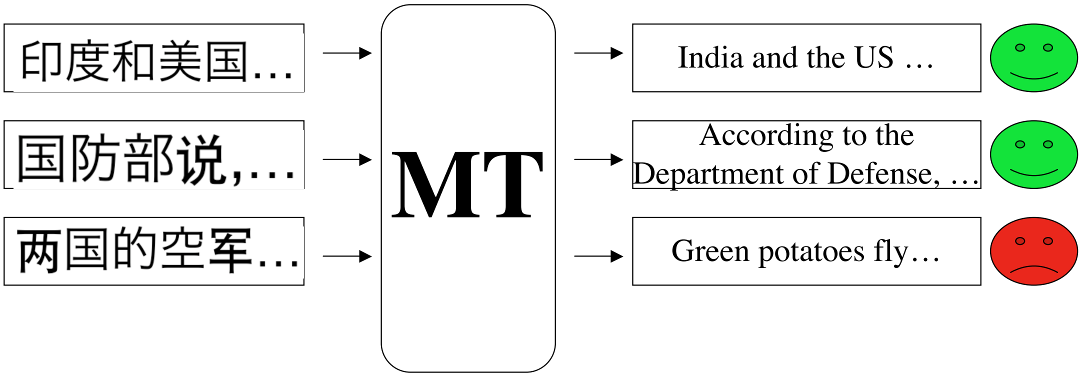
Franz Josef Och
Google, Inc.
September 12, 2005

Statistical Machine Translation: Foundations and Recent Advances

Tutorial at MT Summit 2005
Phuket, Thailand

Franz Josef Och
Google, Inc.
September 12, 2005

Goal: High quality translation of NL text



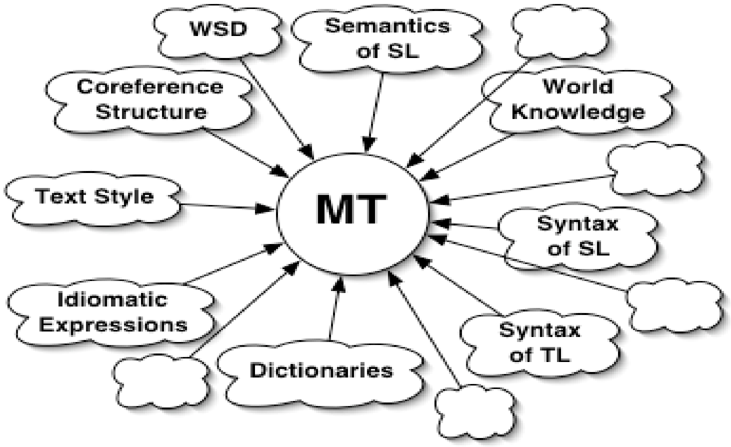
印度和美国… → MT → India and the US … ☺

国防部说,… → MT → According to the Department of Defense, … ☺

两国的空军… → MT → Green potatoes fly… ☹

Goal: Many ☺ Few ☹

Challenges in MT: Complexity + Size



- 100 languages with more than 7 million speakers
- 10.000 MT systems?

What is Statistical Machine Translation? (“The Aachen interpretation”)

Machine translation is decision problem

- Given source language sentence: decide for the best translation

In machine translation we have to make decisions under uncertainty

- Fundamental uncertainties (e.g. ambiguity)
- Simplified model assumptions (e.g. just trigram dependencies)

Goal is to make good, ideally optimal decisions

- General theory: statistics, statistical decision theory

For machine translation: statistical MT

= trying to make good, ideally optimal decisions in MT

Hermann Ney: “statistical approach is expressed by equation

MT = Linguistic Modeling + Statistical Decision Theory”

Overview

Foundations of Statistical Machine Translation

Automatic Evaluation of Machine Translation

Data

Statistical alignment models

Phrase-based models

Search

Discriminative Training

System Building

Results

Outlook

SMT vs. EBMT

Statistical Machine Translation vs. **Example-based Machine Translation (EBMT)**

- Both are empirical approaches
  - As opposed to rule-based machine translation
- EBMT emphasizes ‘learning from examples’
  - Often heuristic scoring/learning methods
- SMT emphasizes ‘making optimal decisions’, statistical methods, …
- SMT and EBMT astonishingly separate research communities
  - SMT researchers often use methods and terminology from speech recognition research
  - Different ‘language’ used in both communities

True probability distribution vs. model

e: target language sentence (English sentence)

f: source language sentence (Foreign/French sentence)

Pr(**e**|**f**): **true probability distribution** that English (target language) sentence **e** is translation of Foreign (source language) sentence **f**

Problem: true probability distribution for translation is not known

p(**e**|**f**): **model** approximating Pr(**e**|**f**)

Assumption: we have the true probability distribution Pr(**e**|**f**)

Question: What is the ‘best translation’ **e** for a specific Foreign **f** sentence?

- What means ‘best translation’?
- Some errors are more ‘costly’ than others

L(**e**,**e’**,**f**): costs (loss) to produce **e’** given that **e** is the correct translation of Foreign sentence **f**?

Reformulated question: What is the translation where our expected loss is minimal?

Minimum loss decision rule; MAP decision rule

(General) Minimum loss decision rule:

$$\hat{\mathbf{e}}(\mathbf{f}) = \mathrm{argmin}_{\mathbf{e}} \sum_{\mathbf{e}'} L(\mathbf{e}, \mathbf{e}', \mathbf{f}) \cdot Pr(\mathbf{e}|\mathbf{f})$$

Assume: L(**e**,**e’**,**f**)=0 for identical **e** and **e**’ and L(**e**,**e’**,**f**)=1 otherwise

- So-called 0-1 loss function
- MAP (maximum a posteriori) decision rule

$$\hat{\mathbf{e}}(\mathbf{f}) = \mathrm{argmax}_{\mathbf{e}} Pr(\mathbf{e}|\mathbf{f})$$

Three tasks: Modeling, Training, Search

Modeling:

- Coming up with a model p(e|f) for Pr(e|f)
- Describing the relevant (linguistic) dependencies

Training (or Parameter Estimation):

- Assigning specific values to model parameters given data
- E.g. EM-algorithm, relative frequency estimate, maximum likelihood

Search (or Decoding):

- Actual translation process
- Find 'best' translation e for input sentence f
- Very large search space: typically heuristic search algorithms

Classification of translation errors in final system

Bayes Error:

- Inherent property of the translation problem; can never be eliminated

Model Error:

- Error due to having an incorrect model (e.g. restricting to bigram dependencies)

Training Error:

- Error due to finite training sample (or 'bad' training criterion)

Search Error:

- Error due to finding a sub-optimal translation due to heuristic search

(Bugs)

Modeling techniques

Naïve approach:

- Introduce model parameter for each (**f**,**e**)-pair
- Not enough data for this!
- Need to introduce further structure

(some) useful modeling techniques:

- Source-channel model (introduces separate target language model)
- Log-linear model
- Generative model
- Markovization

Source-channel model

$$\begin{aligned}
\hat{\mathbf{e}}(\mathbf{f}) &= \operatorname{argmax}_{\mathbf{e}} Pr(\mathbf{e}|\mathbf{f}) \\
&= \operatorname{argmax}_{\mathbf{e}} \frac{Pr(\mathbf{e}) \cdot Pr(\mathbf{f}|\mathbf{e})}{Pr(\mathbf{f})} \\
&= \operatorname{argmax}_{\mathbf{e}} Pr(\mathbf{e}) \cdot Pr(\mathbf{f}|\mathbf{e}) \\
&\approx \operatorname{argmax}_{\mathbf{e}} p(\mathbf{e}) \cdot p(\mathbf{f}|\mathbf{e})
\end{aligned}$$

Framework for early statistical machine translation approaches
([Brown, Cocke et al, 1990])

Source-channel model

p(**e**): language model (LM)

- Judges whether **e** is well-formed English string
- Standard modeling approach: n-gram model

p(**f**|**e**): translation model (TM)

- Judges whether **f** is reasonable translation of **e**
- Standard modeling approach: statistical alignment models (see later)

Problems:

- ‘simple product’ is not necessarily optimal combination of TM + LM
- No straightforward way to add additional knowledge sources
- Assumes 0-1 loss function

Log-linear model

$$p(e|f) = \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(e,f))}{\sum_{e'} \exp(\sum_{m=1}^{M} \lambda_m h_m(e',f))}$$

$h_m(e,f)$: feature function

λ_m: feature function weight

$$\hat{e}(f, \lambda_1^M) = \operatorname{argmax}_e \sum_{m=1}^{M} \lambda_m h_m(e,f)$$

Motivation:

- Feature function score different aspects of translation quality
- Feature function weight: trustworthiness of feature function
- Generalization of source-channel model
- Straightforward to add new knowledge sources

Example feature functions

Scoring 'fluency'; no dependence on f (**language model features**)

- Language model probability (log(p(e))
- Number of words
- Number of phrases

Scoring translation quality (**translation model features)**

- Phrase translation probability: log(p(e|f))
- Phrase translation probability: log(p(f|e))
- Other phrase quality features

Reordering features:

- Degree of non-monotonicity
- Phrase-specific reordering features

Generative models

Introduce intermediate steps into translation process

- E.g. Word alignment **a** : specifies word order
- In general: hidden variables

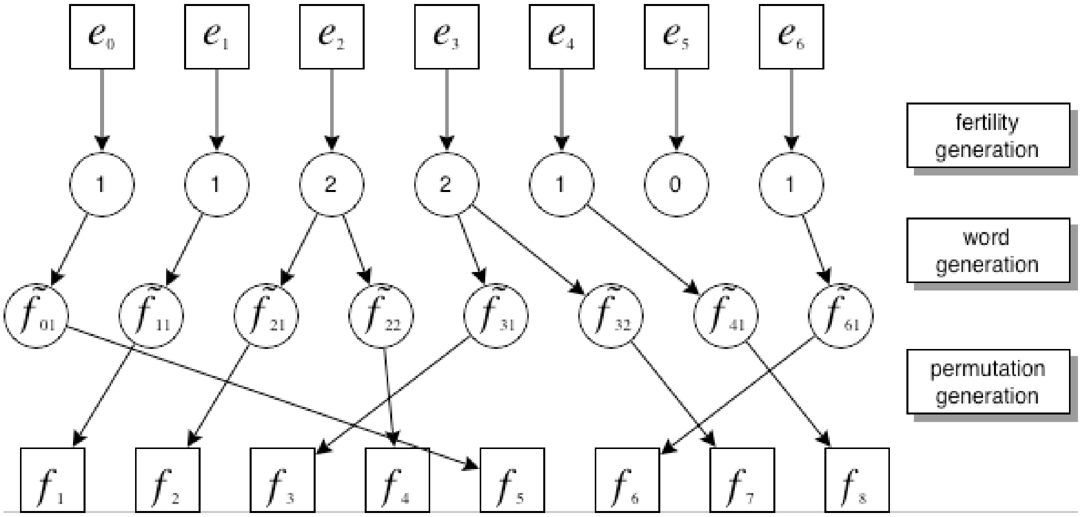
$$Pr(\mathbf{e}|\mathbf{f}) = \sum_{a} Pr(\mathbf{e}, \mathbf{a}|\mathbf{f})$$

Problem: sum over h complicates model

- Maximum approximation: replace sum by maximum

More details later…

Standard Generative Translation Model: Model 3 [Brown, Pietra, Pietra & Mercer, 1993]



Hidden Variables for log-linear models

Refinement for log-linear models: feature functions depend also on hidden variable

$$p(e,a|f) = \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(e,a,f))}{\sum_{e',a'} \exp(\sum_{m=1}^{M} \lambda_m h_m(e',a',f))}$$

$h_m(e,a,f)$: feature function

λ_m: feature function weight

decision rule:

$$\hat{e}(f, \lambda_1^M) = \operatorname{argmax}_e \max_a \sum_{m=1}^{M} \lambda_m h_m(e,a,f)$$

Literature

- [Brown, Cocke et al, 1990]: A statistical approach to machine translation. Computational Linguistics, 16:79–85, 1990.
- [Brown, Pietra, Pietra & Mercer, 1993]: The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, 1993.
- [Och & Ney, 2002]: Discriminative Training and Maximum Entropy Models for Statistical machine Translation, ACL02.
- [Kumar & Byrne, 2004]: Minimum Bayes-Risk Decoding for Statistical Machine Translation, HLT-NAACL04.
- [Duda, Hart & Stork, 2001]: Pattern Classification, John Wiley & Sons, 2001.
- [Ney, 2001]: Stochastic Modeling: From Pattern Classification to Language Translation, Workshop on Data-Driven Machine Translation, 2001.

Overview

Foundations of Statistical Machine Translation

Automatic Evaluation of Machine Translation

Data

Statistical alignment models

Phrase-based models

Search

Discriminative Training

System Building

Results

Outlook

Evaluation of MT? 🙂 Vs. ☹

Ideal criterion: user satisfaction

Problems:

- Expensive, Slow, Inconsistent, Subjective
- Problematic to use in system development

Goal: automatic + objective evaluation of machine translation quality

- Idea: Compute similarity of MT output with good human translations (=reference translations)
- Hope:
  - If MT output is good: similar to good human translations
  - If MT output is bad: very different from human translations

Question: Which similarity metric?

Which output is better?

Human reference translations:

- Reference 1: japan remained the biggest trading partner
- Reference 2: japan is still the largest trade partner
- Reference 3: japan still remains the number one trade partner

MT output:

- MT 1: japan will continue to be partner big
- MT 2: japan is still the biggest trading partner
- MT 3: ukraine won 2:1 against poland

Which output is better?

Human reference translations:

- Reference 1: japan remained the biggest trading partner
- Reference 2: japan is still the largest trade partner
- Reference 3: japan still remains the number one trade partner

MT output:

- MT 1: **japan** will continue to be **partner** big
- MT 2: **japan is still the biggest trading partner**
- MT 3: ukraine won 2:1 against poland

N-gram precision / BLEU

Modified n-gram precision: p_n

- N-gram precision: fraction of N-grams occurring in references
- Modified N-gram precision: same part of reference cannot be 'used' twice

Brevity penalty: BP

- Penalize too short translations
- BP = exp(min(1 - r/c , 0))
- c: length of MT output, r: length of reference translation

BLEUn4 score:

- BLEU = BP exp (0.25 (log(p_1)+ log(p_2)+ log(p_3)+ log(p_4)))

BLEU

- Range of BLEU scores 0.0 to 1.0 (0% to 100%)
- Many studies show high correlation (80-95%) of BLEU scores with subjective human judgments of fluency and adequacy **for existing MT systems**
- But: large test corpora needed
  - Low correlation of sentence-level BLEU scores with sentence level human judgments
- Very problematic to compare BLEU scores for
  - different language pairs
  - different number of references
  - different n-gram sizes
  - In general: different test corpora

BLEU

Important: BLEU scores have limited precison

- Rule of thumb: 20K word corpus: ~1% needed for being statistical significant
  - **Frequent mistake**: believing in very small BLEU score differences
- Scientific approach to evaluation and reporting results
  - Report confidence intervals of results (e.g. using bootstrap method)
  - Do not report results with too many digits of precision (deceiving)
  - Ideally: Report results on standard test corpora (e.g. NIST data sets)
  - Compare to results of state-of-the-art systems or state-of-the-art baseline
  - Use standard evaluation settings (e.g. BLEUr4n4 with case)

Typical BLEU scores (2005 NIST evaluation data)

Arabic-English news translation, 4 references:

- Best statistical (research) system: 51% BLEU score
- (some) commercial systems: 10 - 34% BLEU score
- Estimated human BLEU score: 63% BLEU score (see next slide)

Chinese-English news translation, 4 references:

- Best statistical (research) system: 35% BLEU score
- (some) commercial system: 15% BLEU score
- Estimated human BLEU score: 55% BLEU score (see next slide)

A note on human BLEU scores

Approach used to estimate human BLEU score (given 4 references):

- Round robin score one reference against other 3 references
- (Heuristically) Rescale scores from 3 references to 4 references

Problems:

- Approach depends on ‘agreement’ between human translators
  - Stricter translation guidelines yield more ‘agreement’
- On some test corpora: MT output is better than ‘estimated human BLEU scores’ -- but subjectively still worse
- Conclusion: BLEU cannot rank human quality translations
  - If MT gets better: we might need new evaluation metric

Other metrics: mWER, mPER

NIST score:

- Like BLEU score but using information-weighted n-grams

Multi-reference word error rate (mWER):

- Number of edit operations to transform MT output into one of the references
- Edit operations: Insertions, Deletions, Substitutions
- Strongly penalizes differences in word order

Multi-reference position-independent word error rate (mPER):

- Similar as mWER, but ignoring word order

Literature:

Uses of automatic evaluation metrics before BLEU:

- [Su, Wu & Chang, 1992]: A New Quantitative Quality Measure for Machine Translation System; COLING-92.
- [Tillmann, Vogel, Ney & Zubiaga, 1996]: A DP based search using monotone alignments in statistical translation, ACL97.
- [Alshawi, Bangalore & Douglas, 1998]: Automatic Acquisition of Hierarchical Transduction Models for Machine Translation, COLING-ACL98.

BLEU

- [Papineni, Roukos, Ward & Zhu, 2002]: K. Papineni, S. Roukos, T. Ward, W. Zhu: BLEU: a Method for Automatic Evaluation of Machine Translation; ACL02.

Other automatic evaluation metrics:

- [Melamed, Green & Turian, 2003]: Precision and Recall of Machine Translation, HLT-NAACL03.
- [Lin & Och, 2004]: Orange: a Method for Evaluation Automatic Evaluation Metrics for Machine Translation; COLING04.

Textbook on the ‘bootstrap’ technique (for computing statistical significance):

- [Efron, Tibshirani & Robert, 1993]: An Introduction to the Bootstrap, Chapman & Hall, London.

Overview

Foundations of Statistical Machine Translation

Automatic Evaluation of Machine Translation

Data

Statistical alignment models

Phrase-based models

Search

Discriminative Training

System Building

Results

Outlook

Parallel Data Processing

LDC (Linguistic Data consortium) -- provides easy-to-use cleaned data:

- >150M words parallel AE corpora
- >200M words parallel CE corpora

Major steps to prepare data for Training:

- Find parallel data
- Preprocessing / tokenization / normalization
- Document alignment
- Sentence / chunk alignment

Finding parallel data: sources of data

Multilingual content on the web

- European Union web site (20-lingual corpora)
- United Nations web site (6-lingual corpora)
- Tourism web pages
- Canadian Government (French-English), …

Manual collection of data:

- E.g. in Verbmobil project for speech-to-speech translation in specific domain (appointment scheduling, hotel reservation, dialogue)

Translation memories

LDC

…

Preprocessing / tokenization / normalization

Example of preprocessing steps:

- Input:
  `He said: “The car can’t be red, green or blue.”`
- Tokenization: detect token boundaries (introduce spaces)
  `He said : “ The car can’t be red , green or blue . ”`
- True-casing (alternative: lowercase everything)
  `he said : “ the car can’t be red , green or blue . ”`
- Normalization:
  `he said : “ the car can not be red , green or blue . ”`

Preprocessing is language specific:

- Chinese: segmentation of Character stream into words
- Arabic: normalization of vowels, ligatures, …
- German: compound-splitting

Document alignment

Sometimes: file name matching

- European Union web-site:
  - Replace “de” by “en” to find English translation of German document
- Unfortunately, does not work very often

General approach:

- Find large set of candidate translations
- Compare every document with every other document
  - Find pairs that maximize dictionary overlap
- Computationally expensive

Sentence/chunk alignment

Goal: find corresponding sentences/chunks in aligned documents

- Score sentence alignments using
  - Dictionary overlap
  - Sentence length mismatch

- Assumption:
  - monotone translation of sentences
  - Alignment possibilities: 1-1, 2-1, 1-2, 2-2

- Solution: Dynamic programming search for optimal alignment

Monolingual data

Sources of monolingual data:

LDC Gigaword corpora: Chinese, Arabic, English (~1 billion words)

News corpora

The Web (>> 200 billion words)

Standard use of monolingual data:

Train trigram language model: $p(w_n|w_{n-2},w_{n-1})$

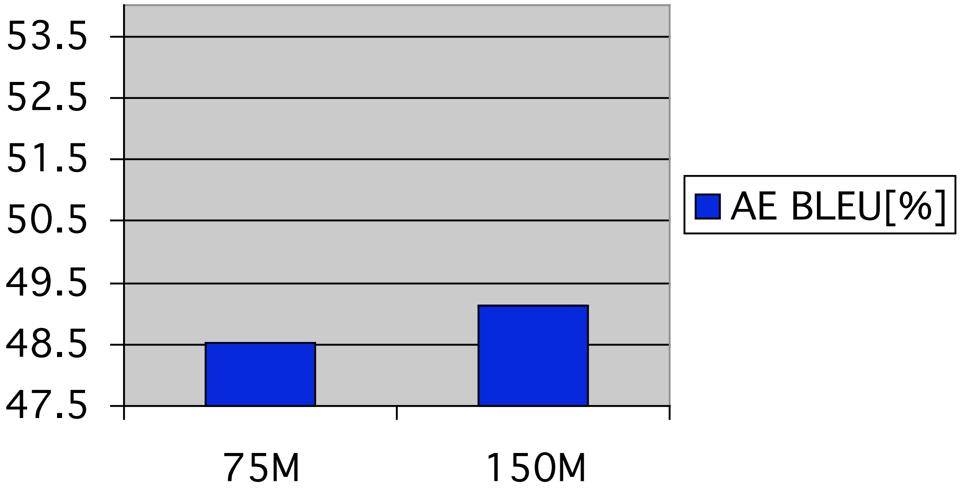
Smoothing methods: linear interpolation, Kneser-Ney, …

How much data is needed?

Answer: MORE

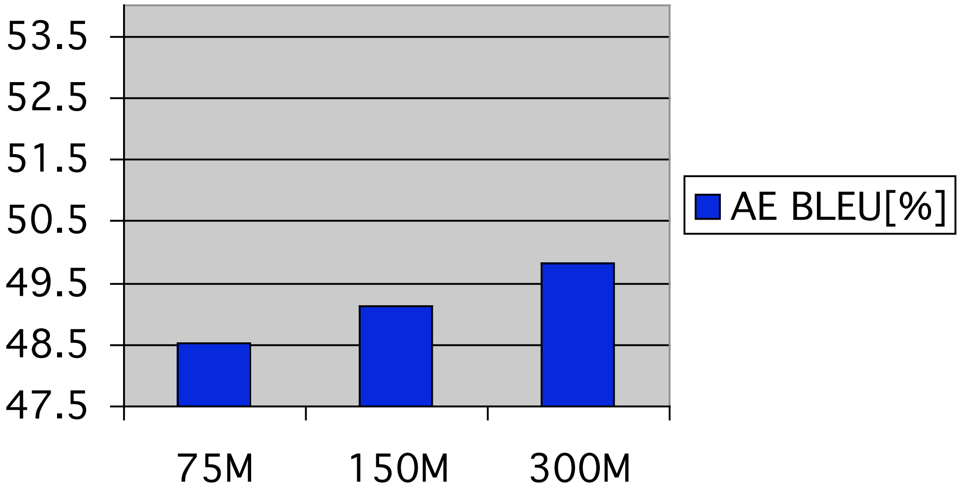
More data is better data…

Impact on size of language model training data (in words) on quality of Arabic-English statistical machine translation system (NIST test data)



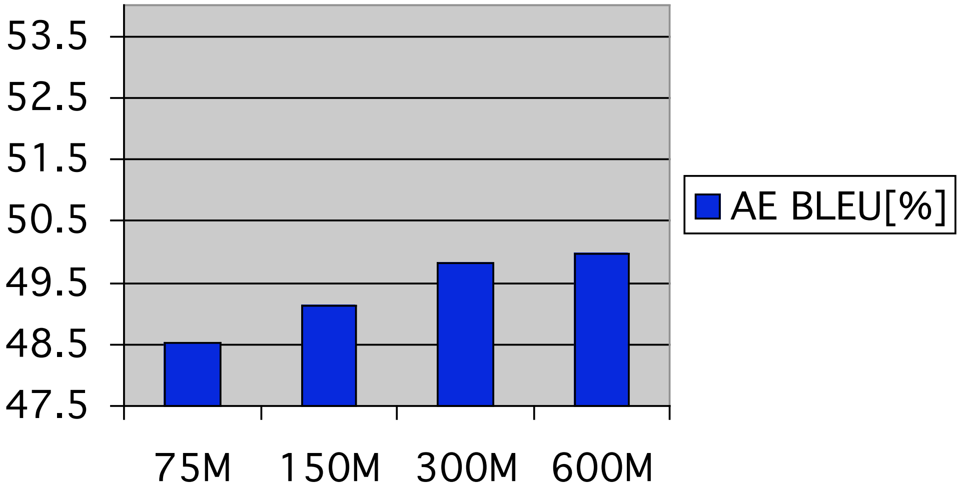
More data is better data…

Impact on size of language model training data (in words) on quality of Arabic-English statistical machine translation system



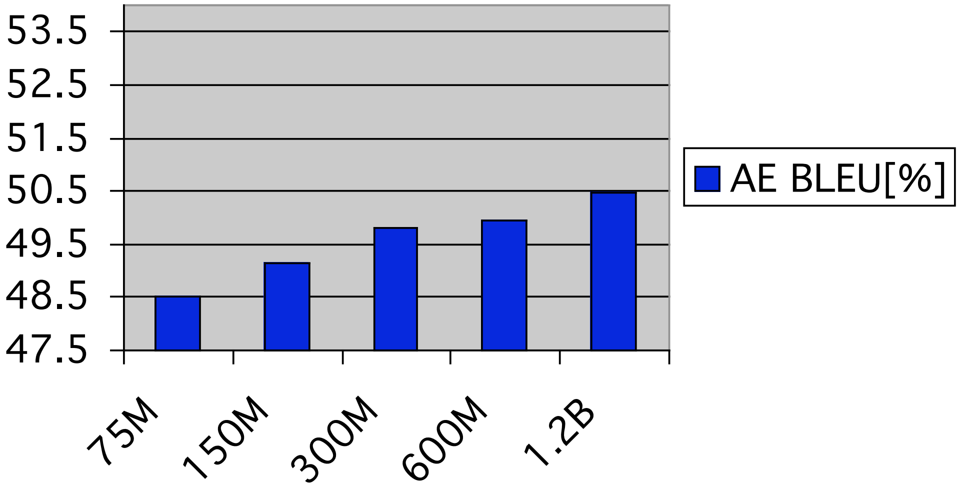
More data is better data…

Impact on size of language model training data (in words) on quality of Arabic-English statistical machine translation system



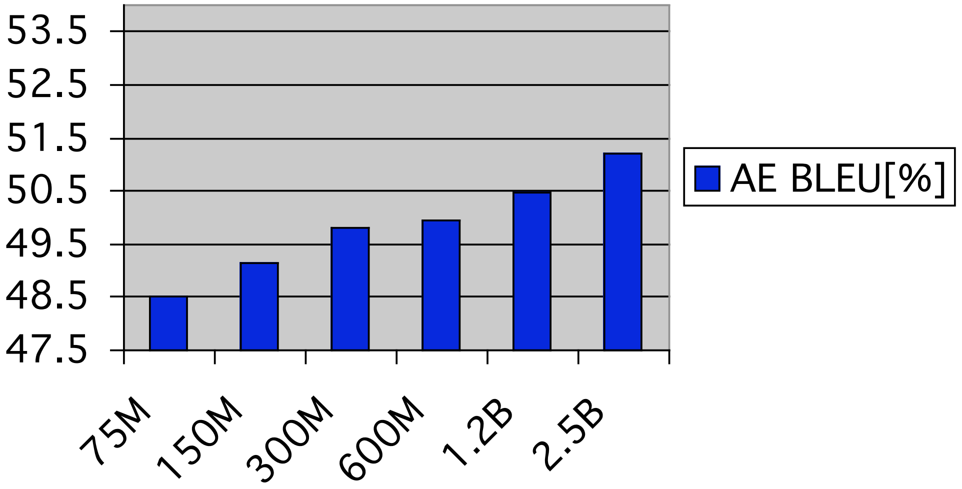
More data is better data…

Impact on size of language model training data (in words) on quality of Arabic-English statistical machine translation system



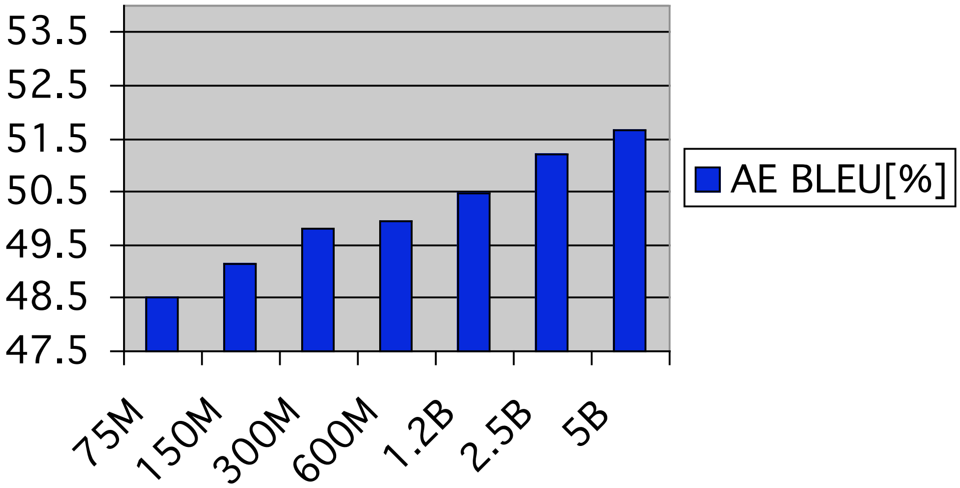
More data is better data…

Impact on size of language model training data (in words) on quality of Arabic-English statistical machine translation system



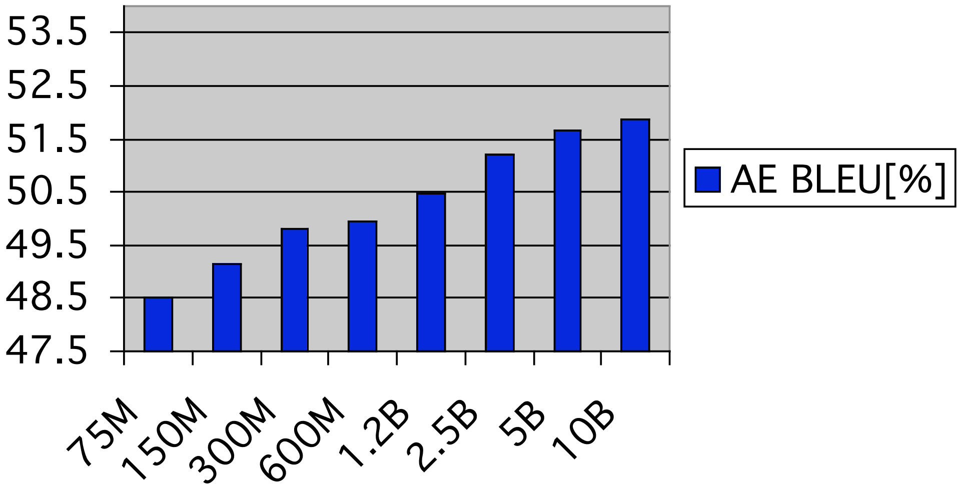
More data is better data…

Impact on size of language model training data (in words) on quality of Arabic-English statistical machine translation system



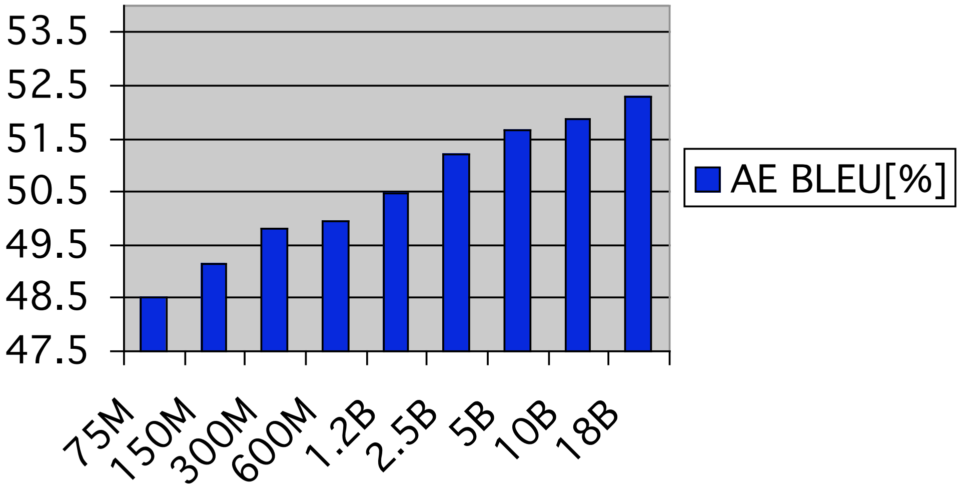
More data is better data…

Impact on size of language model training data (in words) on quality of Arabic-English statistical machine translation system



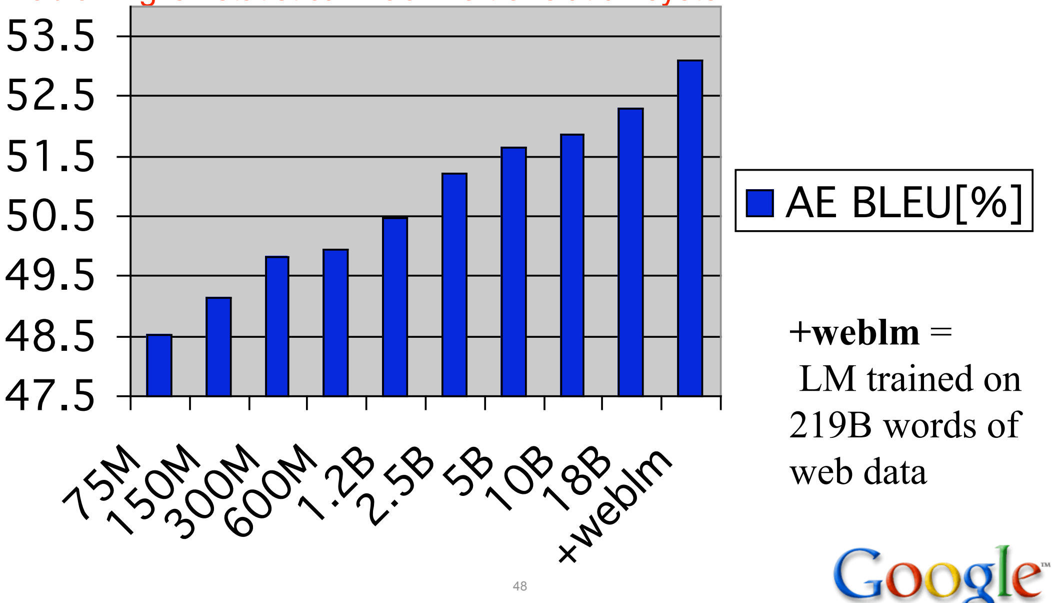
More data is better data…

Impact on size of language model training data (in words) on quality of Arabic-English statistical machine translation system



More data is better data…

Impact on size of language model training data (in words) on quality of Arabic-English statistical machine translation system



+weblm = LM trained on 219B words of web data

More data is better data…

Effect on Arabic-English BLEU score of increasing data by factor of two:

• Double news data for LM corpus:	+0.5% BLEU score
• Double parallel training corpus:	+2.5% BLEU score

If we had:

- 4 times as much parallel training data (600M token)
- 4 times as much monolingual in-domain training data (150B token)

We’d expect BLEU score of 53% + 3% + 3% = 59%

Data: Challenges

Language models trained on > 200 billion words are **huge**

- 45.6 billion 5-grams
- 1.5 terabyte of count data
- 66.5% singletons: but, filtering rare events seems to hurt (tiny bit)

Sophisticated infrastructure necessary

Literature

Data collection:

- [Resnik, 1998]: Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text, AMTA 98.
- [Resnik & Smith, 2003]: The Web as Parallel Corpus, Computational Linguistics 2003.

Document alignment:

- [Nie, Simard, Isabelle & Durand]: Cross-Language Information Retrieval

Sentence alignment:

- [Gale & Church, 1994]: A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics 94.
- [Moore, 2002]: Fast and Accurate Sentence Alignment of Bilingual Corpora, AMTA2002.

Overview

Foundations of Statistical Machine Translation

Automatic Evaluation of Machine Translation

Data

Statistical alignment models

Phrase-based models

Search

Discriminative Training

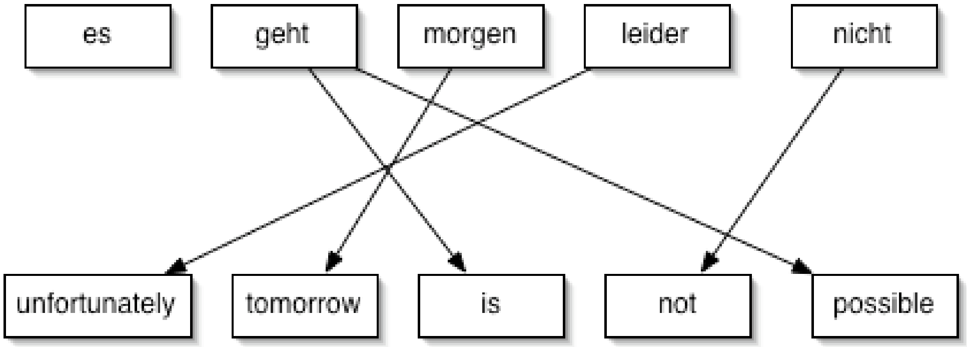
System Building

Results

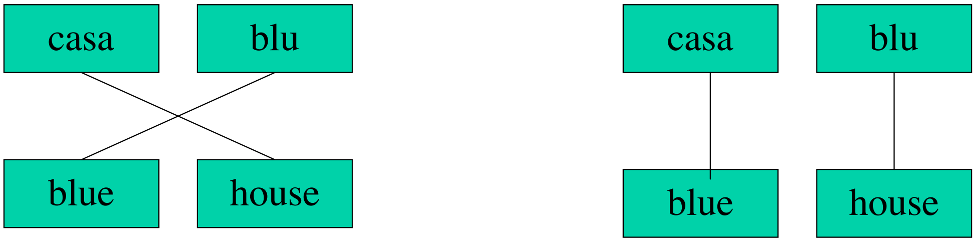
Outlook

Alignment -- Example

Alignment: description of word-to-word correspondences between English and Foreign string



Alignment



Many alignments possible

Some make sense, others don't

Alignment -- Mapping

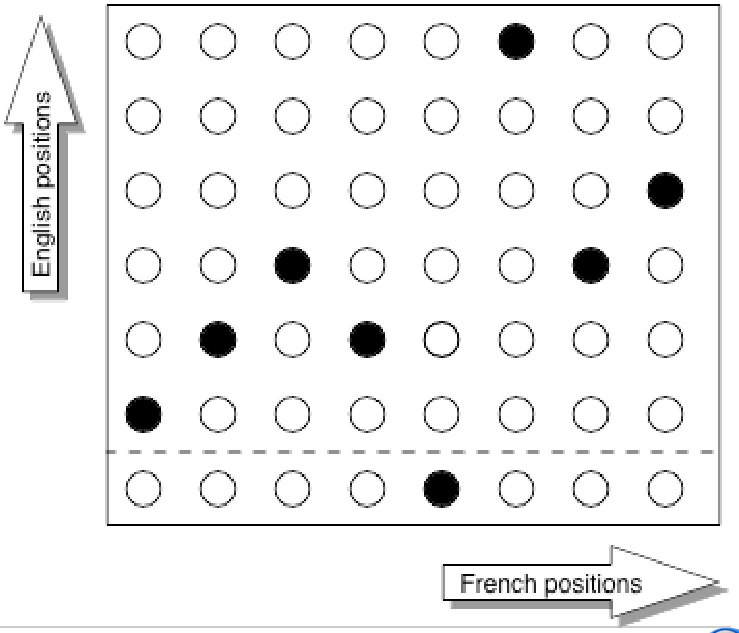
Source language string: $\mathbf{f} = f_1 \ldots f_j \ldots f_J = f_1^J$
Target language string: $\mathbf{e} = e_1 \ldots e_i \ldots e_I = e_1^I$
Aligment mapping: $\mathbf{a} = a_1^J = a_1 \ldots a_j \ldots a_J = a_1^J$

- Alignment is mapping from source language positions to target language positions

  f_j is aligned to e_{aj}

- a_j=0 : f_j is aligned to imaginary empty word e_0
- Note: with this definition it is impossible to represent that one Foreign word aligns to more than one English word
- $(I+1)^J$ different alignments

Alignment -- Mapping



Alignment -- Formal

Alignment is hidden variable of translation:

$$Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} Pr(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

Alignment posterior probability:

$$Pr(\mathbf{a}|\mathbf{f}, \mathbf{e}) = \frac{Pr(\mathbf{f}, \mathbf{a}|\mathbf{e})}{Pr(\mathbf{f}|\mathbf{e})}$$

Decomposition Without Loss Of Generality

$$\begin{aligned}
Pr(f_1^J|e_1^I) &= \sum_{a_1^J} Pr(f_1^J, a_1^J|e_1^I) \\
Pr(f_1^J, a_1^J|e_1^I) &= Pr(J|e_1^I) \cdot Pr(f_1^J, a_1^J|J, e_1^I) \\
&= Pr(J|e_1^I) \cdot Pr(a_1^J|J, e_1^I) \cdot Pr(f_1^J|a_1^J, J, e_1^I) \\
&= Pr(J|e_1^I) \prod_{j=1}^{J} Pr(a_j|a_1^{j-1}, J, e_1^I) \cdot \\
&\quad \prod_{j=1}^{J} Pr(f_j|f_1^{j-1}, a_1^J, J, e_1^I)
\end{aligned}$$

Model 1: length model, uniform alignment model, pairwise lexical dependencies

length model:

$$Pr(J|e_1^I) = p(J|I)$$

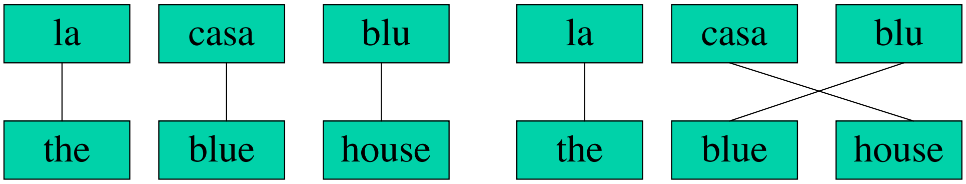
alignment model:

$$Pr(a_j|a_1^{j-1}, J, e_1^I) = \frac{1}{I+1}$$

lexicon model:

$$Pr(f_j|f_1^{j-1}, a_1^J, j, e_1^I) = p(f_j|e_{a_j})$$

Example: Model 1 -- Pr(**a**,**f**|**e**)



P(3|3)

1/4 1/4 1/4

P(la|the)P(casa|blue)

P(blu|house)

P(3|3)

1/4 1/4 1/4

P(la|the)P(casa|house)

P(blu|blue)

Conclusion:
Given
word-to-word-lexicon
we can compute
alignment-probabilities

How to build word-to-word-lexicon?

Parallel corpus:

das Haus ist gelb
the house is yellow

das Auto ist blau
the car is blue

das blaue Haus ist schoen
the blue house is nice

Co-occurrences:

Haus - house : 2
Haus - is : 2
Haus - yellow : 1
Haus - the : 2
Haus - not : 1
Haus - blue : 1
Haus - car : 0
…
ist : is - 3
das : the - 3

How to build word-to-word-lexicon?

Observation:

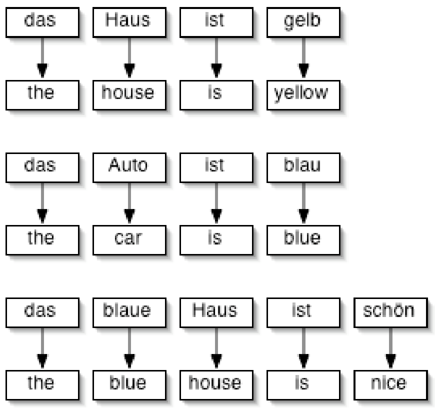
- Co-occurring words: potential translations
- Frequently co-occurring words: likely translations
- Rarely co-occurring words: unlikely translations

Idea:

- estimate translation probabilities using co-occurring counts

Problem: co-occurrences are very noisy

Word-to-word probabilities with **known** alignments



das Haus ist gelb

the house is yellow

das Auto ist blau

the car is blue

das blaue Haus ist schön

the blue house is nice

Haus - house : 2 occurrences
P(*Haus*|house) = 1.0

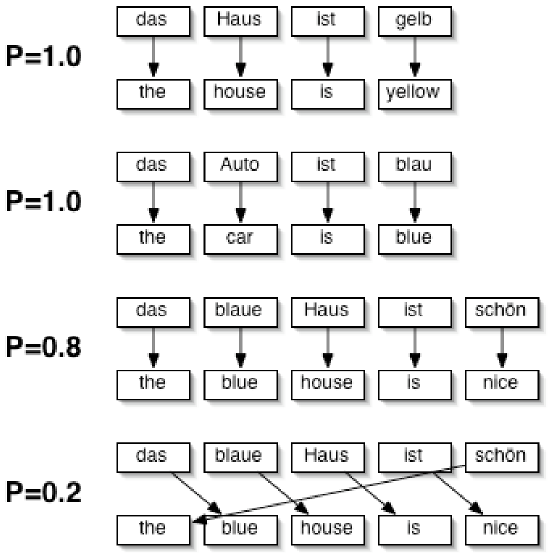
blau - blue : 1
blaue - blue : 1

P(*blau*|blue) = 1/2 = 0.5
P(*blaue*|blue)= 1/2 = 0.5

P(f|e) = N(f,e)/N(e)

Given given alignment information: simple relative frequency

Word-to-word probabilities with **uncertain** alignments



P=1.0
das Haus ist gelb
the house is yellow

P=1.0
das Auto ist blau
the car is blue

P=0.8
das blaue Haus ist schön
the blue house is nice

P=0.2
das blaue Haus ist schön
the blue house is nice

Haus - house : 1.8 times
blaue - house: 0.2 times
P(*Haus*|house) = 1.8/(1.8+0.2)
P(*blaue*|house) = 0.2/(1.8+0.2)

blaue - blue : 0.8
das - blue : 0.2
blau - blue : 1.0
P(*blaue*|blue) = 0.8/2.0=0.4
P(*das*|blue)=0.2/2.0=0.1
P(*blau*|blue)=1.0/2.0=0.5

Word-to-word probabilities with **uncertain** alignments

N(f,e:**a,f,e**): count of alignment between (f,e) in sentence pair **f,e** with alignment **a**

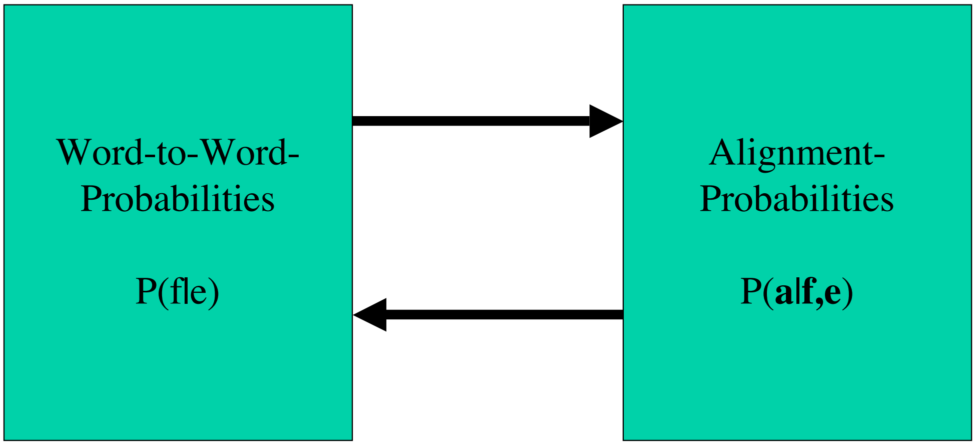
c(f|e): fractional counts -- counts weighted with alignment probability

$$c(f|e) = \sum_{\mathbf{a},\mathbf{e},\mathbf{f}} p(\mathbf{a}|\mathbf{f},\mathbf{e}) N(f,e;\mathbf{a},\mathbf{f},\mathbf{e})$$

$$p(f|e) = \frac{c(f|e)}{\sum_f c(f|e)}$$

Conclusion:
Given
alignment-probabilities
we can compute
word-to-word-lexicon

Chicken-Egg-Problem



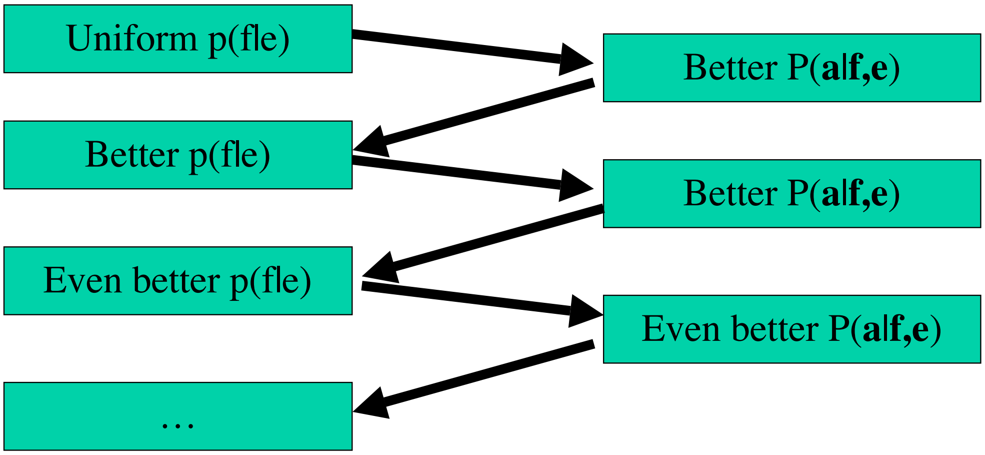
Word-to-Word-Probabilities

P(f|e)

Alignment-Probabilities

P(**a**|**f**,**e**)

Solution: EM-algorithm



Iteratively re-estimate parameters given previous setting

Starting uniformly

Solution: EM-Algorithm

Properties of EM-algorithm:

- Can be shown to converge
- Can be formally derived from **maximum-likelihood principle**
- For Model 1: converges to global optimum
- For more complicated models (later slides):
  - Global optimum not guaranteed
  - Start with a few iterations of Model 1 and then switch to more complicated model

More sophisticated models

IBM Model 2:

- Adds dependence on absolute word positions

$$Pr(a_j|a_1^{j-1}, J, e_1^I) = p(a_j|j, I, J)$$

- Can learn for example that words at the beginning of a sentence are often also translated at the beginning

HMM:

- Adds dependence on relative word positions

$$Pr(a_j|a_1^{j-1}, J, e_1^I) = p(a_j|a_{j-1}, I)$$

- Can learn for example that alignments are often monotone

More sophisticated models

Model 3 (& 4,5):

- Adds new probability distribution p(n|e) for the fertility of words
- Fertility of e: number of Foreign words that e aligns to
- Adds soft coverage constraint for English words
  - Note: Model 1 or 2 has no penalty if English words are ‘unused’

Context-dependent lexicon model

- Takes into account word context

$$Pr(f_j|f_1^{j-1}, a_1^J, j, e_1^I) = p(f_j|e_{a_j}, e_{a_j-1}, e_{a_j+1})$$

- See [Berger, Pietra & Pietra 1996]: further decomposed using maximum entropy modeling

Heuristic alignment combination methods

Observation: alignment from French to English and from English to French provide different views

- f->e alignment (mapping): e-words can align to multiple f-words
- e->f alignment (mapping): f-words can align to multiple e-words

Idea: combine different alignment views into ‘generalized’ alignment by

- Union: produces high recall alignment
- Intersection: produces high precision alignment
- Refined method from [Och & Ney, 2003]: tries to improve precision and recall

Statistical alignment models -- Outlook

Current challenges for statistical alignment models:

- What is the relationship of alignment quality with translation quality?
  - So far: (unfortunately) only very weak relationship

- Existing generative models (Model 1-5, HMM) hard to refine
  - Models get very complicated
  - Promising new methods: log-linear models and discriminative training for word alignment
    e.g. [Fraser & Marcu, 2005]

Literature / GIZA++

Generative statistical alignment models:

- [Brown, Pietra, Pietra & Mercer, 1993]: The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, 1993.
- [Vogel, Ney & Tillmann, 1996]: HMM-based word alignment in statistical translation, COLING96.
- [Och & Ney, 2000]: A Comparison of Alignment Models for Statistical Machine Translation, COLING00.
- [Och & Ney, 2003]: A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, 2003.

Context-dependent lexicon model:

- [Berger, Pietra & Pietra, 1996]: A maximum entropy approach to machine translation, 1996.

Discriminatively trained statistical alignment model:

- [Fraser & Marcu, 2005]: ISI's Participation in the Romanian-English Alignment Task, ACL05.

GIZA++ word alignment tool (GPL):

- www.fjoch.com/GIZA++.html

Overview

Foundations of Statistical Machine Translation

Automatic Evaluation of Machine Translation

Data

Statistical alignment models

Phrase-based models

Search

Discriminative Training

System Building

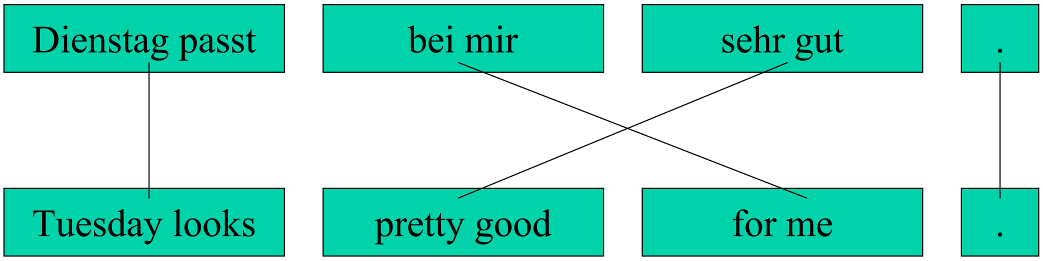
Results

Outlook

Phrase-based statistical machine translation

Basic idea:

- Learn phrase-dictionary by storing **all** aligned phrase pairs in training corpus



- SL and TL are segmented into sequence of phrases
- “Phrase”: any consecutive sequence of words
- Phrases can be reordered (but only as a whole)

Advantages of phrase-based MT (compared to single-word based MT)

Phrases capture local reordering

- Single-word based: needs to be stored in alignment model

Local context useful for disambiguation

- Single-word based: only target language model does disambiguation

Phrases are reordered as a whole

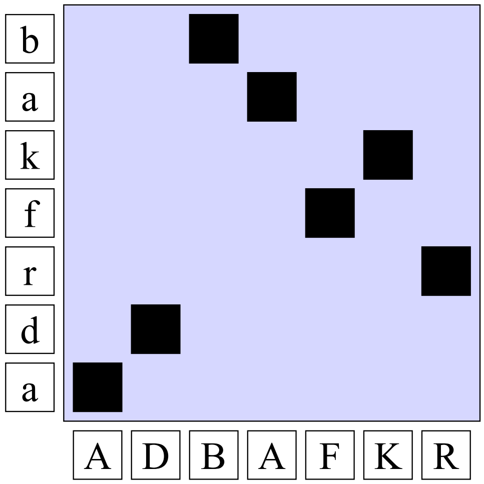
Works well for non-compositional phrases

With a lot of data: sometimes whole sentences can be covered

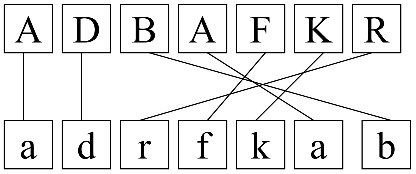
Disadvantage:

- Huge storage space needed to store all phrases

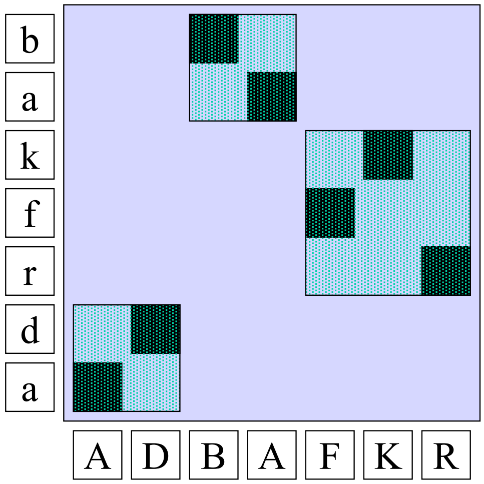
How to train phrases?



Basic idea: Find all aligned phrase pairs which are consistent with the word alignment



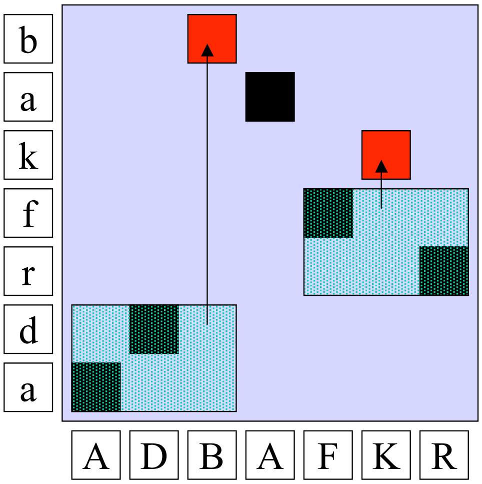
How to train phrases?



Basic idea: Find all aligned phrase pairs which are consistent with the word alignment

Examples of phrases which are consistent with word alignment

How to train phrases?

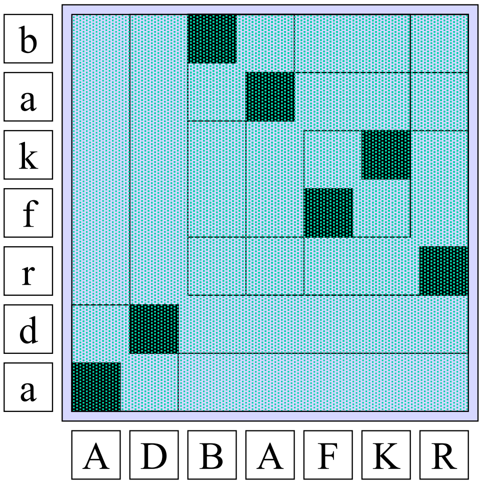


Basic idea: Find all aligned phrase pairs which are consistent with the word alignment

Examples of phrases which are **not** consistent with word Alignment

🟥 out-of-phrase alignments

TM Training: Phrase extraction



Basic idea: Find all aligned phrase pairs which are consistent with the word alignment

Scoring phrase pairs

Provide various signals for assessing ‘quality’ of phrase:

- Phrase translation probability p(e|f), p(f|e)
  - Estimation by relative frequency
- Word translation probabilities
  - Phrase pair scored by Model 1 alignment
- …
- Each signal provides a feature function for overall translation model

Many different scoring methods suggested in literature

Refinements

Alignment templates:

- Generalize phrases by replacing words by word classes
- [Och, 2003]

Hierarchical phrases:

- Allow gaps in phrases
- [Block, 2000]; [Chiang, 2005]

Literature

Foundations of described phrase-based/template-based models:

- [Och, Tillmann & Ney, 1999]: Improved Alignment Models for Statistical Machine Translation, EMNLP99.
- [Och, 2003]: Statistical Machine Translation: From Single-Word based Models to Alignment Templates, PhD thesis, RWTH Aachen, 2003.
- [Och & Ney, 2004]: The Alignment Template Approach to Statistical Machine Translation, Computational Linguistics, 2004.

Other approach: learn phrases and alignment in one-shot

- [Marcu & Wong, 2002]: A Phrase-Based Joint Probability Model for Statistical Machine Translation, EMNLP02.

Approach to reduce storage requirement for phrase-base models:

- [Burch, Bannard & Schroder, 2005]: Scaling Phrase-Based Statistical Machine Translation to Larger Corpora and Longer Phrases, ACL05.

Hierarchical phrases:

- [Block, 2000]: Example-based synchronous interpretation, In: Wolfgang Wahlster (editor): Verbmobil: Foundations of Speech-to-Speech Translation, Springer Verlag.
- [Chiang, 2005]: A hierarchical phrase-based model for statistical machine translation, ACL05.

Overview

Foundations of Statistical Machine Translation

Automatic Evaluation of Machine Translation

Data

Statistical alignment models

Phrase-based models

Search

Discriminative Training

System Building

Results

Outlook

Search problem: log-linear model

$$Pr(e,h|f) = \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(e,h,f))}{\sum_{e',h'} \exp(\sum_{m=1}^{M} \lambda_m h_m(e',h',f))}$$

$h_m(e, h, f)$: feature function

λ_m: feature function weight

decision rule:

$$\hat{e}(f, \lambda_1^M) = \text{argmax}_e \text{max}_h \sum_{m=1}^{M} \lambda_m h_m(e, h, f)$$

Phrase-based model:

- Hidden variable h is sequence of phrase instantiations

Search

Problem: huge search space (all possible English sentences)

- Finding optimal translation for Model 1 and bigram LM is NP-complete [Knight, 1999]
- Exact search (e.g. A*) prohibitive for long sentences

Heuristic search methods:

- Beam-search (described in the following)
- Greedy search

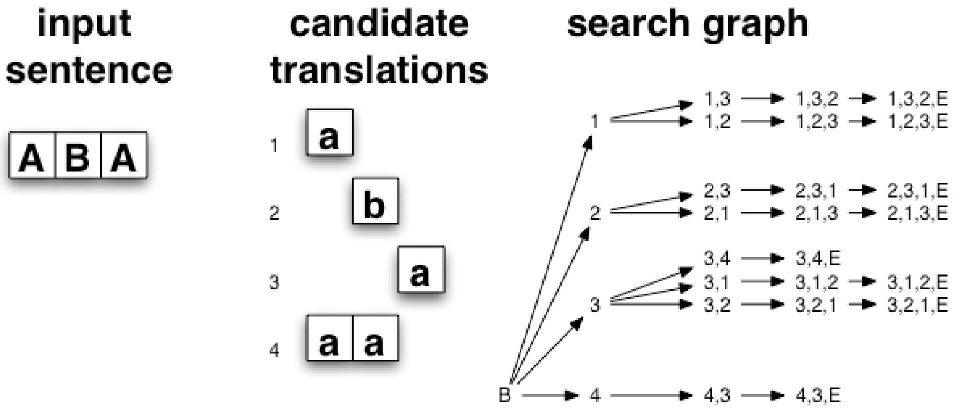
DP for phrase-based models

Phrase-based models:

- **Candidate phrases**: each phrase from learned phrase table that is applicable in sentence to be translated
- Selected phrases: set of phrases chosen to translate sentence
- **Coverage constraint**: selected phrases need to cover each SL position exactly once
- **Hypothesis:** partial sequence of selected phrases (prefix of translation); includes accumulated partial scores
- **Extending a hypothesis:** computing all hypotheses that can be produced by adding one additional phrase to a hypothesis
- **Search space:** Graph with hypotheses as nodes and extensions as edges

Search graph for phrase-based SMT (no recombination)

Example: three word sentence; 4 candidate translations

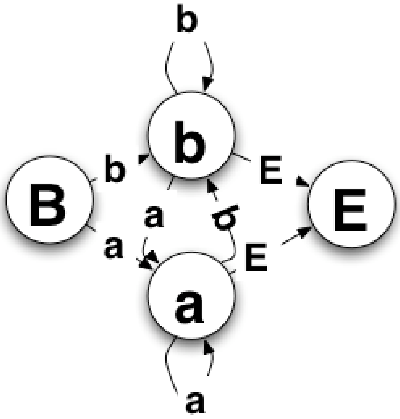


DP Search

Assumption: feature functions are local

- Future decisions depend only on limited history:
  - Dependencies can be represented as finite-state automaton

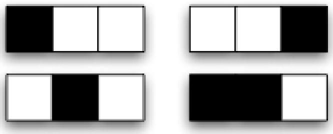
- E.g. state space of bigram language model with symbols ‘a’, ‘b’
- Begin/End symbols: ‘B’/‘E’
- Future bigram probabilities only depend on the most recently produced word

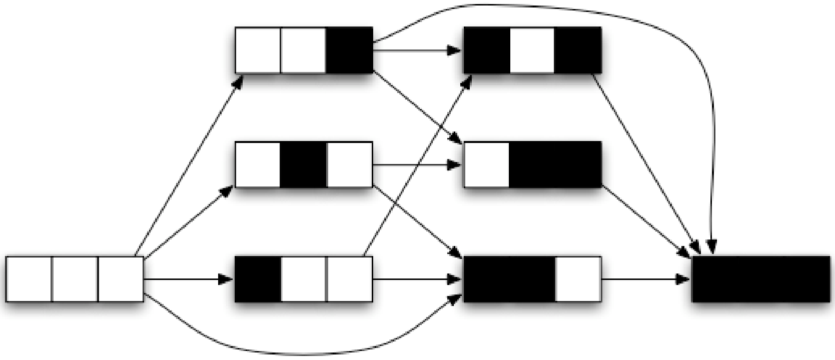


DP Search

State-space of phrase coverage constraint:

available phrases:





DP Search

State-space of ‘jump-distance feature function’:

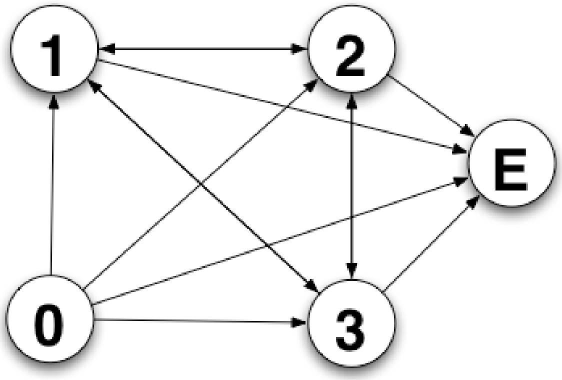
- Feature function sums ‘jump-distances’ for a translation
- Jump: distance between end-position of j-th phrase and first position of (j+1)-th phrase

States:

0: beginning of sentence

1,2,3: end-position of previous translated phrase

E: end of sentence



DP Search

Search problem:

- Find optimal path through combined (composed) search space

Recombination:

- If two (or more) hypotheses are indistinguishable given their state:
  - Just keep best hypothesis

Back-pointer:

- Pointer to previous best hypothesis

Reordering constraints

Even using recombination: search space is too big

- Main reason: reordering feature function
  - # of different coverage vectors: 2^n for n-word-sentence

Frequently used to further reduce search space: additional reordering constraint

- Often: only local reordering (5-8 words)
- Allow reordering for at most m words (m<<n)
  - # of different coverage vectors: $< n\ 2^m$

Beam-search

Beam-Search:

- Prune(=ignore) bad scoring search hypotheses as soon as possible
- Constrain maximal number of hypotheses

For all hypotheses which cover the same number of source positions:

- Allow at most N hypotheses
- Prune all hypotheses which are worse than α times best hypothesis

Refinement:

- Coverage pruning: have separate thresholds for hypotheses that cover the same source positions (not only the same number of positions)

Tuning Beam-Size:
Trade-off between Quality and Efficiency

Example beam tuning from [Tillmann & Ney, 2003]:

- German-English Verbmobil, 331 test sentences
- Search errors are computed as lower bound: using the hypotheses found with widest beam as reference

-log(α)	CPU time [s]	Search errors	mWER[%]
1.0	0.03	287	48.5
2.0	0.06	277	41.9
4.0	0.30	239	34.1
7.5	3.2	106	26.6
10.0	14.2	32	25.1
12.5	42.2	5	24.9
17.5	176.7	0	24.9

Additional methods

Introduction of estimate of future costs for scoring hypotheses in DP beam-search

- Example of (optimistic) future cost estimate: choosing for each uncovered position the most likely translation in the probabilistic lexicon

Greedy search:

- Start with some translation (e.g. monotone translation)
- Iteratively improve translation probability by performing small changes (until convergence)

Literature + Tools

Dynamic Programming Search:

- [Tillmann & Ney, 2003]: Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation, Computational Linguistics 2003.
- [Och & Ney, 2004]: The Alignment Template Approach to Statistical Machine Translation, Computational Linguistics, 2004.

A* search:

- [Och, Ueffing & Ney, 2001]: An Efficient A* Search Algorithm for Statistical Machine Translation, DDMT01.

Decoding is NP-complete:

- [Knight, 1999]: Decoding complexity in word-replacement translation models, Computational Linguistics, 1999.

Greedy Search:

- [Germann, 2003]: Greedy decoding for statistical machine translation in almost linear time, ACL03.

Finite-state implementation for phrase-based models:

- [Kumar & Byrne, 2003]: A weighted finite state transducer implementation of the alignment template model for statistical machine translation, HLT/NAACL'03.

Rewrite decoder -- single-word based decoder for Model 4:

- www.isi.edu/licensed-sw/rewrite-decoder/

Pharaoh decoder -- a beam-search decoder for phrase-based machine translation:

- www.isi.edu/licensed-sw/pharaoh/

Overview

Foundations of Statistical Machine Translation

Automatic Evaluation of Machine Translation

Data

Statistical alignment models

Phrase-based models

Search

Discriminative Training

System Building

Results

Outlook

Goals:

- Optimize model parameters of the overall system (as opposed to optimize language and translation model independently)
- Directly train model parameters for producing good translations (as measured by an automatic evaluation metric)

Advantages:

- Better translation quality
- We can easily build systems that perform well for specific evaluation criteria (e.g. dependent on the specific application)

Minimum Error Training

Decision Rule:

$$\hat{e}(f, \lambda_1^M) = \mathrm{argmax}_e \mathrm{max}_a \sum_{m=1}^{M} \lambda_m h_m(e, a, f)$$

Resulting Minimum-Error training criterion:

$$\hat{\lambda}_1^M = \mathrm{argmin}_{\lambda_1^M} \left\{ \sum_{s=1}^{S} E(r_s, \hat{e}(f_s; \lambda_1^M)) \right\}$$

E(r,e): number of errors comparing translation ‘e’ with reference ‘r’

- Can directly optimize for mWER, mPER, …: simply change E(r,e)
- Straightforward extension possible for BLEU, NIST

Challenges + Solutions

Challenge 1: no gradient descent possible

- Objective function is piecewise constant
- search for good parameter setting: efficiently try out many values

Challenge 2: many local optima

- using multiple starting points

Challenge 3: danger of overfitting

- small number of parameters

Challenge 4: embedded optimization for decision rule (= search)

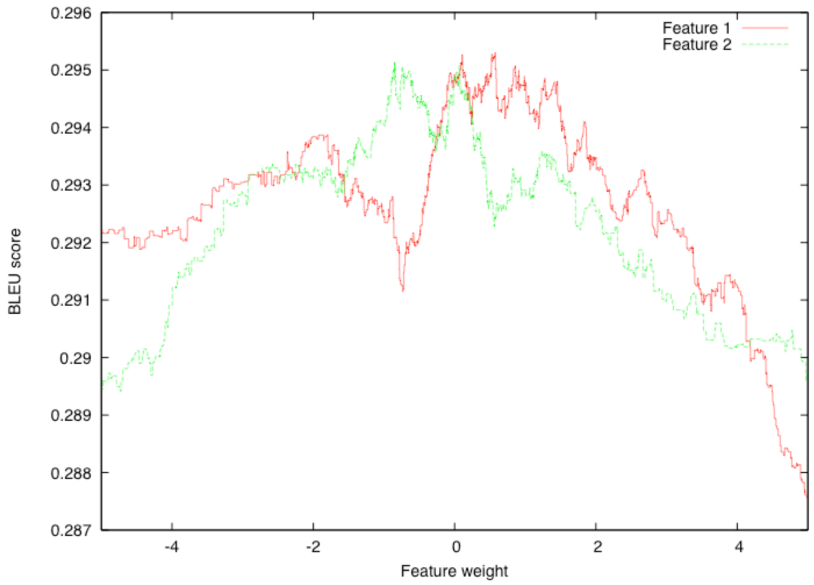
- optimize decision rule only over limited (N=1.000…10.000) list of candidate translations: **N-best list**

M-dim. Optimization Algorithm

DO N times:

- Choose random start point
- While not converged
  - FOR each parameter m=1…M
    - Optimize parameter value with respect to error rate (one-dimensional opt. problem)
  - Change parameter which gives largest improvement

Example objective function for maximum-BLEU training for two feature functions

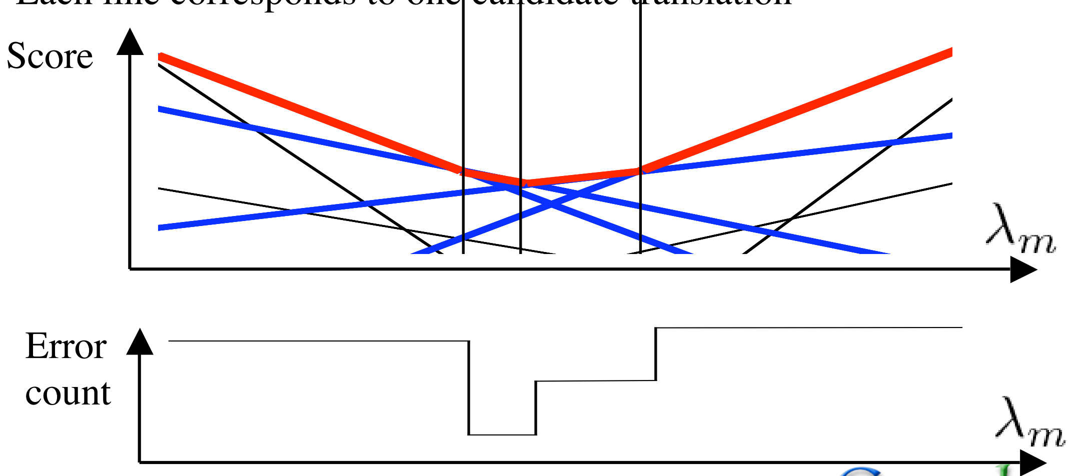


1-dim. Optimization Problem

Optimizing for one parameter = optimize over set of lines

$$\hat{e}(f_s, \lambda_m) = \text{argmax}_{e \in C} \lambda_m h_m(e, f) + H_{-m}(f, e)$$

Each line corresponds to one candidate translation



1-dim. Optimization Algorithm

Algorithm:

- For each sentence:
  - compute sequence of ‘decision boundaries’
    - By maximizing over set of lines(=candidates) and enumerating the intersection points of intersecting lines
  - compute sequence of error count changes
- Traverse all sequences of error count changes and find minimum error count

Adapting to BLEU/NIST straightforward

- Instead of ‘error count change’ compute more sophisticated statistics (for BLEU: # correct n-grams, hypothesis length, reference length)

Handling of N-best lists in minimum-error training

Problem: N-best list might be too small

Solution: iteratively refine N-best lists

1. Initialize parameters
2. Perform search and extract N-best list
3. Optimize model parameters using N-best list
4. Perform new search with optimized model parameters and compute new N-best list
5. Merge old N-best list with new N-best list
6. While not converged goto 3

Typically converges after 7..20 iterations

Example results of maximum BLEU training from [Och 2003]

test	mWER	mPER	BLEU	NIST
train	[%]	[%]	[%]	
mWER	**70**	53	15	5.9
mPER	72	**52**	17	6.6
BLEU	77	55	**20**	6.9
NIST	74	53	19	**7.1**

Conclusion: optimizing for certain metric gives best results for that metric (mWER, mPER: lower = better)

Results

Minimum error/maximum score training gives often significantly better results

- Standard approach in best systems (NIST MT evaluation)
- Without maximum-BLEU training: a lot of manual system tuning necessary

Taking into account evaluation criterion in training is important

- Better evaluation metrics can directly lead to better MT

Literature

Discriminative Training for Machine Translation:

- [Och & Ney, 2002]: Discriminative Training and Maximum Entropy Models for Statistical machine Translation, ACL02.
- [Och, 2003]: Minimum Error Rate Training for Statistical Machine Translation, ACL03.
- [Shen, Sarkar & Och, 2004]: Discriminative Reranking for Machine Translation, HLT/NAACL04.

Discriminative Training of Language Model:

- [Roark, Saraclar, Collins & Johnson, 2004]: Discriminative language modeling with conditional random fields and the perceptron algorithm, ACL04.

Overview

Foundations of Statistical Machine Translation

Automatic Evaluation of Machine Translation

Data

Statistical alignment models

Phrase-based models

Search

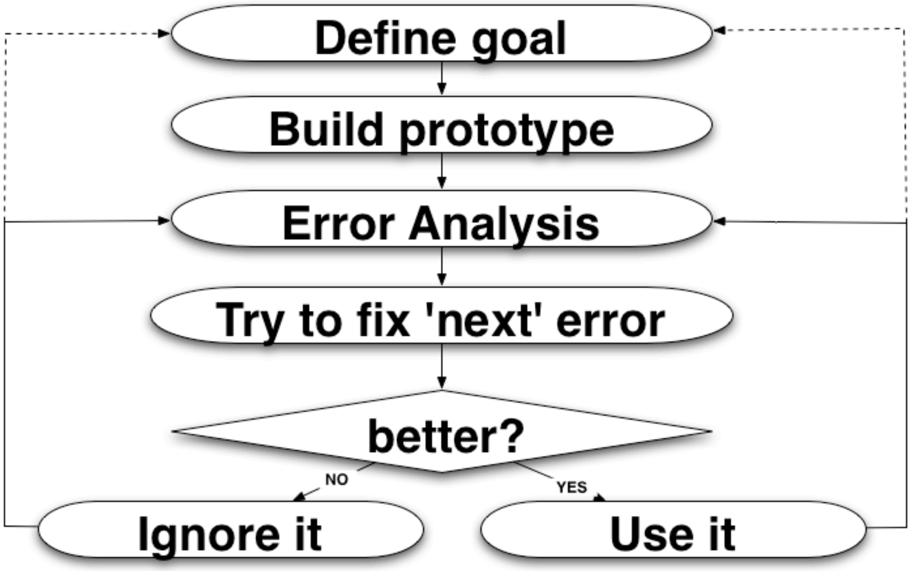
Discriminative Training

System Building

Results

Outlook

Development Cycle for MT Research



Some issues in systems building

Observation: building state-of-the-art research system for statistical machine translation research is non-trivial

- Huge differences in quality for phrase-based statistical machine translation systems in NIST MT evaluation

Major challenges:

- Large quantities of data
  - Slow experimentation cycle
- Many system components interact in complicated ways
- Unclear failure criteria:
  - if MT output is bad: system bug or modeling/training/search error?

Other important components

- Transliteration of foreign language words
  - ‘Simple’ statistical machine translation problem
- Rule-based translation for numbers, dates, bylines
- Postprocessing (de-tokenization, headline-casing, …)
- Sub-sentence aligner (to make task for word alignment easier)
- Error analysis tools

Some tricks…

Problem: Phrase tables for large corpora become **huge** (tens of Gigabyte)

- Solution: compute phrase table only for sentences in development / test corpus
- Note: relative frequency computation of p(f|e) this can be taken into account already in phrase extraction (not for p(e|f) …)

Problem: Language models for large corpora become **huge** (hundreds of Gigabyte)

- Similar solution: compute for each test corpus sentence a sentence-specific language model filtered for the vocabulary of that sentence

More tricks…

How many bits are needed to store probabilities?

- Double: 64 bit?
- Float: 32 bit?
- 16, 8, …?

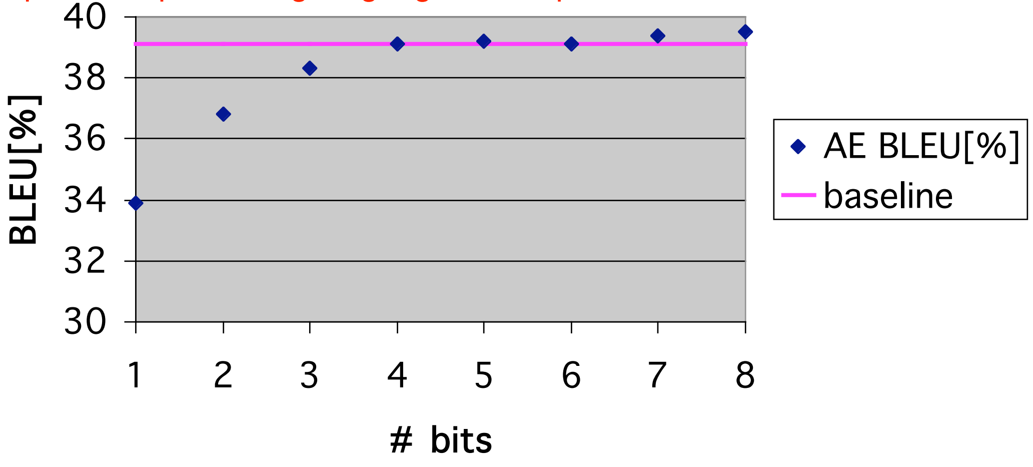
For efficient representation of large models:

- use minimum number of bits that do not lead to loss in performance

Here: quantization method quantizes linearly in log-prob-space between probability 1.0 and smallest occurring probability

How many bits are needed to store probabilities?

Impact of representing language model probabilities with N bits



Astonishing result: 4 bits are enough(!)

More tricks…

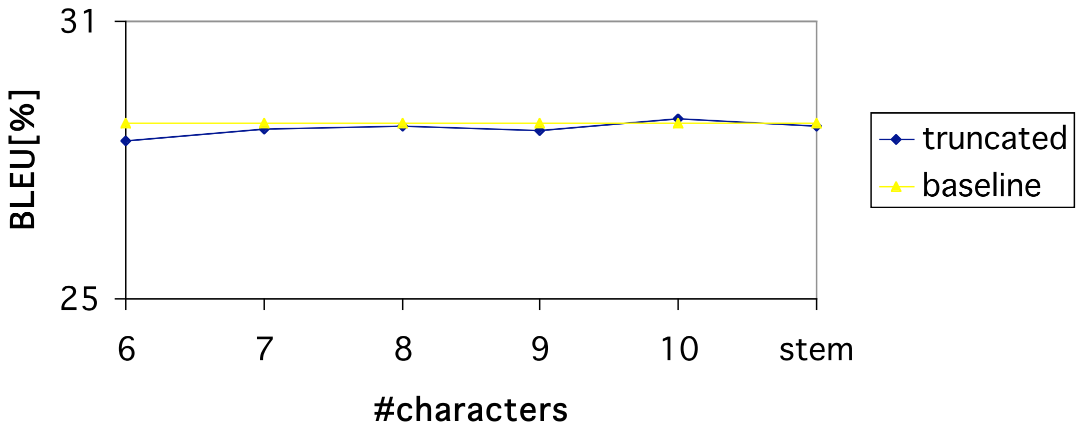
Problem: word alignment (GIZA++) needs a lot of memory because it stores all lexical co-occurrences

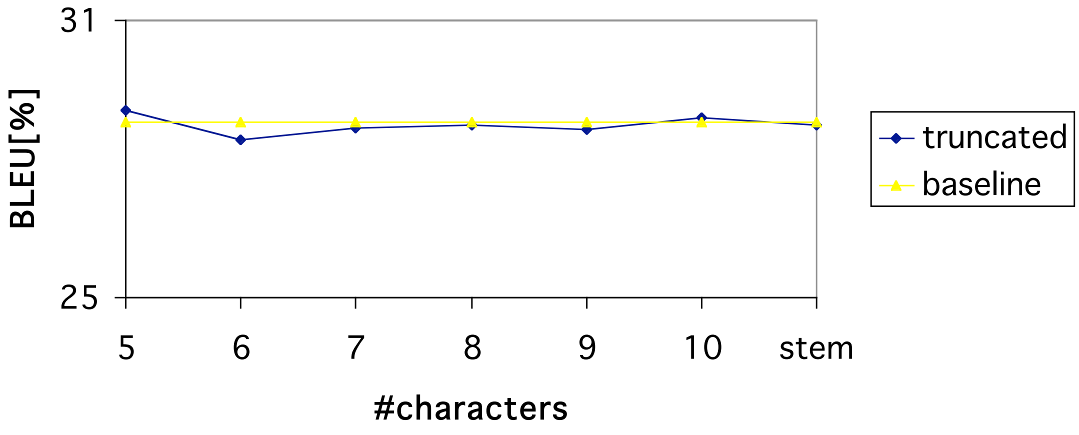
Idea:

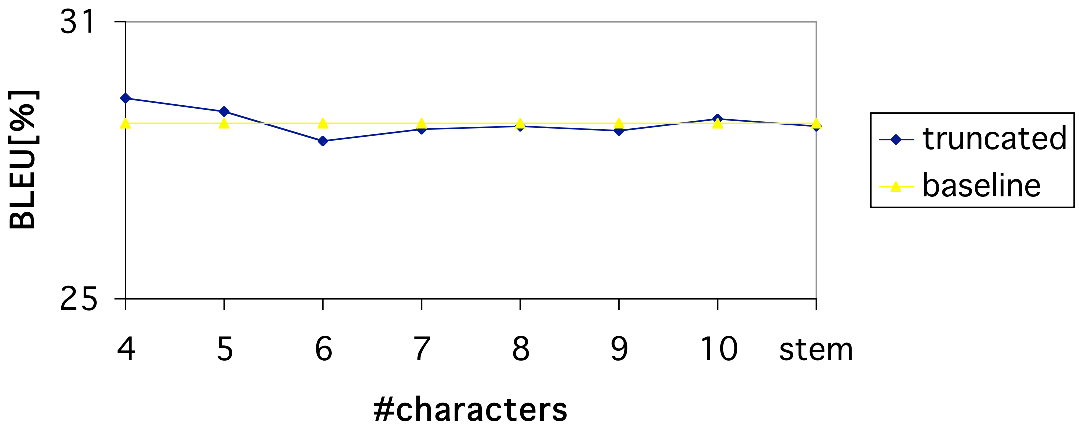
- Replace (for word alignment) words by their word stems
- Even simpler: replace words by their first N characters

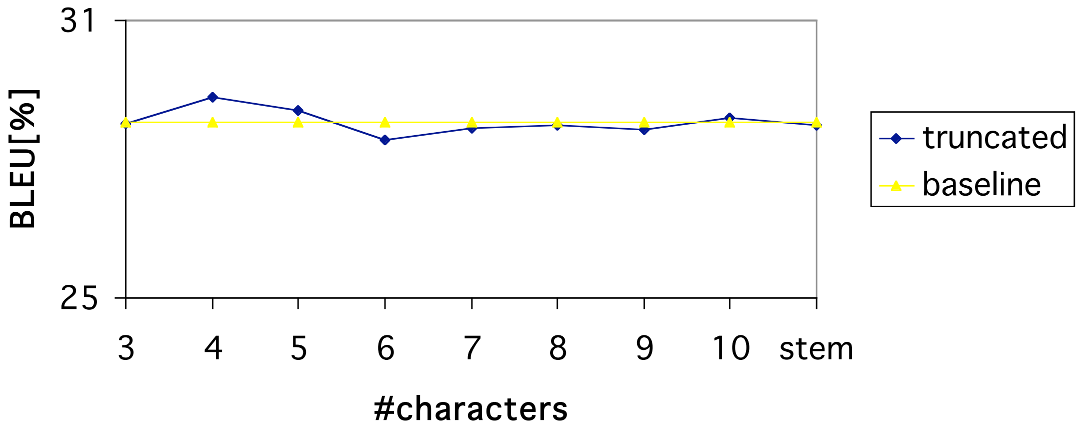
Goal:

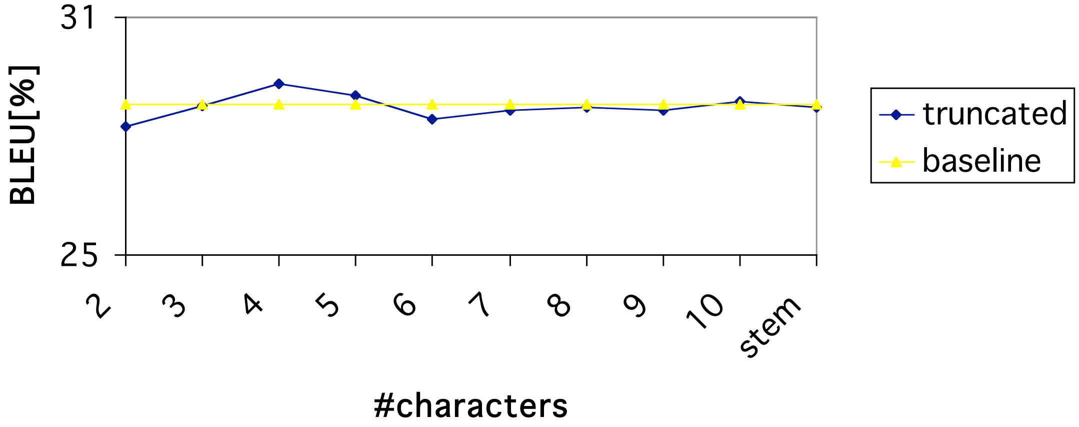
- Use smallest representation that does not hurt alignment quality



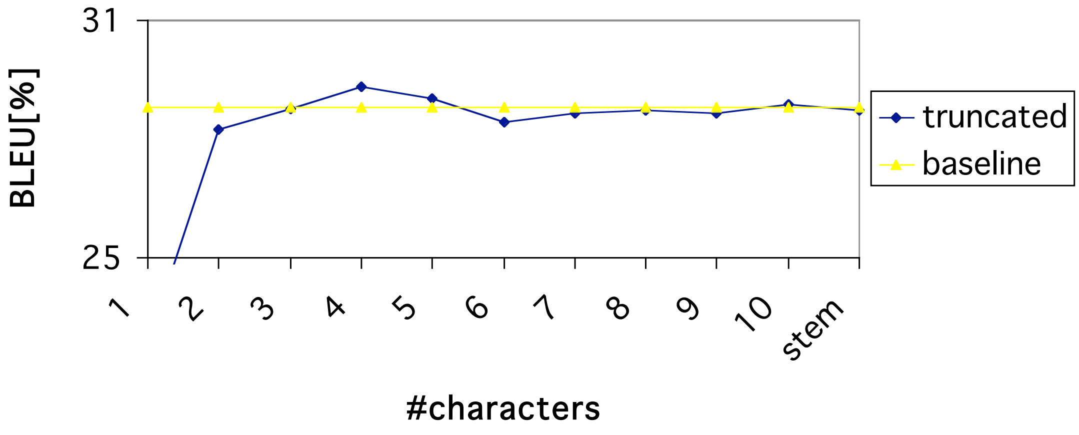

31
25
BLEU[%]
5
6
7
8
9
10
stem
#characters
truncated
baseline


31
25
BLEU[%]
4
5
6
7
8
9
10
stem
#characters
truncated
baseline


31
25
BLEU[%]
3
4
5
6
7
8
9
10
stem
#characters
truncated
baseline


31
25
BLEU[%]
2
3
4
5
6
7
8
9
10
stem
#characters
truncated
baseline

Astonishing result: reducing English to 4/5 characters for word alignment seems to improve BLEU score (though not statistically significant)



Overview

Foundations of Statistical Machine Translation

Automatic Evaluation of Machine Translation

Data

Statistical alignment models

Phrase-based models

Search

Discriminative Training

System Building

Results

Outlook

Translation Example - Arabic-English - I

Ardogan Confirms That Turkey Would Not Accept Any Pressure, Urging Them to Recognize Cyprus

Ankara 1-12 (AFP) - Turkish Prime Minister Recep Tayyip Erdogan announced today, Wednesday, that Ankara will reject any pressure has been exercised by the European Union to urge the recognition of Cyprus, two weeks before the summit of Heads of State and Governments of the European Union who decide to open accession negotiations with Ankara or not.

He Ardogan station "that since" television that "the European Union could not address by imposing new conditions on the Cyprus.

Will discuss this issue during the accession negotiations. ”

He added : "Let me be clear, it may not be me arm Turkey, which can not accept it.”

Translation Example - Arabic-English - II

It is expected to recommend that the European Parliament on 14 December the heads of states and governments to agree to open accession negotiations with Ankara, but with specific conditions.

Greece considers that recognition of Cyprus by Ankara had also become a member of the European Union, "obvious condition" for the European Union gave the green Duah to open accession negotiations with Ankara.

Turkey still refuses to recognize the Republic of Cyprus became a member of the European Union since the first of last May, but the only State in the world that recognized the "Turkish Republic of Northern Cyprus", which declared autonomy in the north of the island.

Ardogan hypothesis also refused to open negotiations to establish a privileged partnership between Turkey and the European Union as a substitute for accession negotiations.

Translation Example - Arabic-English - III

He Ardogan in this topic that "this proposal, which can not accept it, not compatible with the seriousness such as the European Union.”

He added : "We want to start negotiations in the first six months of 2005, or in Section II of the year if it is (...) but we can not determine the date of commencement of negotiations in 2006."

Translation Example - Chinese-English

Euro-Zone Economic Indicators This Week Will Show the Economy Continues Weak

Economists said the euro zone economic indicators will be published this week will show the economy weakening confidence and further evidence of the slowdown in economic output, on the other hand, the British data are expected to show increased industrial activity.

They said that the biggest countries in the euro zone economic output data will continue to show significant monthly fluctuations, but the long-term trends in economic output will show economic growth slowed.

This week will be the most important economic indicators in Germany, the market close attention to the ZEW economic institute's economic expectations indicator scheduled for release on 7 and December will provide the first evidence of economic confidence.

…

Translation Example - Chinese-English

Bahrain Princess U.S. Soldier She Married Five Years Had Broken Marriage

Bahrain Princess Erin Hari married U.S. Marine Corps soldier Strong Health, has won widespread praise for the American TV host, but this particular twists and turns before continuing five years after marriage, it has in the desert Leaving Las Vegas Las Vegas painting of sentence.

Erin is a member of the Gulf State of Bahrain royal family, strong for their lost career risks in this prohibited Lovers of 1991, they were just a fairy tale marriage, has become the subject of a TV movie.

However, according to the Las Vegas review, the report only after five years, two to go, Erin Court to file for divorce, marriage location in running quickly and run faster Leaving Las Vegas divorce.

…

How well does it work?

- Excellent quality compared to state-of-the-art for those language pairs (also: see NIST MT evaluation results)
- Impressive quality of Arabic-English translations
- Chinese-English not as good as Arabic-English
- Very high variance:
  - For some documents: near-perfect translations
  - For other documents: word salad
- To be investigated: impact of genre on translation quality
  - E.g. translating sports news articles seems much harder than political news
- Still a lot of room for improvement…

NIST MT evaluation

- DARPA funded project TIDES
- Machine translation, summarization, information extraction, …
- Yearly evaluation performed by NIST
  - ‘01 (dry run), ‘02 … ‘05
  - Language pairs: Arabic-English, Chinese-English
  - **Highly important** for driving progress in recent years
- This year: 21 participants
  - Universities, industrial research labs, government research labs, companies
  - International: US, Germany, Italy, Britain, China

NIST MT evaluation

Best systems in ‘05 NIST MT evaluation: SMT

- A->E: top-ranked systems are ‘phrase-based statistical’
- C->E: top-ranked systems are ‘phrase-based statistical’

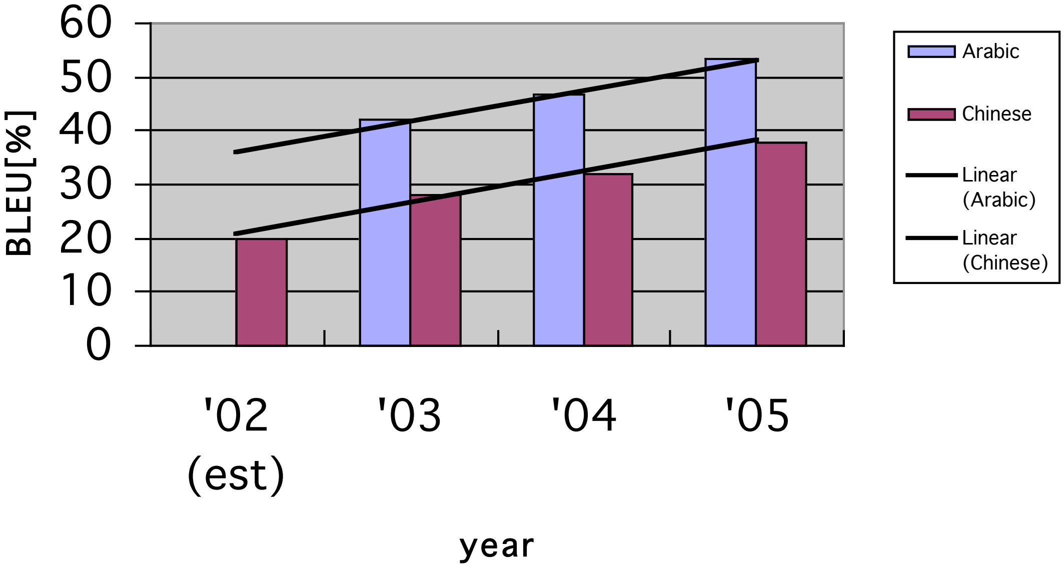
Results:

- http://www.nist.gov/speech/tests/mt/mt05eval_official_results_release_20050801_v3.html
- Or search for: “NIST MT evaluation results”
- Google did well… :-)

Best models are simple (nice!)

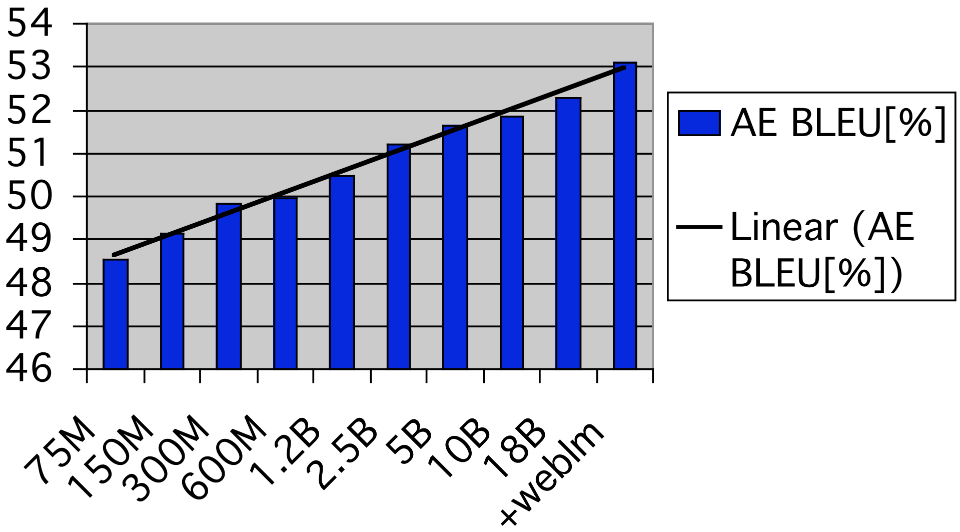
- No parser, POS tagger, NP chunker, WordNet, explicit WSD, …

NIST MT evaluation: Progress over time



More data is better data…

Doubling news LM training corpus size: ~0.5% higher BLEU score



Overview

Foundations of Statistical Machine Translation

Automatic Evaluation of Machine Translation

Data

Statistical alignment models

Phrase-based models

Search

Discriminative Training

System Building

Results

Outlook

Surprises in Statistical Machine Translation

In best systems no use of explicit syntactic models

- No POS tagger, linguistic parsers, shallow parsers, treebanks, …

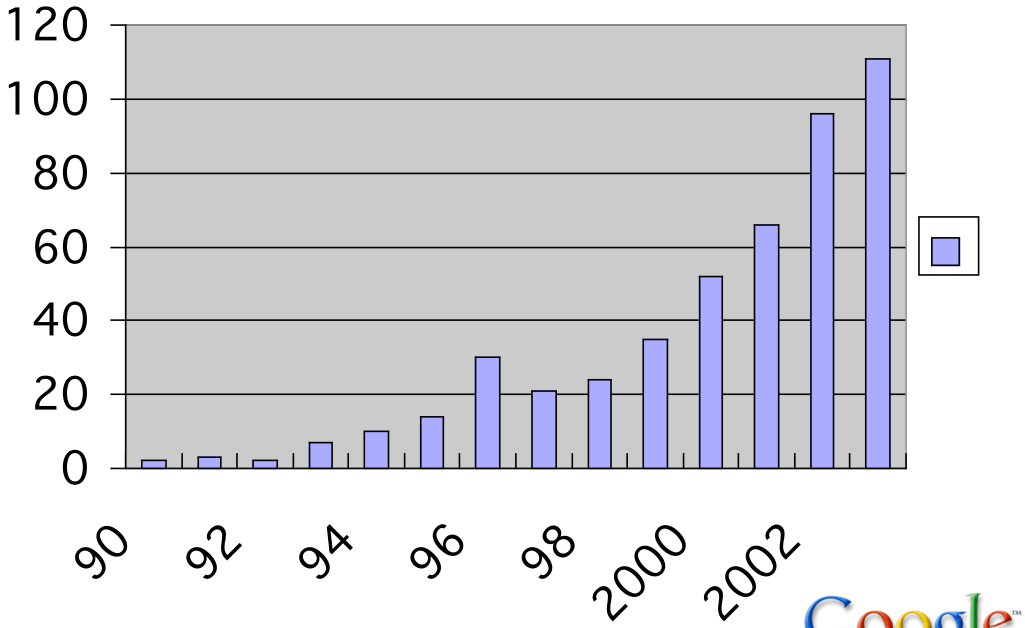
In best systems: no use of explicit semantic models

- No Senseval-style word sense disambiguation, ontologies, PropBank, WordNet, …
- Statistical machine translation is essentially about word sense disambiguation

Arabic: so far no gain of sophisticated morphological analysis

Is there any use of annotated corpora (besides parallel corpora) for MT?

Publications referring to “statistical machine translation”



Outlook: Progress from …

Better Models + Training

- Generalized phrase models (e.g. hierarchical)
- Long-distance dependencies
- Topic adaptation
- Discriminative training with many more features

Much More Data

- Monolingual data: > 1 trillion words
- Bilingual data: > 1 billion words

Better automatic machine translation evaluation (BLEU++)

Better engineering / infrastructure / tools

Translation Example (last document of ae-mteval05)

China, Canada Issued a Joint Statement, Confirming the Sustained Growth

Beijing January 20 / Xinhua / China, Canada issued a joint statement here today, Thursday, they had agreed to unite their efforts to achieve sustainable growth through trade, investment and innovation.

The two ministers stressed the role of the World Trade Organization, and agreed to improve cooperation in the light of this framework.

He said the two sides in the joint statement, "As members of the World Trade Organization and the major trading partners, there is no doubt that the system of multilateral trade rules and the building is vital for the prosperity of both countries."

Translation Example (last document of ae-mteval05)

They agreed to cooperate closely to prepare for the WTO Ministerial Meeting, to be held in Hong Kong later in the year, to promote the successful completion of the Doha Development Agenda at the earliest, according to the statement.

The statement said that China's accession to the World Trade Organization has created unprecedented opportunities for cooperation between companies in both countries, adding, "It will also strengthen cooperation through an ambitious programme of technical assistance to support the full and active participation of China in the World Trade Organization.”

The statement pointed out that bilateral economic relations between China and Canada is strong.

The two sides agreed to strengthen that relationship through a series of concrete initiatives designed to support the continued expansion of partnership in trade, investment and innovation.

Translation Example (last document of ae-mteval05)

In addition, the statement said that the two sides are determined to deal with the problem of global warming in the light of their shared commitment to sustainable development and balanced growth.

Joint Statement issued after the formal talks conducted by the President of the Chinese State Council won Jia Bao with visiting Canadian Prime Minister Paul Martin Thursday afternoon.

Franz Josef Och
Google, Inc.
och@google.com