JHU/CUED Chinese-English Translation System 2005 TC-STAR Evaluation

Shankar Kumar, Yonggang Deng, Bill Byrne

Machine Intelligence Laboratory
Cambridge University Engineering Department

Center for Language and Speech Processing
The Johns Hopkins University

wjb31@eng.cam.ac.uk

22 April 2005

Overview

Similar architecture to JHU'04 Chinese→English NIST Translation System
- Optimized for news text

- Used all bitext sources from LDC: 175M/207M Chinese/English words
- Subsentence alignment by bitext chunking (Y. Deng'03)
- Bitext alignments generated under new modeling framework
  - Alignment quality equals IBM Model 4, with fast, exact estimation and search
  - *no longer need IBM Model 4*
- Phrase pairs extracted from alignments using the usual heuristics
- TTM : WFST implementation (Kumar, Deng, Byrne - JNLE'05)
- MET performed over provided dev sets for text and verbatim conditions (used for ASR)
- 2 pass decoding - LMs estimated with SRI LMTK
  - First pass lattice generation with heavily pruned 3-gram LM
  - Second pass lattice rescoring and generation with 3- and 4-gram LMs
  - MET applied for a contrast system
- MT Model Training - Y. Deng, Decoder Implementation - S. Kumar
- Four days total effort : two days in development, two days in evaluation

We in the JHU/CUED team thank TC-STAR for the opportunity to participate !

Sentence Alignment

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

$K = 3$:

$$\begin{array}{lll} \mathbf{e}_1\mathbf{e}_2 & \mathbf{e}_3\mathbf{e}_4\mathbf{e}_5 & \mathbf{e}_6 \\ \mathbf{f}_1 & \mathbf{f}_2\mathbf{f}_3 & \mathbf{f}_4 \end{array}$$

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

Can describe the alignment of chunks of sentences

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

Sentence Alignment via Divisive Clustering (Y. Deng'03)

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 :

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 。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 :

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 。	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 .
	3
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 ”。	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 。	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

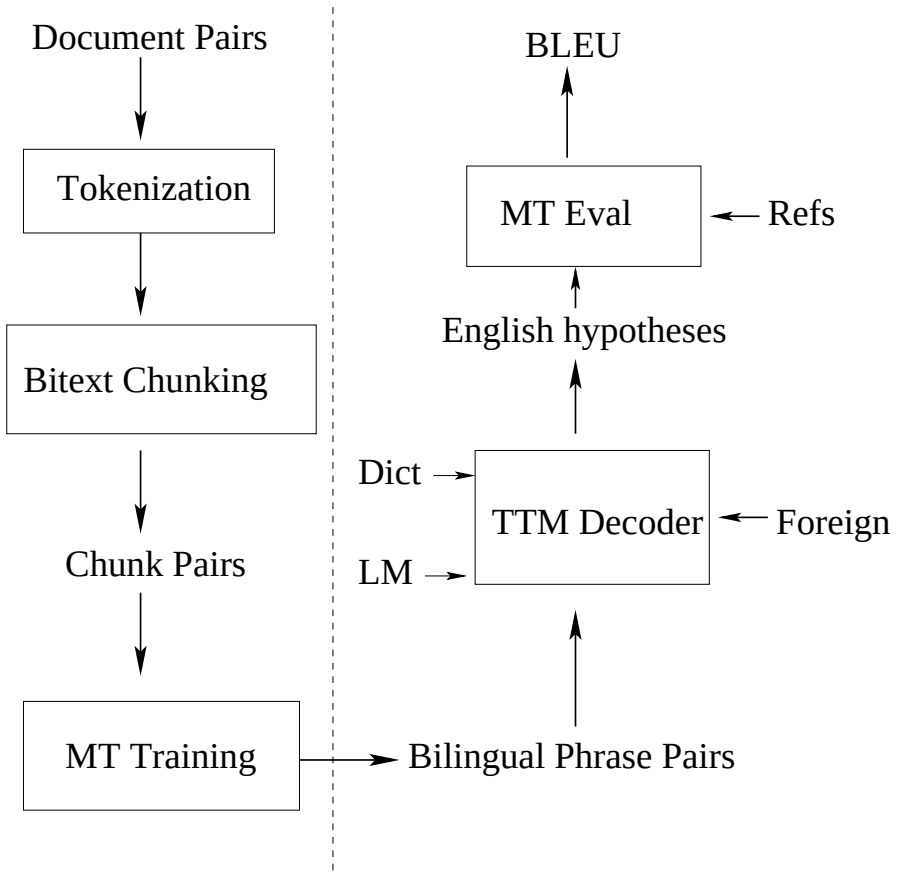
Benefits of good chunking

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

A good alignment procedure :

- fast
- efficient: as little bitext should be discarded as possible
- flat-start
- language independent
- require minimal linguistic knowledge
- subsentence chunks

Coarse monotonic alignment followed by fine divisive clustering works well



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

Current work:

- estimate phrase pairs under the model to improve test set coverage

Future work:

- direct use of models in translation -
  support discriminative training, adaptation, etc.

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

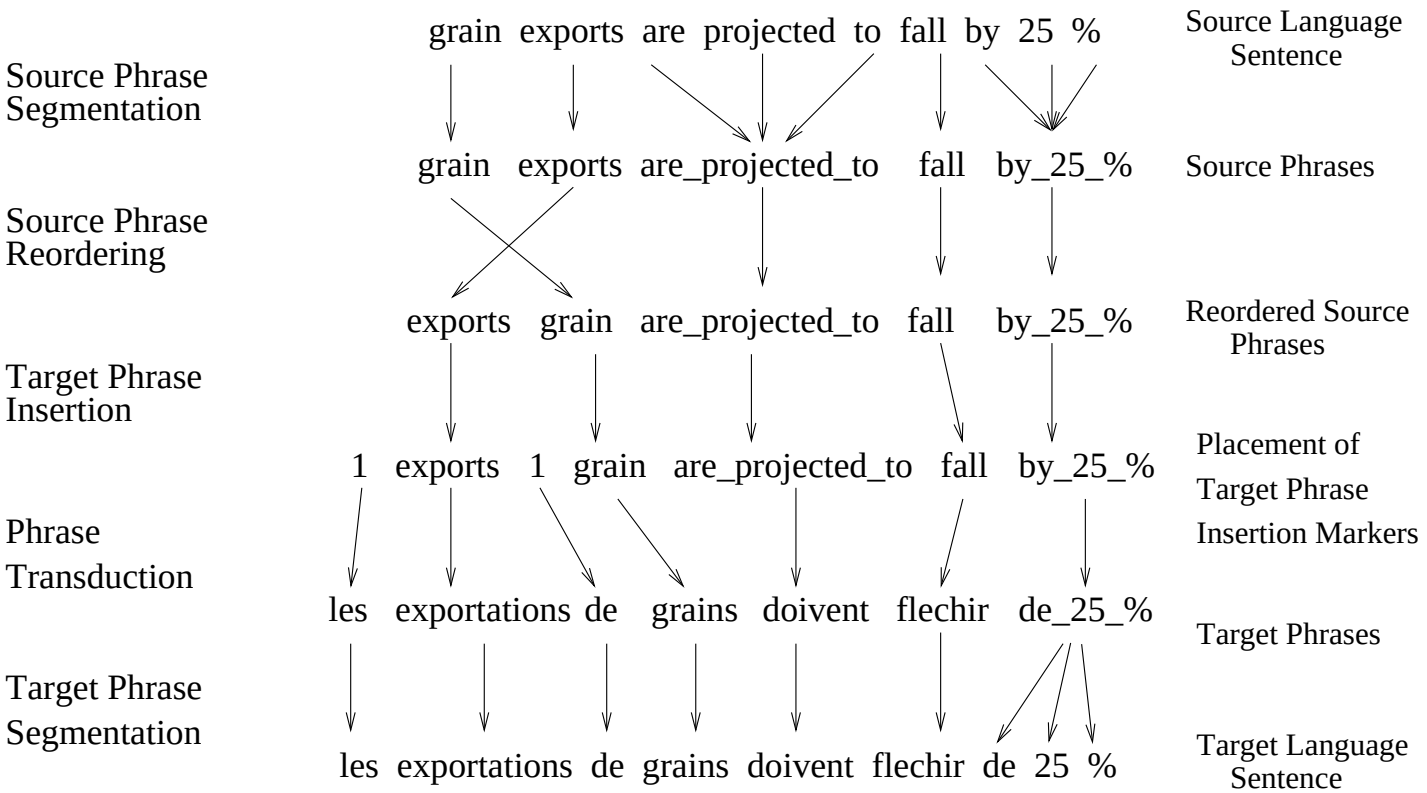
- Can be limiting, but avoids special-purpose decoders
- The value of this modeling approach has been shown in ASR by the systems developed at AT&T
  - Translation is performed using libraries of standard FSM operations
  - Clear formulation

Translation Template Model - TTM (Kumar, Deng, Byrne - JNLE'05)

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
- *Actual* source-channel model - with proper component distributions
- Implementation is direct using standard WFST operations
- Uses publicly available AT&T FSM Tools

TTM and MET

MET can be used to optimize the combination of TTM components

- Cast TTM as a log-linear model with scaling factors $\Lambda = \lambda_1^M$
  $P_{TTM}(E|F) = \prod_{m=1}^{M} p_m(E,F)^{\lambda_m}$
  λ's applied to WFSTs during decoding
- Minimum Error Training (Och 2003) : Estimate parameters of a log-linear model to reduce error count over a development set
- Minimize an Error Function $\mathcal{E}$ (BLEU) over a development corpus:

  N-best lists used for training

  Multidimensional search in M dim space by Powell's algorithm

  MET gives good improvement over a state-of-the-art baseline

Training and Translation Procedures

1. Bitext Chunking
   - 1.1 Monotone alignment into coarse chunks of documents
   - 1.2 Divisive clustering into subsentence chunks
2. Generate word alignments in each translation direction
   - 2.1 No need to partition training sets
   - 2.2 Training speed is several times faster than IBM Model 4, owing to parallelization
3. Extract phrase pairs using the ‘standard’ heuristic
4. Construct component WFSTs for the TTM
5. Translation lattice generation with pruned trigram
6. Translation lattice rescoring with unpruned 4gram →
7. MET under BLEU→
   - 7.1 1000-best lists for both text and ‘verbatim’ conditions

Text Processing & System Building Strategy

Goal: Exploit all available text resources

- Chinese Text segmented into words using LDC segmenter (Linguistic Data Consortium)
- English Text processed using a simple tokenizer

Bitext for Translation Model Training

	Chinese-English
# of chunk pairs (M)	7.6
# of words (M)	175.7/207.4

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

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 Partition	Contribution by Source: En words (M) FBIS	HKNews	XHTS	UN	Total	BLEU (%) 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
Decode with Big3g							
1+2+3						27.7	27.1
Unpartitioned training - new procedure						27.9	27.0

(XHTS = Xinhua, Hansards, Treebank, Sinorama)

TC-STAR Chinese-English Dev and Eval Results

	BLEU		BLEU (cased)		
	Dev		Eval		
	Text	Verbatim	Text	Verbatim	ASR
Vanilla Decoding	16.8	14.6	14.6	12.9	13.4
MET	17.4	15.7	15.7	13.4	13.2

Conclusions

- Good, basic system
  - Performed more-or-less as expected in the Text task
  - Needs tuning for ASR and Verbatim conditions
- Our evaluation system was a 'snapshot' of a work-in-progress
  - Less than 4 days effort overall, including the evaluation
  - Please don't draw any conclusions about the overall system quality

Our motivation: Access to data

- Mandarin ASR (lattices) and reference translations