when optimizing the model scaling factors with respect to the NIST score
- base model scaling factors optimized using a narrow beam
- n-best lists created using a broader beam
- each added feature results in performance gain

System	Error Rates		Accuracy Measures	
	mWER [%]	mPER [%]	BLEU [%]	NIST
baseline	55.2	45.6	34.8	7.76
broad beam	53.4	45.3	33.6	7.63
+ IBM-1 lexicon	50.9	42.1	36.4	8.06
+ deletion model	50.6	42.2	37.1	8.07
+ 9-gram LM	50.6	42.2	38.0	8.14

Rescoring Improvements - Japanese-English

- error rates and scores on the development corpus (CSTAR 2003 test set)
- ITG reordering constraints in search improve the translation quality

System	Error Rates		Accuracy Measures	
	mWER [%]	mPER [%]	BLEU [%]	NIST
baseline	48.7	38.6	44.3	9.10
+ ITG constraints	45.1	36.0	47.3	9.32
+ broad beam	49.5	37.3	45.0	9.32
+ IBM-1 lexicon	44.6	35.7	48.9	9.71
+ deletion model	43.2	34.7	50.1	9.80
+ 5-gram LM	42.6	34.2	51.5	9.92

Conclusions

- translation system based on loglinear model combination
- additional knowledge sources easily integrated as features
- phrasal context and local word reorderings are important

  $\Rightarrow$ captured in the alignment templates model
- direct optimization of base models using minimum error training of model scaling factors
- an additional deletion model feature penalizes too short translations
- scaling factors for additional features optimized using n-best lists of translation hypotheses
- optimization of the RWTH system with respect to the NIST score seems to correspond best to subjective evaluation criteria
- on the BTEC Chinese-English and Japanese-English tasks, translations of good quality were produced