Current Research in Phrase-Based Statistical Machine Translation

*and some links to ASR*

Bill Byrne

Cambridge University Engineering Department

wjb31@eng.cam.ac.uk

22 February 2005

Work done with Shankar Kumar and Yonggang Deng
Outline

Introduction

Five Easy Problems in Statistical Machine Translation
Automatic Measurement of Translation Quality
Bitext for SMT Training: Document and Sentence Alignment
Translation Models: Word-to-Word Alignment
Translation Models: Phrase-to-Phrase Alignment
Translation

2004 JHU-CLSP NIST MT Evaluation Systems
Baseline Systems and Performance

Summary and Conclusion
Automatic Speech Recognition and Machine Translation

Both can be cast in a generative source-channel modeling framework

- Readable source language text is ‘corrupted’ into some other form
SMT is All About Alignment

Relies on the idea that parallel text can be found and aligned:

- Document Level Alignment
- Sentence Level Alignment
- Word and Phrase Level Alignment

NULL Mr. Speaker, my question is directed to the Minister of Transport

Monsieur le Orateur, ma question se adresse à le ministre chargé de les transports

Hierarchical models of translation:

- trained over collections of known translations
- component models form a complete translation model
Sentence Translation via Alignment Models

**Modeling:** Given an English Sentence $e$ and a French sentence $f$, construct a joint distribution over their alignments.

$$P(e, a, f) = \underbrace{P(f|a, e)} P(a|e) \underbrace{P(e)}$$

**Translation Model**

**Alignment Model**

**Language Model**

**Decoding:** Given $f$, find a translation $\hat{e}$ and an alignment $\hat{a}$ as

$$( \hat{e}, \hat{a} ) = \arg\max_{e, a} P(f|a, e) P(a|e) P(e)$$

**Assumptions:** MAP decoding, uni-directional translation models, source text is ‘generated’ independently of translation hypotheses, search relies on a single alignment hypothesis ($P(f|e) \approx \max_{a} P(f, a|e)$), ...
Can We Pretend MT is ASR?

Of course. We just need:

- Alignment models and estimation algorithms
- Training data: bitext, monolingual text
- Search algorithms for translation
- Some way to measure translation quality
Outline

Introduction

Five Easy Problems in Statistical Machine Translation

Automatic Measurement of Translation Quality

Bitext for SMT Training: Document and Sentence Alignment

Translation Models: Word-to-Word Alignment

Translation Models: Phrase-to-Phrase Alignment

Translation

2004 JHU-CLSP NIST MT Evaluation Systems

Baseline Systems and Performance

Summary and Conclusion
Automatic Measurement of Translation Quality

Automatic performance metrics have been central to the development of large statistical language processing systems

- Word/Character Error Rate (WER): ASR, OCR, ...
- Precision/Recall: Information Retrieval, Speaker ID, ...
- Crossing Brackets: Parsing, ...

These are all relative to *human performance* over defined test sets

- human performance is measured *once* over a fixed test set
- system performance can be measured *many times* relative to
  - human performance, directly
  - performance of other systems, indirectly

- allows incremental system improvement
- metrics can be incorporated into estimation and decoding

Evaluation metrics should

- be inexpensive to compute
- require little or no ongoing human involvement
Automatic Metrics for Machine Translation

**BLEU** (Papineni 2001) is an MT metric based on n-gram precision

- A single-reference example:

  **Reference** :  mr. speaker, in absolutely no way.
  **Hypothesis** : in absolutely no way, mr. chairman.

  **BLEU Computation**

<table>
<thead>
<tr>
<th>n-gram matches</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-word 2-word 3-word 4-word</td>
<td>((\frac{7}{8} \times \frac{3}{7} \times \frac{2}{6} \times \frac{1}{5})^{\frac{1}{4}} = 0.3976)</td>
</tr>
</tbody>
</table>

  - Can be generalized to multiple references
  - Typically also includes a length penalty
  - Correlates well with human judgments of translation
BLEU Assigns High Scores to Good Translations

Example translations at various levels of translation performance as measured by the sentence-level BLEU score. Examples selected from the NIST 2002 Ch-En MT evaluation set.

<table>
<thead>
<tr>
<th>Translations</th>
<th>BLEU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghan Earthquake Victims begin to rebuild their homes.</td>
<td>66.1</td>
</tr>
<tr>
<td>Prior to this, the ANC has issued a statement calling for the international community to respect the choice of the people and help them survive.</td>
<td>66.0</td>
</tr>
<tr>
<td>Statistics show that since 1992, a total of 204 UN personnel have been killed, but only 15 criminals have been arrested.</td>
<td>64.4</td>
</tr>
<tr>
<td>Chavez emphasized that Venezuela needs peace, stability and reason for all parties should make joint efforts to end the conflict.</td>
<td>62.2</td>
</tr>
<tr>
<td>London Financial Times Index Friday at closing newspaper 5,292.70 points, up 31.30 points.</td>
<td>61.0</td>
</tr>
</tbody>
</table>

readable and fairly plausible
## BLEU Assigns Low Scores to Poor Translations

<table>
<thead>
<tr>
<th>Translation</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japan Telecom company in 2000 to spend 5.5 billion dollars buy back</td>
<td>0.0</td>
</tr>
<tr>
<td>However, the voting result shows that Zhu because there is no reason</td>
<td>0.0</td>
</tr>
<tr>
<td>to be losing power by NPC deputies desolate</td>
<td></td>
</tr>
<tr>
<td>77 private manufacturing enterprises also reported a foreign trade management right</td>
<td>0.0</td>
</tr>
<tr>
<td>Identification Department found that college students of the certificate, many of them were fake</td>
<td>0.0</td>
</tr>
<tr>
<td>The European Union would be implemented in steel imports temporary protective measures to discuss with the Chinese side,</td>
<td>0.0</td>
</tr>
<tr>
<td>Georgia from a section of the great mountains Canyon withdrawal,</td>
<td>0.0</td>
</tr>
</tbody>
</table>

barely readable and probably misleading
BLEU Can Be Used (Carefully) For SMT Development

BLEU is not an absolute measure of translation performance
  ▶ not like Word Error Rate

Improvements in BLEU can be trusted when:
  ▶ many different systems are developed under BLEU, and results throughout development are published on standard test sets
  ▶ periodically validated against human judgments

Current test sets used in the NIST MT evaluations
  ▶ \( \sim 1K \) sentences / \( \sim 25K \) words,
  ▶ four independently produced reference translations

There is extensive effort in improving/ extending/ replacing BLEU
Outline

Introduction

Five Easy Problems in Statistical Machine Translation

Automatic Measurement of Translation Quality

Bitext for SMT Training: Document and Sentence Alignment

Translation Models: Word-to-Word Alignment

Translation Models: Phrase-to-Phrase Alignment

Translation

2004 JHU-CLSP NIST MT Evaluation Systems

Baseline Systems and Performance

Summary and Conclusion
Bitext and Bitext Alignment

**Bitext** is a collection of text in two languages that is known or suspected to contain translations.

**Bitext Alignment**: finds the regions that are translations

- There are **levels** of alignment: document, sentence, or word
- different models and algorithms are used for each level

Some bitexts...

- Hong Kong Hansards in Chinese-English
- Canadian Hansards in French-English
- Europarl corpus
  [http://www3.europarl.eu.int](http://www3.europarl.eu.int)
Document Level Alignment

The definition of a document varies. Depending on the domain, a document could be a book, a news story, a numbered chapter, paragraph, or verse, ...

Parallel Documents are created purposefully by human translators, for example in creating a foreign language edition of a newspaper.

- correspondence between the translations and the original source language documents is maintained
- these are the most valuable, but are expensive to create and to obtain
- large collections of parallel documents are available from LDC

Some interest in searching for Parallel Documents ‘in the wild’ by crawling the web and looking for hints that documents encountered might be translations of each other.

... assume we have parallel documents
Sentence Level Alignment Models

Goal: align sentences across a pair of parallel documents

- English Document: \( e_1 \cdots e_m \)
- French Document: \( f_1 \cdots f_n \)

Two underlying processes
- **Segmentation**: the bitext is *chunked* into \( K \) segments
- **Alignment**: chunks of sentence are aligned across the documents

\( K = 3: \)

\[
E_1E_2E_3E_4E_5E_6 \\
F_1F_2F_3F_4
\]

\( a^K_1: a_1 = (1, 2, 1, 1), a_2 = (3, 5, 2, 3), a_3 = (6, 6, 4, 4) \)

Can describe the alignment of chunks of sentences

\[
P(f^n_1, a, K|e^m_1) = \prod_{k=1}^{K} P^{(w)}(f_k|e_k) \quad P(a^k_1|m, n, k) \quad \beta(K|m, n)
\]

- Translation Model (Coarse)
- Alignment Model
- Chunk Count Model
Sentence Alignment via Monotone Chunk Alignment

\[
\{ \hat{K}, \hat{a}_1^K \} = \arg \max_{K,a^K_1} P(f_1^n, a, K|e_1^m)
\]

DP search: alignment model \( P(a^K_1|m, n, k) \) allows only monotone alignments (Gale & Church '91)

Performance can be good if *if monotone alignment order is adequate*
Sentence Alignment via Divisive Clustering (Y. Deng)
Proceeds from coarse to fine and allows chunk reordering
Non-monotone alignment process $P(a^K_{m,n})$

Since the Korean Peninsula was split into two countries, the Republic of Korea has, while leaning its back on the "big tree" of the United States for security, carefully and consistently sought advanced weapons from the United States in a bid to confront the Democratic People's Republic of Korea. An informed source in Seoul revealed to the Washington Post that the United States had secretly agreed to the request of South Korea earlier this year to "extend its existing missile range" to strike Pyongyang direct. This should have elated South Korea. But since the situation surrounding the peninsula has changed dramatically and the two heads of state of the two Koreas have met with each other and signed a joint statement, what should South Korea do now? It has no choice but spit back the "greasy meat" from its mouth and put the "missile expansion plan" on the back burner. A knowledgeable South Korean speaks the truth: "Because of the summit meeting, we have shelved our own missile plan. If we go ahead with it, it will spoil the excellent situation opened up by the summit meeting."
Sentence Alignment via Divisive Clustering (Y. Deng)
Proceeds from coarse to fine and allows chunk reordering.
At each iteration, the single most likely splitting point is chosen.

Since the Korean Peninsula was split into two countries, the Republic of Korea has, while leaning its back on the "big tree" of the United States for security, carefully and consistently sought advanced weapons from the United States in a bid to confront the Democratic People's Republic of Korea. An informed source in Seoul revealed to the Washington Post that the United States had secretly agreed to the request of South Korea earlier this year to "extend its existing missile range" to strike Pyongyang direct. This should have elated South Korea. But since the situation surrounding the peninsula has changed dramatically and the two heads of state of the two Koreas have met with each other and signed a joint statement, what should South Korea do now? It has no choice but spit back the "greasy meat" from its mouth and put the "missile expansion plan" on the back burner. A knowledgeable South Korean speaks the truth:

"Because of the summit meeting, we have shelved our own missile plan. If we go ahead with it, it will spoil the excellent situation opened up by the summit meeting.
Sentence Alignment via Divisive Clustering (Y. Deng)

Proceeds from **coarse** to **fine** and allows chunk reordering.

At each iteration, the single most likely splitting point is chosen.

<table>
<thead>
<tr>
<th><strong>At the Summit Meeting</strong></th>
<th><strong>Proceedings</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>自从 朝鲜半岛被分裂成两个国家以来，韩国在背靠美国这棵大树以求自安的同时，还小心翼翼但却坚持不懈地向美国寻求先进武器，以至于抗衡朝鲜。据汉城的消息灵通人士向《华盛顿邮报》透露，今年早些时候，美国已秘密而不动声地同意韩国“可以扩展其现有导弹的射程”，使之能够直捣朝鲜首都平壤。这本应是韩国感到欣喜的事件，可眼下半岛局势有了重大变化，朝韩首脑面对面地会了晤，并签署了联合声明。韩国怎么办？只好把到嘴的“肥肉”先吐出来，搁置自己的“导弹射程扩展计划”。</td>
<td></td>
</tr>
<tr>
<td>自从朝鲜半岛被分裂成两个国家以来，韩国在背靠美国这棵大树以求自安的同时，还小心翼翼但却坚持不懈地向美国寻求先进武器，以至于抗衡朝鲜。据汉城的消息灵通人士向《华盛顿邮报》透露，今年早些时候，美国已秘密而不动声地同意韩国“可以扩展其现有导弹的射程”，使之能够直捣朝鲜首都平壤。这本应是韩国感到欣喜的事件，可眼下半岛局势有了重大变化，朝韩首脑面对面地会了晤，并签署了联合声明。韩国怎么办？只好把到嘴的“肥肉”先吐出来，搁置自己的“导弹射程扩展计划”。</td>
<td></td>
</tr>
<tr>
<td>Since the Korean Peninsula was split into two countries, the Republic of Korea has, while leaning its back on the &quot;big tree&quot; of the United States for security, carefully and consistently sought advanced weapons from the United States in a bid to confront the Democratic People’s Republic of Korea. An informed source in Seoul revealed to the Washington Post that the United States had secretly agreed to the request of South Korea earlier this year to &quot;extend its existing missile range&quot; to strike Pyongyang directly. This should have elated South Korea. But since the situation surrounding the peninsula has changed dramatically and the two heads of state of the two Koreas have met with each other and signed a joint statement, what should South Korea do now? It has no choice but spit back the &quot;greasy meat&quot; from its mouth and put the &quot;missile expansion plan&quot; on the back burner.</td>
<td></td>
</tr>
<tr>
<td>Since the Korean Peninsula was split into two countries, the Republic of Korea has, while leaning its back on the &quot;big tree&quot; of the United States for security, carefully and consistently sought advanced weapons from the United States in a bid to confront the Democratic People’s Republic of Korea. An informed source in Seoul revealed to the Washington Post that the United States had secretly agreed to the request of South Korea earlier this year to &quot;extend its existing missile range&quot; to strike Pyongyang directly. This should have elated South Korea. But since the situation surrounding the peninsula has changed dramatically and the two heads of state of the two Koreas have met with each other and signed a joint statement, what should South Korea do now? It has no choice but spit back the &quot;greasy meat&quot; from its mouth and put the &quot;missile expansion plan&quot; on the back burner.</td>
<td></td>
</tr>
<tr>
<td>One knowledgeable South Korean speaks the truth:</td>
<td></td>
</tr>
<tr>
<td>&quot;Because of the summit meeting, we have shelved our own missile plan. If we go ahead with it, it will spoil the excellent situation opened up by the summit meeting.&quot;</td>
<td></td>
</tr>
</tbody>
</table>
Sentence Alignment via Divisive Clustering (Y. Deng)
Proceeds from coarse to fine and allows chunk reordering
At each iteration, the single most likely splitting point is chosen.

Since the Korean Peninsula was split into two countries, the Republic of Korea has, while leaning its back on the "big tree" of the United States for security, carefully and consistently sought advanced weapons from the United States in a bid to confront the Democratic People's Republic of Korea.

An informed source in Seoul revealed to the Washington Post that the United States had secretly agreed to the request of South Korea earlier this year to "extend its existing missile range" to strike Pyongyang directly. This should have elated South Korea. But since the situation surrounding the peninsula has changed dramatically and the two heads of state of the two Koreas have met with each other and signed a joint statement, what should South Korea do now? It has no choice but spit back the "greasy meat" from its mouth and put the "missile expansion plan" on the back burner.

A knowledgeable South Korean speaks the truth:
"Because of the summit meeting, we have shelved our own missile plan. If we go ahead with it, it will spoil the excellent situation opened up by the summit meeting.
Sentence Alignment via Divisive Clustering (Y. Deng)

Proceeds from coarse to fine and allows chunk reordering.

At each iteration, the single most likely splitting point is chosen.

Since the Korean Peninsula was split into two countries, the Republic of Korea has, while leaning its back on the "big tree" of the United States for security, carefully and consistently sought advanced weapons from the United States in a bid to confront the Democratic People's Republic of Korea.

An informed source in Seoul revealed to the Washington Post that the United States had secretly agreed to the request of South Korea earlier this year to "extend its existing missile range" to strike Pyongyang directly.

This should have elated South Korea. But since the situation surrounding the peninsula has changed dramatically and the two heads of state of the two Koreas have met with each other and signed a joint statement, what should South Korea do now?

A knowledgeable South Korean speaks the truth:

"Because of the summit meeting, we have shelved our own missile plan. If we go ahead with it, it will spoil the excellent situation opened up by the summit meeting."

Introduction

Five Easy Problems in Statistical Machine Translation

2004 JHU-CLSP NIST MT Evaluation Systems

Summary and Conclusion
Bitext Alignment Goal: Better Translation Models

Benefits of good chunking

- better training set alignment improves translation performance
- smaller chunk pairs leads to faster translation model training

A good alignment algorithm should be ...

- fast, so multiple iterations are possible
- efficient: as little bitext should be discarded as possible
- initialized from a flat-start
- language independent
- require minimal linguistic knowledge
- able to work at the subsentence level
  - coarse monotonic alignment followed by fine divisive clustering works well
  - splitting points can depend on the language pairs
Outline

Introduction

Five Easy Problems in Statistical Machine Translation

Automatic Measurement of Translation Quality
Bitext for SMT Training: Document and Sentence Alignment
Translation Models: Word-to-Word Alignment
Translation Models: Phrase-to-Phrase Alignment
Translation

2004 JHU-CLSP NIST MT Evaluation Systems
Baseline Systems and Performance

Summary and Conclusion
Word Alignment

An English-French Sentence Pair: $(e_1^l, f_1^l)$

NULL Mr. Speaker, my question is directed to the Minister of Transport

Monsieur le Orateur, ma question se adresse à le ministre chargé de les transports

- **Alignment Links**: $b = (i, j) : f_j$ linked to $e_i$
- **Alignment is defined by a Link Set** $B = \{b_1, b_2, ..., b_J\}$
- Some links are NULL links
- Introduce a word alignment process $a_j$

\[
 b = (i, j) \implies a_j = i \implies f_j \leftrightarrow e_{a_j}
\]
IBM Model 1 & 2 and HMM Word Alignment Models

Translation has ‘direction’, e.g. English generates French

- any English word can generate any French word

\[
P(f, a, m | e) = P(f | a, m, e) \quad P(a | m, e) \quad P(m | e) \\
= \prod_{j=1}^{m} P_t(f_j | e_{a_j}) \quad \prod_{j=1}^{m} P_a(a_j | j, l, m) \quad P_e(m | l)
\]

- EM is possible

\[
P(f|e) = \sum_a P(f, a|e) = P_e(m|l) \prod_{j=1}^{m} \sum_{i=0}^{l} P_t(f_j|e_i) P_a(i|j, l, m)
\]

- HMM Alignment Model - EM is possible here, too

\[
P(a|m, e) = \prod_{j=1}^{m} P(a_j|a_{j-1}, l, m)
\]
IBM Model 4 for Word Alignment

Model 4 is a powerful model of word alignment within sentence pairs

Features:

- lexical model, as in Model 1, and an alignment model
- **Fertility**: probability distribution over the number of French words each English word can generate
  - an approximation to **phrase modeling**
- **NULL Translation Model**: allows French words to be generated without a corresponding source on the English side
- **Distortion Model**: describes how French words are distributed throughout the French sentence when generated from a single English word

Neither estimation nor decoding is easy under Model 4 -

- \( \sum_a P(f, a|e) \) and \( \max_a P(f, a|e) \) are not easy to compute
- **no EM, no parallel training**
Word-by-Word Translation Does Have Limitations

A reasonable example:

NULL Mr. Speaker, my question is directed to the Minister of Transport

Monsieur le Orateur, ma question se adresse à le ministre chargé de les transports

A more troublesome example:

English: ... you’re just pulling my leg ...
Italian: ... mi stai prendeno per il naso ...

Question: When should leg be translated as nose?
Answer: Whenever the bitext says so.
Phrase-to-Phrase Translation Models

Alignment Template Models (Och et al. 1999)

- derived from ‘good’ word-level alignments, typically from IBM-4

\[
\text{ce bill met de le baume sur une blessure}
\]

\[
\text{this bill places salve on a sore wound}
\]

- Phrase Pairs are extracted to cover patterns of word alignments found in the training bitext

- Probability distributions are defined over phrase pair sequences
Translation is ‘trivial’ once the component distributions are defined.

\[ \hat{e} = \arg\max_{e} P(e|f) \]

One approach is to construct specialized search algorithms

- Depending on the underlying models, search can be by Viterbi, A*, and other specialized search procedures

But decoder design and implementation is complex

- Small model changes might require large changes to a decoder
- Any approximations made during search lead to inexact implementation of the model
- Decoder implementation takes effort away from ‘modeling’
Translation via Weighted Finite State Transducers

*Translation with Finite State Devices*, Knight & Al Onaizan, AMTA’98

- Implements the IBM models as WFSTs
  - word-to-word translation, word fertility, and permutation (reordering)

If the component models can be implemented as WFSTs which can be composed, building a decoder is trivial

- Can be limiting, but avoids special-purpose decoders
- The value of this modeling approach has been shown in ASR by the systems developed at AT&T
  - Translation is performed using libraries of standard FSM operations
  - Clear formulation
Translation Template Model - TTM

Generative source-channel model of machine translation

- Takes the best of
  - Och&Ney's Phrase-based translation models
  - Knight&Al Onaizan WFST description of translation via IBM models
TTM Component Distributions

- Transformations via stochastic models implemented as WFSTs
- Implementation is direct using standard WFST operations
Translation Template Model Features

- Bitext word alignment and translation under the model can be performed using standard WFST operations
  - Modular Implementation
  - No necessity for a specialized decoder
  - Can easily generate translation lattices and N-best lists
- Should be easy to translate ASR lattices
Outline

Introduction

Five Easy Problems in Statistical Machine Translation
Automatic Measurement of Translation Quality
Bitext for SMT Training: Document and Sentence Alignment
Translation Models: Word-to-Word Alignment
Translation Models: Phrase-to-Phrase Alignment
Translation

2004 JHU-CLSP NIST MT Evaluation Systems
Baseline Systems and Performance

Summary and Conclusion
Training and Translation Procedures

1. Bitext Chunking
   1.1 Monotone alignment into coarse chunks of documents
   1.2 Divisive clustering into subsentence chunks

2. Partition the bitext into training sets of manageable size

3. For each partition
   For each translation direction
   Train the following models with Giza++:
   3.1 IBM Model 1
   3.2 HMM word alignment model
   3.3 IBM Model 4

4. Merge word aligned bitexts and extract phrase pairs

5. Construct component WFSTs for the TTM

6. Translation lattice generation with pruned trigram

7. Translation lattice rescoring with unpruned 4gram

8. Minimum Bayes Risk rescoring under BLEU
Text Processing & System Building Strategy

**Preliminary Analysis:** Every model component suffered data sparcity

**Goal:** Exploit all available text resources

- **Chinese Text** segmented into words using LDC segmenter (Linguistic Data Consortium)
- **Arabic Text** processing pipeline
  - Modified Buckwalter analyzer
  - split conjunctions, prepositions, Al- and pronouns
  - Al- and w- deletion (maybe wrong decision!)
- **English Text** processed using a simple tokenizer

**Bitext for Translation Model Training**

<table>
<thead>
<tr>
<th></th>
<th>Chinese-English</th>
<th>Arabic-English</th>
</tr>
</thead>
<tbody>
<tr>
<td># of sent. pairs (K)</td>
<td>-</td>
<td>68.0</td>
</tr>
<tr>
<td># of chunk pairs (M)</td>
<td>7.6</td>
<td>5.1</td>
</tr>
<tr>
<td># of words (M)</td>
<td>175.7/207.4</td>
<td>123.0/132.5</td>
</tr>
</tbody>
</table>
## English Language Model Training Data

By source - in Millions of words

<table>
<thead>
<tr>
<th>Source</th>
<th>Xin</th>
<th>AFP</th>
<th>PD</th>
<th>FBIS</th>
<th>UN</th>
<th>AR-news</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small 3g</td>
<td>4.3</td>
<td>-</td>
<td>16.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>20.5</td>
</tr>
<tr>
<td>Big 3g</td>
<td>155.7</td>
<td>200.8</td>
<td>16.2</td>
<td>10.5</td>
<td>-</td>
<td>-</td>
<td>373.3</td>
</tr>
<tr>
<td>Big 4g</td>
<td>155.7</td>
<td>200.8</td>
<td>16.2</td>
<td>10.5</td>
<td>-</td>
<td>-</td>
<td>373.3</td>
</tr>
<tr>
<td>A-E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small 3g</td>
<td>63.1</td>
<td>200.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.1</td>
<td>266.0</td>
</tr>
<tr>
<td>Big 3g</td>
<td>83.0</td>
<td>210.0</td>
<td>-</td>
<td>-</td>
<td>131.0</td>
<td>3.6</td>
<td>428.0</td>
</tr>
<tr>
<td>Big 4g</td>
<td>83.0</td>
<td>210.0</td>
<td>-</td>
<td>-</td>
<td>131.0</td>
<td>3.6</td>
<td>428.0</td>
</tr>
</tbody>
</table>

Available from LDC:

- Xinhua, Agence France Press, People’s Daily, FBIS, United Nations collections, AR-news
Translation Performance : MT Bitext Size

### Chinese-English - Decode with Small3g LM

<table>
<thead>
<tr>
<th>Bitext Partition</th>
<th>Contribution by Source: En words (M)</th>
<th>BLEU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBIS</td>
<td>HKNews</td>
<td>XHTS</td>
</tr>
<tr>
<td>1</td>
<td>10.5</td>
<td>16.3</td>
</tr>
<tr>
<td>2</td>
<td>10.5</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>10.5</td>
<td>16.3</td>
</tr>
<tr>
<td>1+2+3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(XHTS = Xinhua, Hansards, Treebank, Sinorama)

### Arabic-English - Decode with Small3g LM

<table>
<thead>
<tr>
<th>Bitext Partition</th>
<th>Contribution by Source: En words (M)</th>
<th>BLEU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN News</td>
<td>Total</td>
<td>eval02</td>
</tr>
<tr>
<td>1</td>
<td>65.0</td>
<td>3.5</td>
</tr>
<tr>
<td>2</td>
<td>64.0</td>
<td>3.5</td>
</tr>
<tr>
<td>1+2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Vanilla Translation System Performance

<table>
<thead>
<tr>
<th>System #</th>
<th>Decoding Method</th>
<th>BLEU%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Chinese-English</td>
</tr>
<tr>
<td></td>
<td></td>
<td>e01  e02  e03  e04 (c)</td>
</tr>
<tr>
<td>1</td>
<td>JHU-UMD ’03</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>AE-Primary ’04</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>Small3g</td>
<td>28.1</td>
</tr>
<tr>
<td>4</td>
<td>Big3g</td>
<td>28.7</td>
</tr>
<tr>
<td>5</td>
<td>Big4g</td>
<td>29.6</td>
</tr>
<tr>
<td>6</td>
<td>Nbest from 4</td>
<td>29.7</td>
</tr>
</tbody>
</table>

- Eval04 results are case-sensitive BLEU
  - Truecasing was performed using HMM capitalizer
- good improvements from 2003 to 2004
- lattice rescoring framework works well
- translation benefits if the system can support long-span LMs
Outline

Introduction

Five Easy Problems in Statistical Machine Translation
Automatic Measurement of Translation Quality
Bitext for SMT Training: Document and Sentence Alignment
Translation Models: Word-to-Word Alignment
Translation Models: Phrase-to-Phrase Alignment
Translation

2004 JHU-CLSP NIST MT Evaluation Systems
Baseline Systems and Performance

Summary and Conclusion
Summary of SMT Evaluation System Development

- An SMT system based on Bitext Chunking and TTM (and Giza++)
  - WFST architecture supports generation and rescoring of lattices
- Similar architectures for C-E and A-E systems
  - C-E system developed in 1 month - by 1 Indian graduate student
  - A-E system developed in 10 days - by 1 Chinese graduate student
  - Bitext chunking thoroughly exploits almost all available bitext
  - Translation tuned to get the best LM contribution
- Large gains relative to JHU-CLSP 2003 eval system
- Respectable performance in Chinese-English and Arabic-English MT

Upcoming Evaluations
- NIST C-E and A-E evaluations – May ’05
- TC-STAR C-E Speech-to-Text – Spring ’05
Summary - Current Research in SMT

Have discussed five problems in statistical machine translation
- These are not necessarily independent research topics
  ▶ **Divisive Clustering** integrates ‘text preprocessing’ issues – sentence and document alignment – into overall MT system training
  ▶ **Minimum Bayes Risk Alignment and Translation** integrates Translation Performance Metrics into search algorithms
  ▶ **WFST Translation Models** blur the distinction between translation models and translation algorithms – if the models are constructed appropriately, translation is ‘straightforward’

Overall Objectives: efficient, exact models and aggressive use of bitext
Word Alignment Models - New!

Goal: Replace IBM Model 4 by ‘simpler’ alignment models

- careful and exact model formulation
- equal alignment and translation performance to Model 4
- efficient training via EM - parallelization: 3 days on 60 CPUs vs 2 weeks on 3 CPUs
- a single set of models over all available bitext - avoid partitioning the bitext training data
- estimate phrase pairs under the model to improve test set coverage
- direct use of models in translation - support discriminative training, adaptation, etc.
The Machine Translation Pyramid

Casts the problem in familiar terms
- strengths and weaknesses of the formulation are obvious
The Statistical Machine Translation Pyramid

- Simple Models & Fast Algorithms
- Evidence
- Linguistic Knowledge
- Explanation
- Machine Learning
- Elegance

Introduction
Five Easy Problems in Statistical Machine Translation
2004 JHU-CLSP NIST MT Evaluation Systems
Summary and Conclusion
Martin Luther versus The Lemmings

Need to be aware of ongoing arguments about translation:

- Translation is impossible.
- Martin Luther: “If I followed those lemmings the literalists ...”
  Should a translation be faithful or should it be accessible?
  - domesticating vs. foreignising, modern vs. archaic, transliteration, ...
- Translation is not annotation (or maybe for SMT it could be)
- What’s the relationship between translation and interpretation?
  - generation followed by translation vs. simultaneous generation
- ...

One’s position often depend on the intended application

SMT assumes people have found workable solutions to these problems
  - it’s enough simply to mimic them (or not ...)


Thank you!