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

- 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

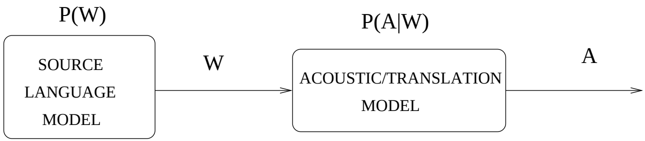
- Baseline Systems and Performance

Summary and Conclusion

Automatic Speech Recognition and Machine Translation

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

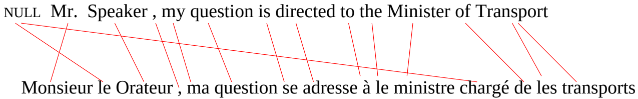
- Readable source language text is ‘corrupted’ into some other form



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.

$$P(e, a, f) = \underbrace{P(f|a, e)}_{\text{Translation Model}} \; \underbrace{P(a|e)}_{\text{Alignment Model}} \; \underbrace{P(e)}_{\text{Language Model}}$$

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

$$(\,\widehat{e}, \widehat{a}\,) = \operatorname*{argmax}_{e,a} \; P(f|a, e) \; P(a|e) \; P(e)$$

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
- 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

n-gram matches				BLEU
1-word	2-word	3-word	4-word	$\left(\frac{7}{8} \times \frac{3}{7} \times \frac{2}{6} \times \frac{1}{5}\right)^{\frac{1}{4}} = 0.3976$
7/8	3/7	2/6	1/5	

- 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.

Translations	BLEU (%)
	60 – 70%
Afghan Earthquake Victims begin to rebuild their homes .	66.1
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 .	66.0
Statistics show that since 1992 , a total of 204 UN personnel have been killed , but only 15 criminals have been arrested .	64.4
Chavez emphasized that Venezuela needs peace , stability and reason for all parties should make joint efforts to end the conflict .	62.2
London Financial Times Index Friday at closing newspaper 5,292.70 points , up 31.30 points .	61.0

readable and fairly plausible

BLEU Assigns Low Scores to Poor Translations

	$0 - 10\%$
Japan Telecom company in 2000 to spend 5.5 billion dollars buy back .	0.0
However , the voting result shows that Zhu because there is no reason to be losing power by NPC deputies desolate .	0.0
77 private manufacturing enterprises also reported a foreign trade management right .	0.0
Identification Department found that college students of the certificate , many of them were fake .	0.0
The European Union would be implemented in steel imports temporary protective measures to discuss with the Chinese side ,	0.0
Georgia from a section of the great mountains Canyon withdrawal ,	0.0

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
- 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

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

$$\begin{array}{rl} \text{English Document :} & \mathbf{e}_1 \cdots \mathbf{e}_m \\ \text{French Document :} & \mathbf{f}_1 \cdots \mathbf{f}_n \end{array}$$

Two underlying processes

- Segmentation : the bitext is *chunked* into K segments
- Alignment : chunks of sentence are aligned across the documents

$$K = 3: \qquad \begin{array}{ccc} \mathbf{e}_1\mathbf{e}_2 & \mathbf{e}_3\mathbf{e}_4\mathbf{e}_5 & \mathbf{e}_6 \\ \mathbf{f}_1 & \mathbf{f}_2\mathbf{f}_3 & \mathbf{f}_4 \end{array}$$

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

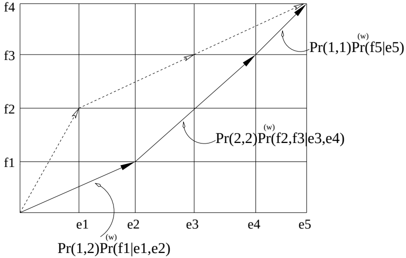
Can describe the alignment of chunks of sentences

$$P(\mathbf{f}_1^n, \mathbf{a}, K | \mathbf{e}_1^{\mathbf{m}}) = \quad \underset{\substack{\text{Translation} \\ \text{Model (Coarse)}}}{\prod_{k=1}^K P^{(w)}(f_{a_k}|e_{a_k})} \quad \underset{\substack{\text{Alignment} \\ \text{Model}}}{P(a_1^k|m,n,k)} \quad \underset{\substack{\text{Chunk Count} \\ \text{Model}}}{\beta(K|m,n)}$$

Sentence Alignment via Monotone Chunk Alignment

$$\{\hat{K}, \hat{a}_1^{\hat{K}}\} = \operatorname*{argmax}_{K, a_1^K} P(\mathbf{f}_1^n, \mathbf{a}, K | \mathbf{e_1^m})$$

DP search: alignment model $P(a_1^k | m, n, k)$ allows only monotone alignments (Gale & Church '91)

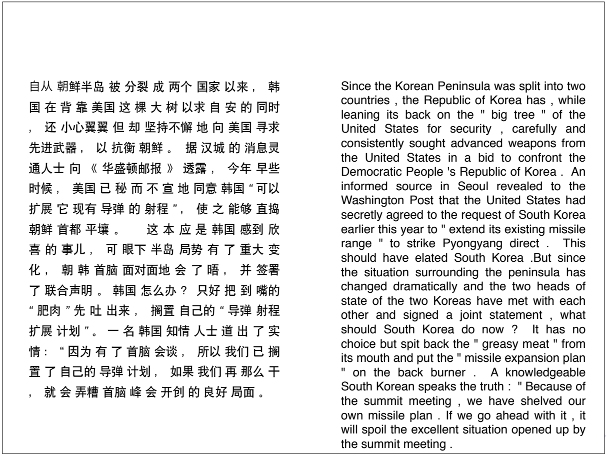

f4
f3
f2
f1
e1
e2
e3
e4
e5
$\Pr(1,1)\overset{(w)}{\Pr}(f5|e5)$
$\Pr(2,2)\overset{(w)}{\Pr}(f2,f3|e3,e4)$
$\Pr(1,2)\overset{(w)}{\Pr}(f1|e1,e2)$

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 $P(a_1^K | m, n)$



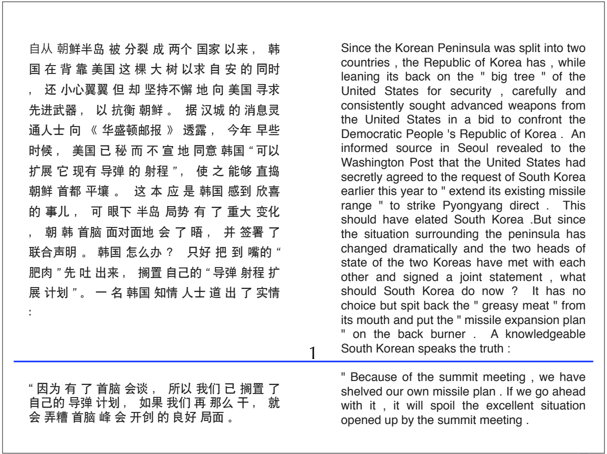
自从 朝鲜半岛 被 分裂 成 两个 国家 以来， 韩国 在 背 靠 美国 这 棵 大 树 以求 自 安 的 同时 ， 还 小心翼翼 但 却 坚持不懈 地 向 美国 寻求 先进武器， 以 抗衡 朝鲜 。 据 汉城 的 消息灵通人士 向 《 华盛顿邮报 》 透露， 今年 早些 时候， 美国 已 秘 而 不 宣 地 同意 韩国 “ 可以 扩展 它 现有 导弹 的 射程 ”， 使 之 能够 直捣 朝鲜 首都 平壤 。 这 本 应 是 韩国 感到 欣喜 的 事儿， 可 眼下 半岛 局势 有 了 重大 变化， 朝 韩 首脑 面对面地 会 了 晤， 并 签署 了 联合声明 。 韩国 怎么办？ 只好 把 到 嘴的 “ 肥肉 ” 先 吐 出来， 搁置 自己的 “ 导弹 射程 扩展 计划 ”。 一 名 韩国 知情 人士 道 出 了 实情： “ 因为 有 了 首脑 会谈， 所以 我们 已 搁置 了 自己的 导弹 计划， 如果 我们 再 那么 干 ， 就 会 弄糟 首脑 峰 会 开创 的 良好 局面 。

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.



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 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 而 不 宣 地 同意 韩国 “ 可以 扩展 它 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 事儿， 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 ”。 一 名 韩国 知情 人士 道 出 了 实情 :

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 :

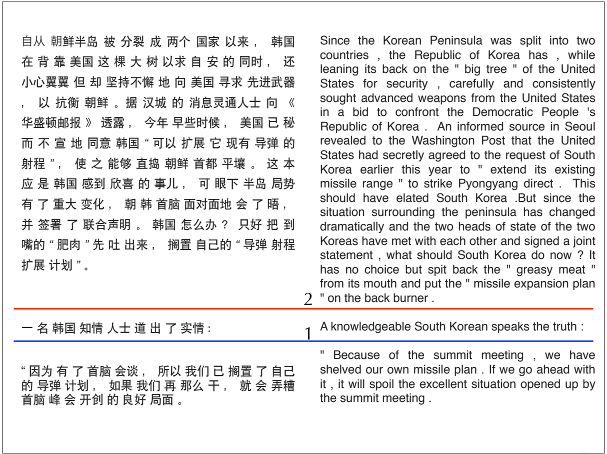
“ 因为 有 了 首脑 会谈， 所以 我们 已 搁置 了 自己的 导弹 计划， 如果 我们 再 那么 干， 就 会 弄糟 首脑 峰 会 开创 的 良好 局面 。

" 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.



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 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 而 不 宣 地 同意 韩国 “ 可以 扩展 它 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 事儿 ， 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 ” 。

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 .

2

一 名 韩国 知情 人士 道 出 了 实情 :

A knowledgeable South Korean speaks the truth :

1

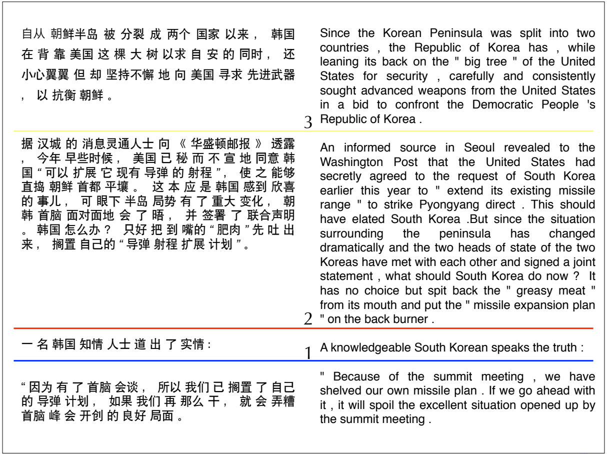
“ 因为 有 了 首脑 会谈 ， 所以 我们 已 搁置 了 自己 的 导弹 计划 ， 如果 我们 再 那么 干 ， 就 会 弄糟 首脑 峰 会 开创 的 良好 局面 。

" 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.

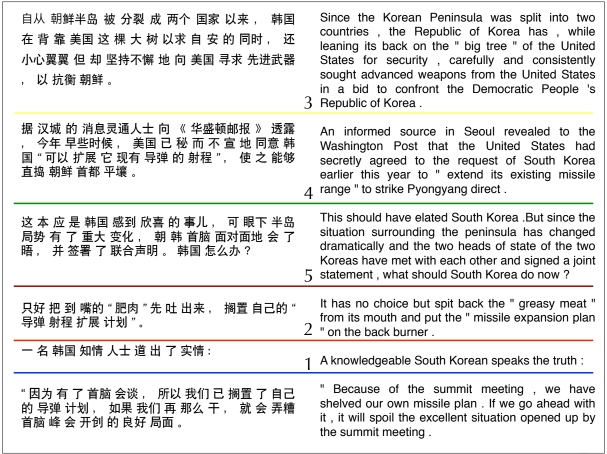


自从 朝鲜半岛 被 分裂 成 两个 国家 以来， 韩国 在 背 靠 美国 这 棵 大 树 以求 自 安 的 同时， 还 小心翼翼 但 却 坚持不懈 地 向 美国 寻求 先进武器 ， 以 抗衡 朝鲜 。	3	Since the Korean Peninsula was split into two countries , the Republic of Korea has , while leaning its back on the " big tree " of the United States for security , carefully and consistently sought advanced weapons from the United States in a bid to confront the Democratic People 's Republic of Korea .
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 而 不 宣 地 同意 韩国 “ 可以 扩展 它 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 事儿， 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 ”。	2	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 .
一 名 韩国 知情 人士 道 出 了 实情：	1	A knowledgeable South Korean speaks the truth :
“ 因为 有 了 首脑 会谈， 所以 我们 已 搁置 了 自己 的 导弹 计划， 如果 我们 再 那么 干， 就 会 弄糟 首脑 峰 会 开创 的 良好 局面 。		" 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.



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 抗衡 朝鲜 。	3	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 .
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 而 不 宣 地 同意 韩国 “ 可以 扩展 它 现有 导弹 的 射程 ” ， 使 之 能够 直捣 朝鲜 首都 平壤 。	4	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 .
这 本 应 是 韩国 感到 欣喜 的 事儿 ， 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 ？	5	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 ?
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 ” 。	2	It has no choice but spit back the " greasy meat " from its mouth and put the " missile expansion plan " on the back burner .
一 名 韩国 知情 人士 道 出 了 实情 :	1	A knowledgeable South Korean speaks the truth :
“ 因为 有 了 首脑 会谈 ， 所以 我们 已 搁置 了 自己 的 导弹 计划 ， 如果 我们 再 那么 干 ， 就 会 弄糟 首脑 峰 会 开创 的 良好 局面 。		" 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 .

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
- Introduction

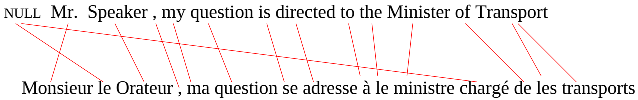
Five Easy Problems in Statistical Machine Translation
- Automatic Measurement of Translation Quality
- Bitext for SMT Training: Document and Sentence Alignment
- Translation Models: Word-to-Word Alignment
- Translation Models: Phrase-to-Phrase Alignment
- Translation

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

Summary and Conclusion

Word Alignment

An English-French Sentence Pair: (e_1^I, f_1^J)



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

$$\begin{array}{rcccc} P(\mathbf{f},\mathbf{a},m\,|\,\mathbf{e}) & = & P(\mathbf{f}\,|\,\mathbf{a},m,\mathbf{e}) & P(\mathbf{a}\,|\,m,\mathbf{e}) & P(m\,|\,\mathbf{e}) \\ & = & \prod_{j=1}^{m} P_t(f_j\,|\,e_{a_j}) & \prod_{j=1}^{m} P_a(a_j\,|\,j,l,m) & P_\epsilon(m\,|\,l) \\ & & \text{Word Translation} & \text{Word Alignment} & \text{Length} \\ & & \text{(table lookup)} & \text{(flat in Model 1)} & \end{array}$$

- EM is possible

$$P(\mathbf{f}|e) = \sum_{\mathbf{a}} P(\mathbf{f},\mathbf{a}|\mathbf{e}) = P_\epsilon(m|l)\prod_{j=1}^{m}\sum_{i=0}^{l} P_t(f_j|e_i)\,P_a(i|j,l,m)$$

- HMM Alignment Model - EM is possible here, too

$$P(\mathbf{a}|m,\mathbf{e}) = \prod_{j=1}^{m} P(a_j|a_{j-1},l,m)$$

IBM Model 4 for Word Alignment

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

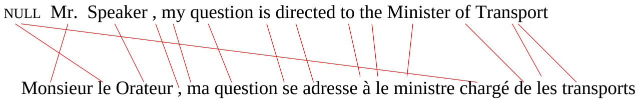
- lexical model, as in Model 1, and an alignment model
- Fertility: probability distribution over the number of French words each English word can generate
    - an approximation to *phrase modeling*
- NULL Translation Model: allows French words to be generated without a corresponding source on the English 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 -

- $\sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a} | \mathbf{e})$ and $\max_{\mathbf{a}} P(\mathbf{f}, \mathbf{a} | \mathbf{e})$ 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:

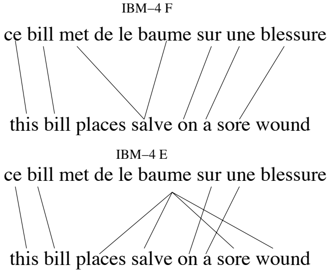
English: *... you're just pulling my leg ...*
Italian: *... mi stai prendeno per il naso ...*

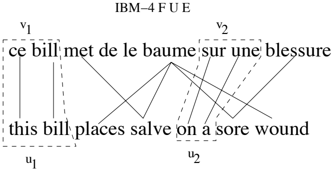
Question: When should leg be translated as nose ?
Answer: Whenever the bitext says so.

Phrase-to-Phrase Translation Models

Alignment Template Models (Och et al. 1999)

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


IBM–4 F
ce bill met de le baume sur une blessure
this bill places salve on a sore wound
IBM–4 E
ce bill met de le baume sur une blessure
this bill places salve on a sore wound


IBM–4 F U E
v1
v2
ce bill met de le baume sur une blessure
this bill places salve on a sore wound
u1
u2

- Phrase Pairs are extracted to cover patterns of word alignments found in the training bitext
- Probability distributions are defined over phrase pair sequences

Translation

Translation is ‘trivial’ once the component distributions are defined.

Its just search ... $\hat{\mathbf{e}} = \mathrm{argmax}_{\mathbf{e}} P(\mathbf{e}|\mathbf{f})$

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

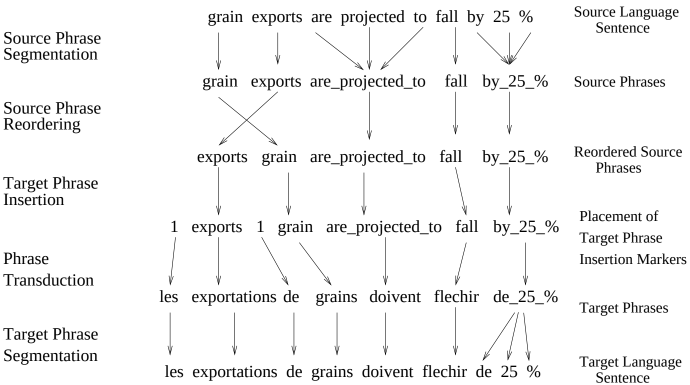
- Can be limiting, but avoids special-purpose decoders
- The value of this modeling approach has been shown in ASR by the systems developed at AT&T
  - Translation is performed using libraries of standard FSM operations
  - Clear formulation

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


grain exports are projected to fall by 25 %
Source Language Sentence
Source Phrase Segmentation
grain exports are_projected_to fall by_25_%
Source Phrases
Source Phrase Reordering
exports grain are_projected_to fall by_25_%
Reordered Source Phrases
Target Phrase Insertion
1 exports 1 grain are_projected_to fall by_25_%
Placement of Target Phrase Insertion Markers
Phrase Transduction
les exportations de grains doivent flechir de_25_%
Target Phrases
Target Phrase Segmentation
les exportations de grains doivent flechir de 25 %
Target Language Sentence

- 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

- 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

	Chinese-English	Arabic-English
# of sent. pairs (K)	-	68.0
# of chunk pairs (M)	7.6	5.1
# of words (M)	175.7/207.4	123.0/132.5

English Language Model Training Data

By source - in Millions of words

Source	Xin	AFP	PD	FBIS	UN	AR-news	Total
C-E							
Small 3g	4.3	-	16.2	-	-	-	20.5
Big 3g	155.7	200.8	16.2	10.5	-	-	373.3
Big 4g	155.7	200.8	16.2	10.5	-	-	373.3
A-E							
Small 3g	63.1	200.8	-	-	-	2.1	266.0
Big 3g	83.0	210.0	-	-	131.0	3.6	428.0
Big 4g	83.0	210.0	-	-	131.0	3.6	428.0

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

Bitext	Contribution by Source: En words (M)					BLEU (%)	
Partition	FBIS	HKNews	XHTS	UN	Total	eval02	eval03
1	10.5	16.3	-	26.7	53.5	25.5	24.1
2	10.5	-	-	85.0	95.5	25.9	25.0
3	10.5	16.3	43.0	25.8	95.6	25.8	24.9
1+2+3						26.6	25.5

(XHTS = Xinhua, Hansards, Treebank, Sinorama)

Arabic-English - Decode with Small3g LM

Bitext	Contribution by Source: En words (M)			BLEU (%)	
Partition	UN	News	Total	eval02	eval03
1	65.0	3.5	68.5	35.0	37.4
2	64.0	3.5	67.5	35.7	37.6
1+2				35.8	37.8

Vanilla Translation System Performance

System #		Decoding Method	BLEU%						
			Chinese-English				Arabic-English		
			e01	e02	e03	e04 (c)	e02	e03	e04 (c)
1	JHU-UMD '03	-	-	16.0	17.0	-	-	-	-
2	AE-Primary '04	-	-	-	-	-	-	38.0	30.6
3	Small3g		28.1	26.6	25.5	-	35.8	37.8	-
4	Big3g		28.7	27.7	27.1	26.5	38.1	40.1	-
5	Big4g	Nbest from 4	29.6	28.2	27.3	27.5	-	-	-
6		Lattice from 4	29.7	28.5	27.4	27.6	39.4	42.1	36.1

- 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

- 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

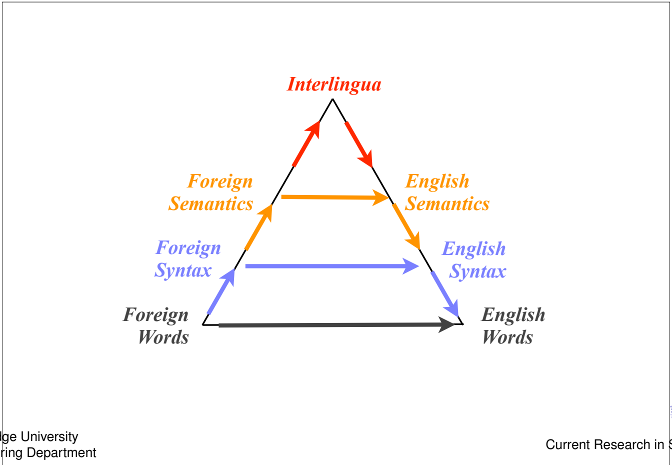
Word Alignment Models - New !

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

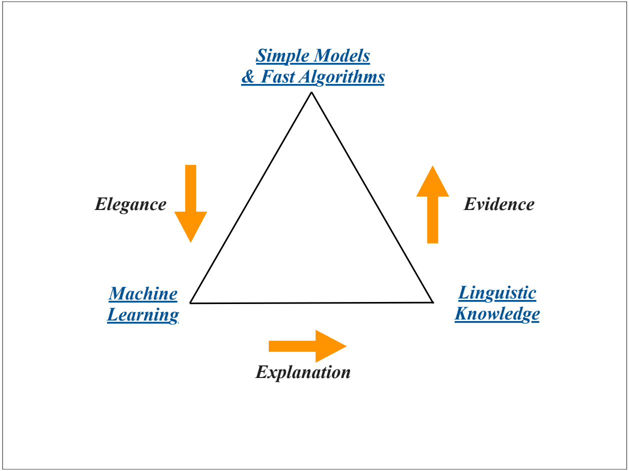
- careful and exact model formulation
- equal alignment and translation performance to Model 4
- efficient training via EM -
  parallelization: 3 days on 60 CPUs vs 2 weeks on 3 CPUs
- a single set of models over all available bitext -
  avoid partitioning the bitext training data
- estimate phrase pairs 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



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 ...)

- Bangalore, S. and Riccardi, G. (2001) A finite-state approach to machine translation. *Proceedings of the 2nd meeting of the North American Chapter of the Association for Computational Linguistics*. Pittsburgh, PA, USA.
- Brown, P. F., Cocke, J., Della Pietra, S. A., and Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L. and Roossin, P. S. (1990) A Statistical Approach to Machine Translation. *Computational Linguistics* **16(2)**: 79–85.
- Brown, P. F., Della Pietra, S. A., and Della Pietra, V. J. and Mercer, R. L. (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* **19(2)**: 263–311.
- Brown, P. F., Della Pietra, S. A., and Della Pietra, V. J. and Mercer, R. L. (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* **19(2)**: 263–311.
- Byrne, W., Khudanpur, S., Kim, W., Kumar, S., Pecina, P., Virga, P., Xu, P. and Yarowsky, D. (2003) The Johns Hopkins University 2003 Chinese-English Machine Translation System. *Proceedings of MT Summit IX*, pp. 447–450. New Orleans, LA, USA.
- S. F. Chen. Aligning sentences in bilingual corpora using lexical information. In *Meeting of the Association for Computational Linguistics*, pages 9–16. 1993.
- Deng, Y. and Byrne, W. (2004) Bitext Chunk Alignment for Statistical Machine Translation. Research Note, Center for Language and Speech Processing, Johns Hopkins University.
- Doddington, G. (2002) Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Proceedings of the Conference on Human Language Technology, pp. 138–145, San Diego, CA. USA,
- W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184. 1991.
- Germann, U., Jahr, M., Knight, K., Marcu, D. and Yamada, K. (2001) Fast Decoding and Optimal Decoding for Machine Translation. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 228–235. Toulouse, France.
- M. Haruno and T. Yamazaki. High-performance bilingual text alignment using statistical and dictionary information. In *Proceedings of ACL '96*, pages 131–138. 1996.
- Knight, K. and Al-Onaizan, Y. (1998) Translation with Finite-State Devices, *Proceedings of the AMTA Conference*, pp. 421–437. Langhorne, PA, USA.
- Koehn, P., Och, F. and Marcu, D. (2003) Statistical Phrase-Based Translation. *Proceedings of the Conference on Human Language Technology*, pp. 127–133. Edmonton, Canada.
- Kumar, S. and Byrne, W. (2002) Minimum Bayes-Risk Alignment of Bilingual Texts. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 140–147. Philadelphia, PA, USA.
- Kumar, S. and Byrne, W. (2003) A Weighted Finite State Transducer Implementation of the Alignment Template Model for Statistical Machine Translation. *Proceedings of the Conference on Human Language Technology*, pp. 142–149. Edmonton, Canada.
- Kumar, S. and Byrne, W. (2004) A Weighted Finite State Transducer Translation Template Model for Statistical Machine Translation. Research Note No. 48, Center for Language and Speech Processing, Johns Hopkins University.
- Kumar, S. and Byrne, W. (2004) Minimum Bayes-Risk Decoding for Statistical Machine Translation *Proceedings of the Conference on Human Language Technology*, pp. 169–176. Boston, MA, USA.
- Kumar, S. *Minimum Bayes-Risk Techniques in Automatic Speech Recognition and Machine Translation*, Ph.D. Dissertation, Johns Hopkins Univ. Electrical and Computer Engineering Dept., 2004
- Kumar, S and Byrne, W. (2005) A weighted finite state transducer translation template model for statistical machine translation, *Journal of Natural Language Engineering*, to appear.
- *LDC Chinese Segmenter*. http://www.ldc.upenn.edu/Projects/Chinese. LDC, 2002.
- Marcu, D. and Wong, W. (2002) A Phrase-Based, Joint Probability Model for Statistical Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 133-139. Philadelphia, PA, USA.
- I. D. Melamed. A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th Annual Conference of the Association for Computational Linguistics.*, pages 305–312. 1997.
- Mohri, M., Pereira, F. and M. Riley (1997), ATT General-purpose finite-state machine software tools. http://www.research.att.com/sw/tools/fsm/
- NIST (2004) The NIST Machine Translation Evaluations. http://www.nist.gov/speech/tests/mt/.
- R. C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceeding of 5th Conference of the Association for Machine Translation in the Americas*, pages 135–244.
- Och, F. (2002) Statistical Machine Translation: From Single Word Models to Alignment Templates. PhD. Thesis, RWTH Aachen, Germany,
- Och, F. and Ney, H. (2000) Improved statistical alignment models. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* pp. 440–447. Hong Kong, China.
- Och, F., Tillmann, C. and Ney, H. (1999) Improved Alignment Models for Statistical Machine Translation. *Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 20–28. College Park, MD, USA.
- Och, F., Ueffing, N. and Ney, H. (2001) An Efficient A* Search Algorithm for Statistical Machine Translation. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 55–62. Toulouse, France.
- Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2001) Bleu: a Method for Automatic Evaluation of Machine Translation. Tech Report RC22176 (W0109-022), IBM Research Division.
- M. Simard and P. Plamondon. Bilingual sentence alignment: Balancing robustness and accuracy. In *Machine Translation*, volume 13, pages 59–80. 1998.
- D. Wu. Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Meeting of the Association for Computational Linguistics*, pages 80–87. 1994.
- Stolcke, A. (2002) SRILM - An Extensible Language Modeling Toolkit. http://www.speech.sri.com/projects/srilm/, *Proceedings of the International Conference on Spoken Language Processing*, pp. 901–904. Denver, CO, USA.
- Tillmann, C. and Ney, H. (2003) Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics* **29(1)**: 97–133.
- Tillmann, C. (2003) A Projection Extension Algorithm for Statistical Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan.
- Ueffing, N., Och, F. and Ney, H. (2002) Generation of Word Graphs in Statistical Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 156–163. Philadelphia, PA, USA.
- Vogel, S., Ney, H. and Tillmann, C. (1996) HMM Based Word Alignment in Statistical Translation. *Proceedings of the 16th International Conference on Computational Linguistics* pp. 836–841. Copenhagen, Denmark.
- Wang, Y. and Waibel, A. (1997) Decoding Algorithm in Statistical Machine Translation. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pp. 366–372. Madrid, Spain.
- JHU (2003), Syntax for Statistical Machine Translation, Final Report, JHU Summer Workshop. http://www.clsp.jhu.edu/ws2003/groups/translate/.
- Zens, R. and Ney, H. (2004) Improvements in Phrase-based Statistical Machine Translation. *Proceedings of the Conference on Human Language Technology*, pp. 257–264. Boston, MA, USA.
- Zhang, Y. Vogel, S. and Waibel A. (2003), Integrated Phrase Segmentation and Alignment Model for Statistical Machine Translation. *Proceedings of the Conference on Natural Language Processing and Knowledge Engineering*. Beijing, China.

Thank you !