The Statistical Machine Translation System of the University of Edinburgh

Philipp Koehn

pkoehn@inf.ed.ac.uk

School of Informatics

University of Edinburgh

Outline

- **Overview: SMT at Edinburgh**
- Baseline System
- Improvements
- Evaluation
- Related Recent Work in SMT

People Working On SMT at Edinburgh

- Philipp Koehn (lecturer)
- Miles Osborne (lecturer)
- Amittai Axelrod (graduate student)
- Alexandra Birch Mayne (graduate student)
- Chris Callison-Burch (graduate student, Linear-B)
- David Talbot (graduate student)
- Michael White (researcher)

MT Eval 2005 Effort

- 3-month effort building on previous work at MIT
  - improved system performance
  - introduced other researchers to the system
- Focus on Arabic-English:
  - deal with more data
  - various feature improvements

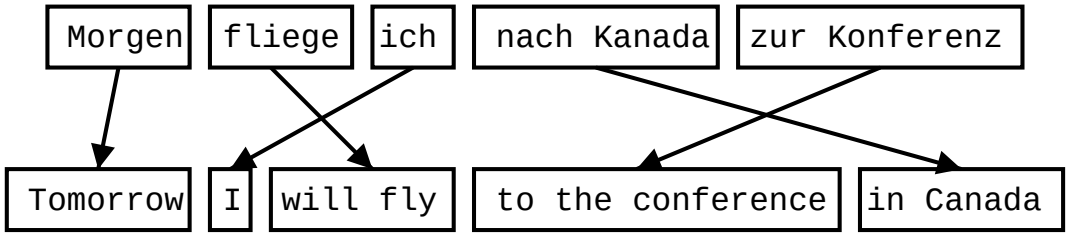
$\Rightarrow$ It is never finished...

  - did not train on new data
  - some changes not completed on time

Outline

- Overview: SMT at Edinburgh
- **Baseline System**
- Improvements
- Evaluation
- Related Recent Work in SMT

Phrase-Based Translation



- Phrase model similar to other groups' model
  - word align corpus, using GIZA++ and Och's refined method
  - collect phrase pairs consistent with word alignment
  - log-linear model to combine model components
  - parameter tuning by minimum error rate training
  - decoder Pharaoh (http://www.isi.edu/licensed-sw/pharaoh/)

System Components

- reordering model linear reordering cost, max. 4 word movement
- language model trigram LM trained using SRILM toolkit
- phrase translation model f→e
- phrase translation model e→f
- word translation model f→e
- word translation model e→f
- word penalty
- phrase penalty

Outline

- Overview: SMT at Edinburgh
- Baseline System
- **Improvements**
  - more training data (+2% BLEU)
  - bigger language model (+2% BLEU)
  - minor model improvements (+2% BLEU)
- Evaluation
- Related Recent Work in SMT

More Training Data

- All of the data (instead of half)
  - maximum sentence length 40 words
  - break up corpus in 2-3 parts
  - run snt2cooc separately, merge
  - combined GIZA++ run (3-5 days CPU time)
- Chunking
- Splitting

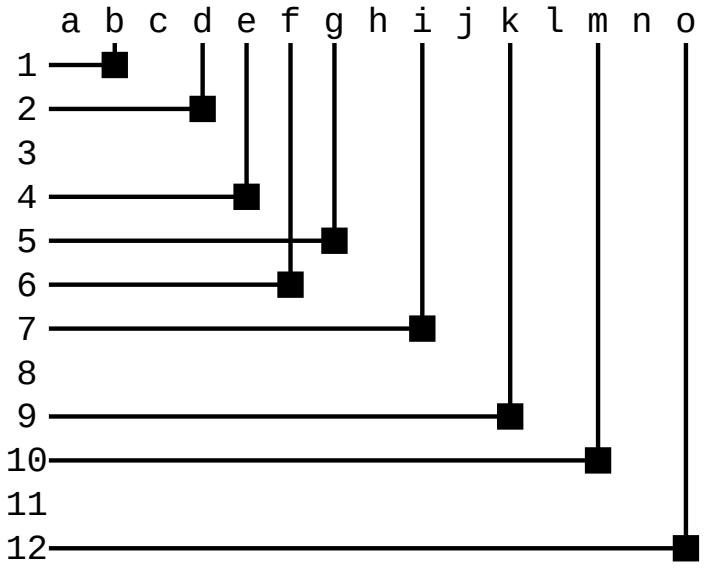
Chunking

- Break up along comma, semicolon, colon, etc.
- Sentence-align smaller units
- 63.9 $\rightarrow$ 100.3 million words used

Splitting

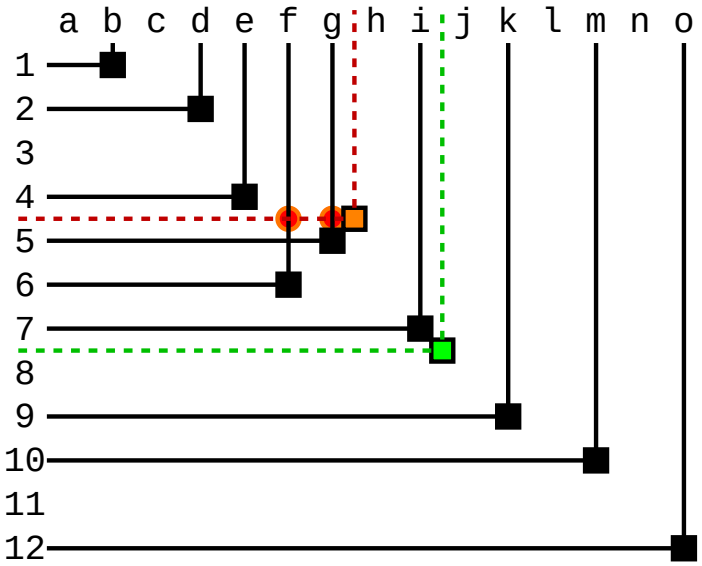
- Break up longer sentences
  - minimum number of crossed word alignments
  - cut sentences in the middle third
  - cut as central as possible
- 100.3 $\rightarrow$ 130.3 million words used

Splitting II



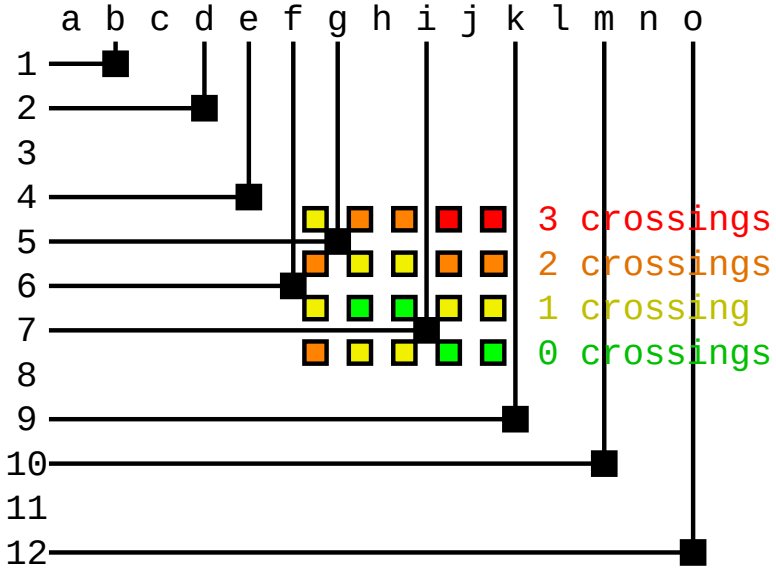
- Aligned sentences using lexical t-table with $p > 0.03$ threshold, eliminate multiple aligned words

Splitting III



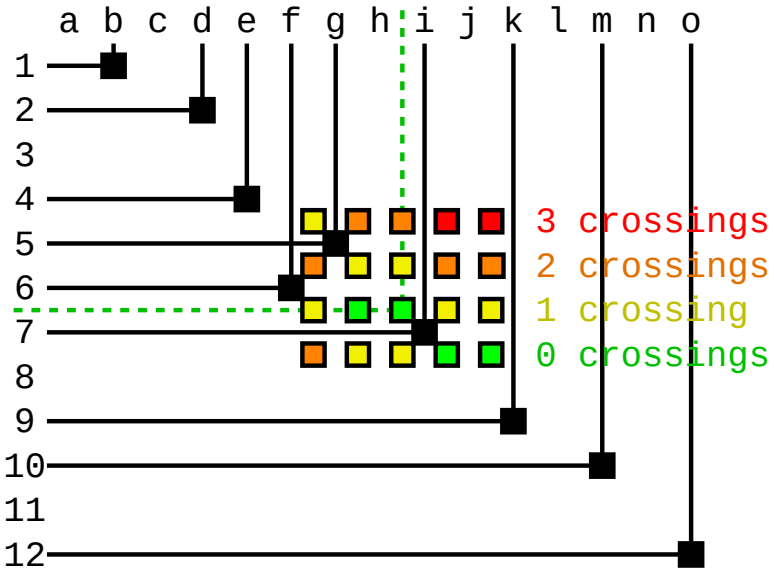
- Good and bad (2 crossings) split points

Splitting IV



- Quality of split points in the middle third

Splitting V



- Find most central best split point

Bigger Language Model

- Dealing with memory limitations in training
- Dealing with memory limitations in decoding
- Multiple language models

Memory Limitations in Training

- A lot of monolingual English text is available
  - English half of parallel text: 130 million words
  - English gigaword corpus: 1.78 billion words
  - the web: 1 trillion words ?
- SRILM training keeps all n-grams in memory (2-4 GB limit)
- Practically limited to:
  - 800 million words (training + part of Gigaword)
  - ignored trigram singletons
  - digits ('0'-'9') replaced by '5'

Memory Limitations in Decoding

- Pruning possible?
  - only need to consider words that can be produced
  - translation model can be cut down to a few (1-2) percent

	Unigrams	Bigrams	Trigrams
Entire LM (trained on 130m)	291,767	4,991,346	7,881,122
1000 sent.	13,792	2,850,983	6,540,940
1000 sent, top 20 transl.	9,860	2,251,111	5,590,783
10 sent, top 20 transl.	871	127,552	488,694

$\Rightarrow$ High overhead in filtering LM

Multiple Language Models

- Pharaoh allows multiple language models:
- Large LM
  - trained on 800 million words (training + part of Gigaword)
  - ignored trigram singletons
  - digits ('0'-'9') replaced by '5'
- Specialized LM
  - trained on 1.1 million words (news training corpus)
  - including all singletons
  - no special treatment of numbers
- Weights of LM determined by discriminative training

Minor Model Improvements

- dropping unknown words during decoding
- delete word feature
- limited changes to the recapitalizer
- limited post-editing of the output
- limited changes to the tokenization of Arabic

Outline

- Overview: SMT at Edinburgh
- Baseline System
- Improvements
- **Evaluation**
- Related Recent Work in SMT

Evaluation for Arabic-English

- Improvements for Arabic-English:

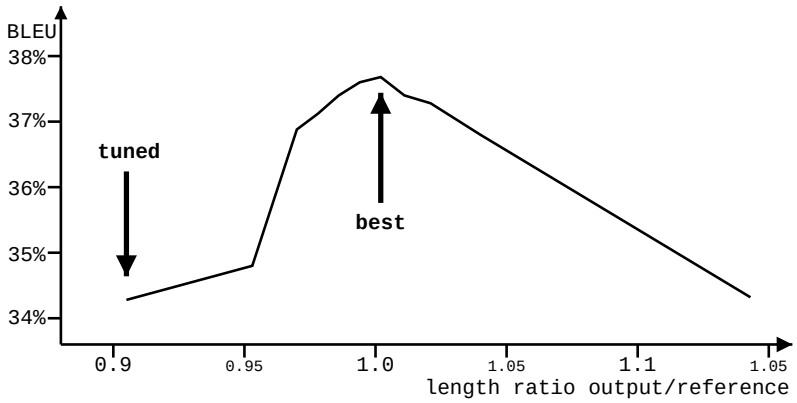
Eval set	'04 system	'05 system
Eval 2002 (partial)	34.4% BLEU	40.4% BLEU
Eval 2004	34.1% BLEU	34.3% BLEU
Eval 2005	35.6% BLEU	40.5% BLEU

Why so Little Improvement on Eval 2004?

- Model optimized on first 300 sentences of Eval 2002

$\Rightarrow$ very short output (length ratio 0.905)

- Word penalty feature allows tuning of output length:



- Manual adjustment: 34.3% $\Rightarrow$ 37.7% BLEU

Evaluation for Chinese-English

- Improvements for Chinese-English
- System changes:
  - bigger language model (800 million words)
  - debugged number translator

Eval set	'04 system	'05 system
Eval 2002 (partial)	26.1% BLEU	27.2% BLEU
Eval 2004	27.1% BLEU	28.1% BLEU
Eval 2005	24.4% BLEU	25.1% BLEU

Outline

- Overview: SMT at Edinburgh
- Baseline System
- Improvements
- Evaluation
- **Related Recent Work in SMT**
  - clause restructuring [Collins,Koehn,Kucerova, 2005]
  - Euromatrix [Koehn, 2005]
  - shared task at ACL workshop [Koehn and Monz, 2005]

Clause Level Restructuring

- Why clause structure?
  - languages differ vastly in their clause structure
    (English: SVO, Arabic: VSO, German: fairly free order;
    a lot details differ: position of adverbs, sub clauses, etc.)
  - large-scale restructuring is a problem for phrase models
- Restructuring
  - reordering of constituents (main focus)
  - add/drop/change of function words
- Ongoing work
  - collaboration with Michael Collins and Ivona Kucerova
  - currently German-English
  - see ACL paper for details

Clause Structure

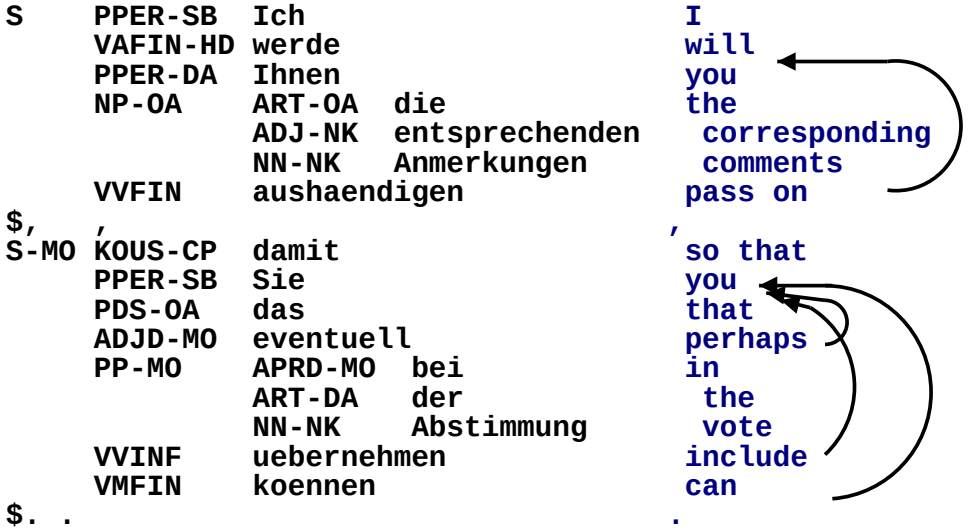
```
S   PPER-SB  Ich    I
    VAFIN-HD werde    will
    VP-OC    PPER-DA Ihnen   you
             NP-OA   ART-OA  die   the
                     ADJ-NK  entsprechenden    corresponding
                     NN-NK   Anmerkungen   comments
             VVFIN   aushaendigen     pass on
             $,      ,       ,
             S-MO    KOUS-CP damit   so that
                     PPER-SB Sie    you
                     VP-OC   PDS-OA  das   that
                             ADJD-MO eventuell    perhaps
                             PP-MO   APRD-MO  bei   in
                                     ART-DA   der   the
                                     NN-NK    Abstimmung  vote
                             VVINF   uebernehmen    include
                     VMFIN   koennen   can
$. .            .
```

MAIN CLAUSE

SUB-ORDINATE CLAUSE

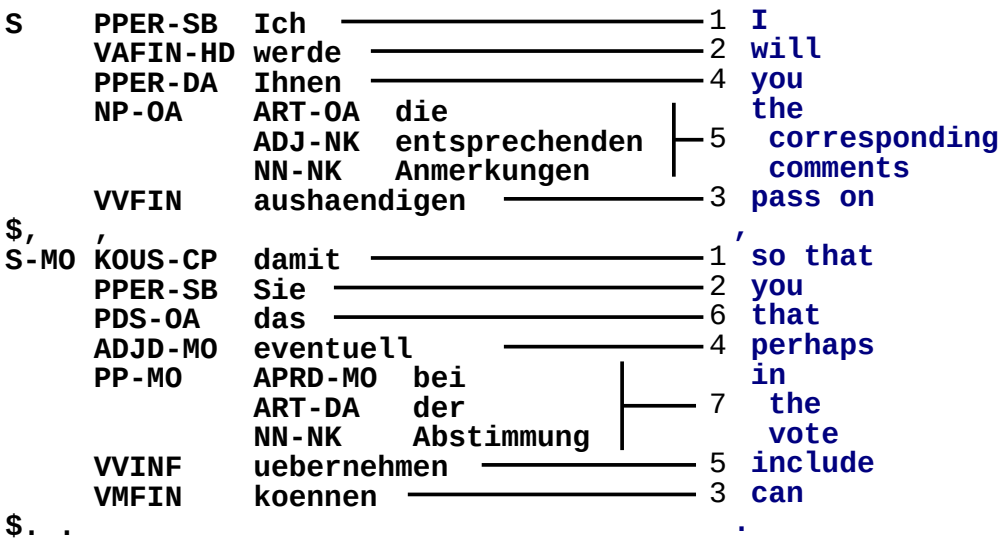
- Syntax tree from German parser
  - statistical parser by Amit Dubey, trained on TIGER treebank

Reordering When Translating



- Reordering when translating into English
  - tree is flattened
  - clause level constituents line up

Clause Level Reordering



- Clause level reordering is a well defined task
  - label German constituents with their English order
  - done this for 300 sentences, two annotators, high agreement

Systematic Reordering German → English

- Many types of reorderings are systematic
  - move verb group together
  - subject - verb - object
  - move negation in front of verb

⇒ Write rules by hand

  - apply rules to test and training data
  - train standard phrase-based SMT system

System	BLEU
baseline system	25.2%
with manual rules	26.8%

Euromatrix

- Proceedings of the European Parliament
  - translated into 11 official languages
  - entry of new members in May 2004: more to come...
- Europarl corpus
  - collected 20-30 million words per language

  → 110 language pairs
- 110 Translation systems
  - 3 weeks on 16-node cluster computer

  → 110 translation systems

Quality of Translation Systems

- Scores for all 110 systems

	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	31.2	32.1	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	30.1	31.1	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	30.5	-	40.2	12.5	32.3	21.4	35.9	23.9
fr	23.7	18.5	26.1	30.0	38.4	-	12.6	32.4	21.1	35.3	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	34.0	36.0	11.0	-	20.0	31.2	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	30.1	37.9	39.0	11.9	32.0	20.2	-	21.9
sv	30.3	18.9	22.8	30.2	28.6	29.7	15.3	23.9	21.9	25.9	-

Translate into vs. out of a Language

- Some languages are easier to translate into that out of

Language	From	Into	Diff
da	23.4	23.3	0.0
de	**22.2**	**17.7**	**-4.5**
el	23.8	22.9	-0.9
en	**23.8**	**27.4**	**+3.6**
es	26.7	29.6	+2.9
fr	26.1	31.1	+5.1
fi	19.1	12.4	-6.7
it	24.3	25.4	+1.1
nl	19.7	20.7	+1.1
pt	26.1	27.0	+0.9
sv	24.8	22.1	-2.6

Backtranslations

- Checking translation quality by back-transliteration
- *"The spirit is willing, but the flesh is weak"*
- English $\rightarrow$ Russian $\rightarrow$ English
- *"The vodka is good but the meat is rotten"*

Backtranslations II

- Does not correlate well with unidirectional performance

Language	From	Into	Back
da	28.5	25.2	56.6
de	25.3	17.6	48.8
el	27.2	23.2	**56.5**
es	30.5	30.1	52.6
fi	21.8	13.0	44.4
it	27.8	25.3	49.9
nl	23.0	21.0	46.0
pt	30.1	27.1	**53.6**
sv	30.2	24.8	54.4

Shared Task at ACL 2005 Workshop

- Given
  - parallel text, word alignment
  - language model
  - decoder Pharaoh
- Task:
  - build SMT system (at least: probabilistic phrase table)
  - French-English, Spanish-English, Finnish-English, German-English
- Participation
  - 11 teams from 8 institutions
  - several new research groups

Thank You!

- Questions?