Improving the ISL System by using results from commercial systems

Jürgen Reichert, Muntsin Kolss
Universität Karlsruhe

TC-STAR OpenLab, Trento, Mar 30-Apr 2, 2006

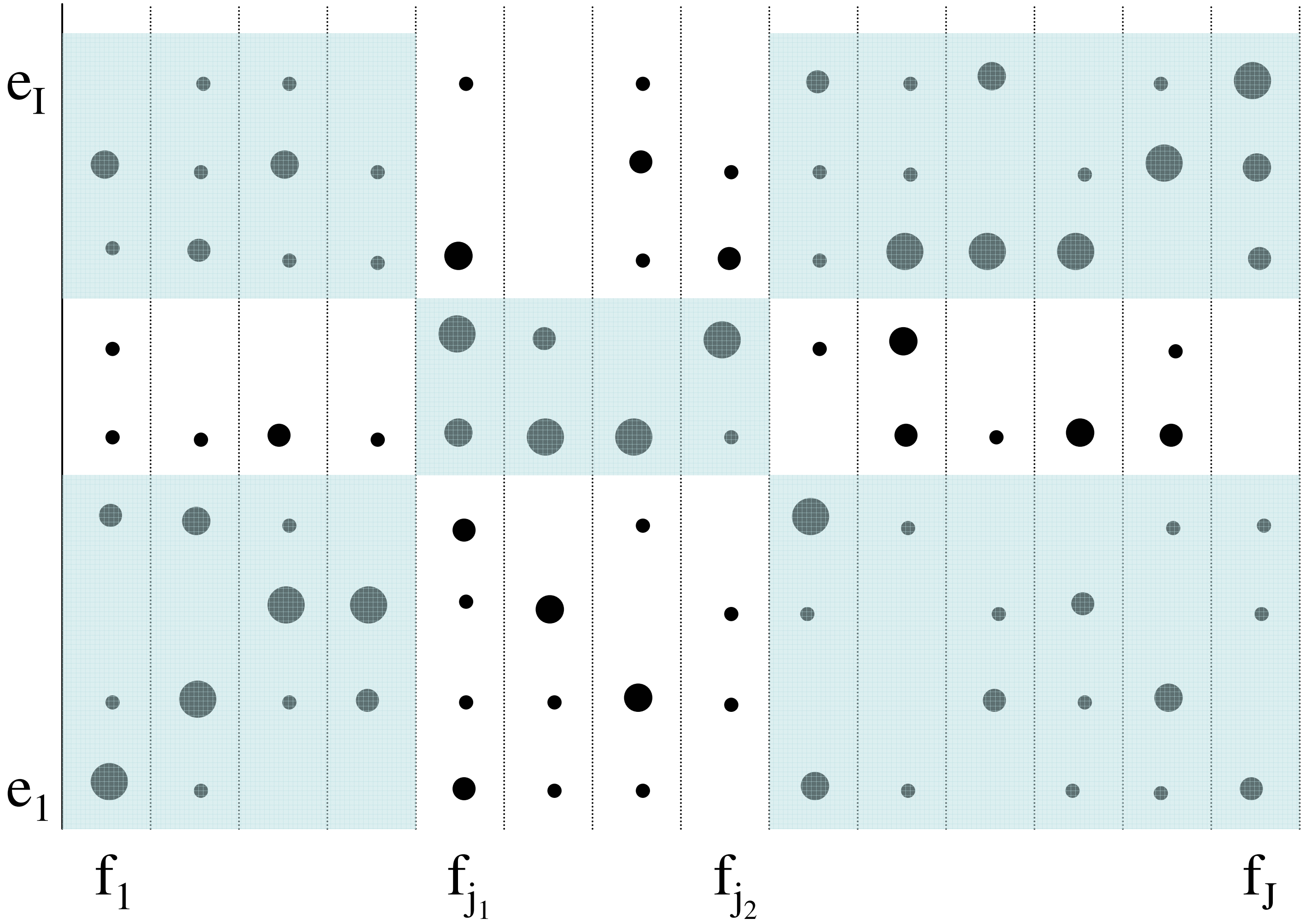
Overview

- The ISL statistical machine translation system
  - STTK developed at CMU/UKA
  - Phrase Translation
  - Decoding
  - OpenLab shared task T1
- System combination with commercial systems

Phrase Translation Approaches

- Train word alignment model and extract phrase-to-phrase translations from Viterbi path
  - IBM model 4 alignment
  - HMM alignment
  - Bilingual Bracketing
- Phrase translation models
  - Integrated segmentation and alignment (ISA)
  - Phrase Pair Extraction via full (constrained) Sentence Alignment (PESA)

Phrase Extraction via Sentence Alignment



Phrase Extraction via Sentence Alignment

- Calculate modified IBM1 word alignment: don't sum over words in 'forbidden' areas

$$\Pr_{(i_1,i_2)}(\vec{t} \mid \vec{s}) = \prod_{j=1}^{j_1-1} \Big(\sum_{i_1 \notin (i_1 \ldots i_2)} \Pr(s_j \mid t_i) \Big) \prod_{j=j_1}^{j_2} \Big(\sum_{i \in (i_1 \ldots i_2)} \Pr(s_j \mid t_i) \Big) \prod_{j=j_2+1}^{J} \Big(\sum_{i_1 \notin (i_1 \ldots i_2)} \Pr(s_j \mid t_i) \Big)$$

- $\Pr(s_j | t_i)$ are normalized over columns, i.e.

$$\sum_{i=1}^{I} \Pr(s_j \mid t_i) = 1$$

- Select target phrase boundaries which maximize sentence alignment probability

$$(i_1, i_2) = \text{argmax}_{(i_1,i_2)} \{ \Pr_{(i_1,i_2)}(s|t) \}$$

ISL Phrase Translation

- Use all translation candidates with scores close to the best one
- Looking from both sides
  - calculate alignment from both sides
  - alignment in reverse direction
  - Interpolation factor tuned on development set
- On-the-fly phrase extraction
  - use suffix array to index source part of corpus
  - Space efficient
  - Search requires binary search
  - Finds n-grams up to any n, within sentence boundaries

Phrase Translation Probabilities

- Most long phrases are seen only once or twice, no good statistics possible
- Want to have phrase translation probabilities close to word translation probabilities
- Use multiple lexical scores as word and phrase translation probabilities:
    - forward and reverse IBM1 at phrase level
    - forward and reverse IBM1 at sentence level
    - relative phrase frequencies
    - can use any statistical lexicon: IBM1-4, HMM, …

Knowledge Sources for Decoding

- Lexical information
  - Statistical lexicon
  - Manual lexicon
  - Phrase translations
  - Named entities
- Language model: standard n-gram
- Position alignment model for word reordering
- Word and phrase count models
- Word fertilities (e.g. from GIZA++)
- Minimum error training (MER) for optimizing model scaling factors

Decoding

- Build translation lattice
  - Run left-to-right over source sentence
  - Search for matching phrases between source sentence and transducer
  - For each translation, insert edges into lattice
  - Lattice input: run over all source lattice edges
- First-best search
  - Run left-to-right over lattice
  - Apply language model
  - Combine translation model score and language model score
  - Recombine and prune hypotheses
  - At sentence end, add sentence length model score
  - Trace back best hypothesis (or n-best hypotheses)

Reordering and Pruning

- Word and phrase reordering within a given window
  - From first un-translated source word next k positions
  - Window length 1: monotone decoding
  - Restrict total number of reordering (typically 3 per 10 words)
- Recombination and pruning of hypotheses
  - Of two hypothesis, keep only better one if no future information can switch their ranking
  - Example: last two word are the same for both hypotheses when a 3-gram LM is used
  - beam search: remove hypotheses which are worse than best hypothesis by a factor k

Evaluation Data and Training

- Training data
  - Spanish/English EPPS: provided T1 corpus, 35? million words
- Preprocessing
  - Some rule-based translation of number and date expressions
  - Some disfluency cleaning (de-stuttering etc.)
  - Tokenization (punctuation marks), lowercasing
  - Splitting of long sentences, limit sentence length
- Postprocessing
  - Remove or keep untranslated words
  - Correct punctuation
  - Mixed Case

Sentence Splitting

- Split long training sentences
  - Improved lexical probabilities
  - Runtime
- Define split points in source and target sentence
  - punctuation marks, brackets
- Choosing split points
  - calculate p_{not_split} = (source sentence | target sentence)
  - calculate p_{split} = p(source left | target left) * p(splitp left | splitp right) *p(source right | target right)
  - in each iteration, re-calculate lexicon and split best N sentence pairs

Combining the ISL system with commercial systems

- ISL system is phrase-based statistical machine translation system
- Commercial systems usually very different from SMT, e.g. grammar/rule based
- Subjective evaluation: comparable translation quality, even though worse when worse NIST/Bleu scores
- Can SMT system profit from this/be improved?

Results, individual systems

T1, Dev-Set	NIST	BLEU	$NIST_{CS}$	$BLEU_{CS}$
UKA/ISL	10.4682	0.5356	10.2179	0.5154
Commercial system 6	9.5855	0.4789	9.5747	0.4818
Commercial system 1	9.4589	0.4587	9.4088	0.4526
Commercial system 7	9.4511	0.4584	9.4008	0.4523
Commercial system 3	9.3926	0.4570	9.3785	0.4521
Commercial system 5	9.3744	0.4551	9.3739	0.4516
Commercial system 4	8.4240	0.4033	8.4080	0.4002
Commercial system 2	8.1513	0.3491	8.1450	0.3450

CS = case sensitive

Results, individual systems

T1, Test-Set	NIST	BLEU	$NIST_{CS}$	$BLEU_{CS}$
UKA/ISL	10.3844	0.5272	10.1403	0.5071
Commercial system 6	9.5608	0.4731	9.5589	0.4701
Commercial system 3	9.4699	0.4570	9.4482	0.4534
Commercial system 5	9.4519	0.4573	9.4335	0.4539
Commercial system 1	9.3338	0.4439	9.2471	0.4342
Commercial system 7	9.3268	0.4437	9.2412	0.4341
Commercial system 4	8.4497	0.4040	8.4150	0.3995
Commercial system 2	8.3189	0.3529	8.2639	0.3468

CS = case sensitive

System selection at the sentence level

- Translate training data by all systems
- Calculate different confidence measures for each utterance
- Calculate NIST/Bleu score for each sentence
- Train classifier (class: best system, parameter vector (confidence measures)
- Translate test sentence by all systems
- Trained classifier selects „best“ hypothesis

Oracle system combination at the sentence level

What is the best we can reach?

Number of systems	NIST optimized		Bleu optimized	
	NIST	BLEU	NIST	BLEU
N=1	NIST=10.8407	BLEU=0.5683	NIST=10.7411	BLEU=0.5694
N=3	NIST=11.0298	BLEU=0.5817	NIST=10.8944	BLEU=0.5859
N=7	NIST=11.1092	BLEU=0.5880	NIST=10.9647	BLEU=0.5931

Oracle system combination at the sentence level

Number of systems	NIST optimized	Bleu optimized
N=1	Systems 0 : 537 counts Systems 1 : 303 counts	Systems 0 : 531 counts Systems 1 : 309 counts
N=3	Systems 0 : 418 counts Systems 1 : 185 counts Systems 2 : 123 counts Systems 3 : 114 counts	Systems 0 : 413 counts Systems 1 : 186 counts Systems 2 : 119 counts Systems 3 : 122 counts
N=7	Systems 0 : 395 counts Systems 1 : 147 counts Systems 2 : 97 counts Systems 3 : 12 counts Systems 4 : 94 counts Systems 5 : 0 counts Systems 6 : 57 counts Systems 7 : 38 counts	Systems 0 : 391 counts Systems 1 : 155 counts Systems 2 : 92 counts Systems 3 : 11 counts Systems 4 : 96 counts Systems 5 : 0 counts Systems 6 : 62 counts Systems 7 : 33 counts

Selection criteria

- OOV estimation
  - Training corpus OOV, Cognate count (lowercase, real words) → not strong enough
- Sentence similarity (n-gram)
  - Generate pool of translated sentences with better scores than SMT system
  - For test sentence, look for best matching sentence in sentence pool
  - If similarity is higher than some threshold, use system which translated the best matching sentence
- Language model score
  - Normalized to sentence length
  - Threshold for each sentence length score
- Sentence length deviation

Results, combined systems

T1, Test-Set	NIST	BLEU
UKA/ISL (baseline)	10.3844	0.5272
All classifiers, 1+7 systems	10.4880	0.5401
Oracle, 1+7 systems	11.1092	0.5880

- NIST improvement 0.10
- Bleu improvement 0.013

Example sentences

593 3,818->8,042

src: Es una iniciativa que merece la pena .

ref This is a worthwhile initiative .

sys0: This is an initiative which deserves the penalty.

sys6: It is an initiative that is worth it.

610 9,049->9,722

src: A este fin hay que desarrollar tecnologías europeas de carbón limpio y captación de dióxido de carbono .

ref: To this end , we have to develop European clean carbon and carbon dioxide sequestering technologies .

sys0: To this end we must develop technologies of the European coal and apprehension clean carbon dioxide.

sys4: To this end one must develop European technologies of clean coal and carbon dioxide collecting.

Example sentences

38 3,833->7,791

src: El pueblo cubano no necesita payasos pasados de moda ni cómplices que le rían las gracias .

ref: The Cuban people do not need out-of-date clowns or accomplices to prop up the regime and pat it on the back .

sys0: The Cuban people not needs buffoons past fashion nor accomplices that you rían thanks.

sys1: The Cuban people do not need not even complicit old-fashioned clowns that laugh it the graces.

817 6,755->15,227

src: Esta es una Comisión mejor .

ref: This is a better Commission .

sys0: This is a Commission that is better.

sys1: This is a better Commission.

Further Work

- Train classifier on more training data
- Better post-processing of system output
- Adapt systems to domain
- More (commercial) systems
- More/different/better selection criteria
- Selection on phrase level