



(Have we found the Holy Grail?)

Panel at MT-Summit 2003



# The HKUST Leading Question Translation System

v0.1 alpha (experimental)

---

**Dekai Wu**

Human Language Technology Center

Department of Computer Science

Hong Kong University of Science and Technology



## The panel chair has put forth 9 questions to be addressed...

---

- Just fed each question to our experimental system
- Will show n-best translations for each
- Caveat: somewhat flaky
- Suggest we try to answer the questions on the fly here



# Question 1

---

**Reference** Have we found the ultimate solution to MT's long quest? If not, is the Holy Grail just around the corner?

**Translation** Are we just about done?



## Question 2(a)

---

**Reference** Will progress in data-driven MT continue unabated?

**Translation**  $\lim_{t \rightarrow \infty} \max_{m \in M_t} \text{BLEU}'(m) = \infty$  ?



## Question 2(b)

---

**Reference** Is there an inherent ceiling on MT quality that will resist even the most sophisticated data-driven methods?

**Translation 1** Are data-driven methods excluded from making use of linguistic or semantic features?

**Translation 2** Is there an inherent ceiling on MT quality that will resist even the most sophisticated methods?



# Question 3

---

**Reference** Has the data-driven paradigm been able to model information that was not present in rule-based systems? Or has it `simply' been able to model the same kind of information more thoroughly and efficiently?

**Translation 1** In data-driven models, do we find `rules' (categories, collocations, templates, decision algorithms)?

**Translation 2** In rule-based systems, do we find massive amounts of fine-grained information on lexical chain preferences (n-grams), collocational correlations, interacting lexical choice factors that support consistent evidence combination, etc?



# Question 4

---

**Reference** Was the metric used to rank participating systems in the NIST competition fair, or was it somehow biased in favor of data-driven systems?

**Translation 1** Did MTEval's scores rate all competing models fairly, or were they preferential in some way to corpus-based models?

**Translation 2** Was the metric used to rank participating systems in the NIST competition somehow biased in favor of data-driven systems, or was it fair?





# Question 5

---

**Reference** Even if the evaluation metric used at NIST was somewhat biased, can we still assume that SMT has indeed surpassed traditional rule-based systems? And if so, at what exactly?

**Translation 1** We may assume anything we want (as long as we state our assumptions). But if we don't like the NIST result, what is it that we wish to prove instead?

**Translation 2** Can we conclude that the rate of improvement of SMT has surpassed the rate of improvement of traditional rule-based systems? And if so, whether at the current rate of progress this will soon no longer be an interesting question?



# Question 6

---

**Reference** Are there niche applications for which the new data-driven techniques are particularly well suited?

**Translation 1** Are there applications for which the traditional techniques are *not* particularly well suited?

**Translation 2** Is there anything else we can work on besides translators' tools and/or intelligence gathering?



# Question 7

---

**Reference** Is there a danger that SMT's recent success may lead the public – and worse yet, the funding agencies – to believe that the MT problem has finally been solved, and so to reduce the level of R&D grants to our field? If so, what can we do to combat this misperception?

**Translation 1** Should we launch a campaign blitz to get the public/funders to 'Take the MT Challenge' and test drive our current clunkers for themselves? If so, will they all swear off MT forever in disgust?

**Translation 2** How do we define a metric that correlates with human judgment at least as well as BLEU, but generates much lower numeric values that imply we have a long way to go?



# Question 8

---

**Reference** Would the results of the NIST competition have been different if the languages involved had been English and French? If so, why?

**Translation 1** Does French's more complex morphology hinder some of us? Does the large number of cognates and similar conceptual structure to English help some of us?

**Translation 2** Have the groups working on English/French had a long time to fine-tune their components, features, and resources?



# Question 9

---

**Reference** In previous debates on this question (TMI-92) many people concluded that hybrid systems were the way of the future. What role do rule-based components play in today's leading data-driven systems, and what are the prospects for their future contribution?

**Translation** <#@%\$&^#\$??!/>



# In the beginning...

	<b>SMT</b>	<b>EBMT</b>	<b>Transfer</b>
Hypothesis probabilities	1		
Collocation blexicons		1	
Transduction rules			1



# But then...

---

- SMT plants trees
  - got collocation bilexicons?
    - learning: Wu & Xia 1995, Smadja 1996, Och et al. 1999, Koehn et al. 2003
    - decoding: Wu 1996, Och et al. 1999, Koehn et al. 2003
- EBMT gets serious about template abstraction
  - got transduction rules?
- Transfer models string out
  - got collocation bilexicons?



# So then...

	<b>SMT</b>	<b>EBMT</b>	<b>Transfer</b>
Hypothesis probabilities	0.6		
Collocation bilexicons	0.4	0.6	0.4
Transduction rules		0.4	0.6





# And then...

---

- SMT plants trees
  - got collocation bilexicons?
    - learning: Wu & Xia 1995, Smadja 1996, Och et al. 1999, Koehn et al. 2003
    - decoding: Wu 1996, Och et al. 1999, Koehn et al. 2003
  - got 'real linguistic' transduction grammar rules?
    - eg: Wu & Wong 1998, Alshawi et al. 1998, Yamada & Knight 2001, Melamed 2003, Schafer & Yarowsky 2003
- EBMT gets real about scoring
  - got probabilities?
    - eg: Brown et al. 2003
- Transfer models soften up
  - got scores, backoff, stronger decoders?
    - eg: Lavie et al. 2003



# So then...

	<b>SMT</b>	<b>EBMT</b>	<b>Transfer</b>
Hypothesis probabilities	0.5	0.2	0.2
Collocation bilexicons	0.3	0.5	0.3
Transduction rules	0.2	0.3	0.5



# Convergence

	<b>SMT</b>	<b>EBMT</b>	<b>Transfer</b>
Hypothesis probabilities	0.33	0.33	0.33
Collocation bilexicons	0.33	0.33	0.33
Transduction rules	0.33	0.33	0.33



# The future of MT is...

---

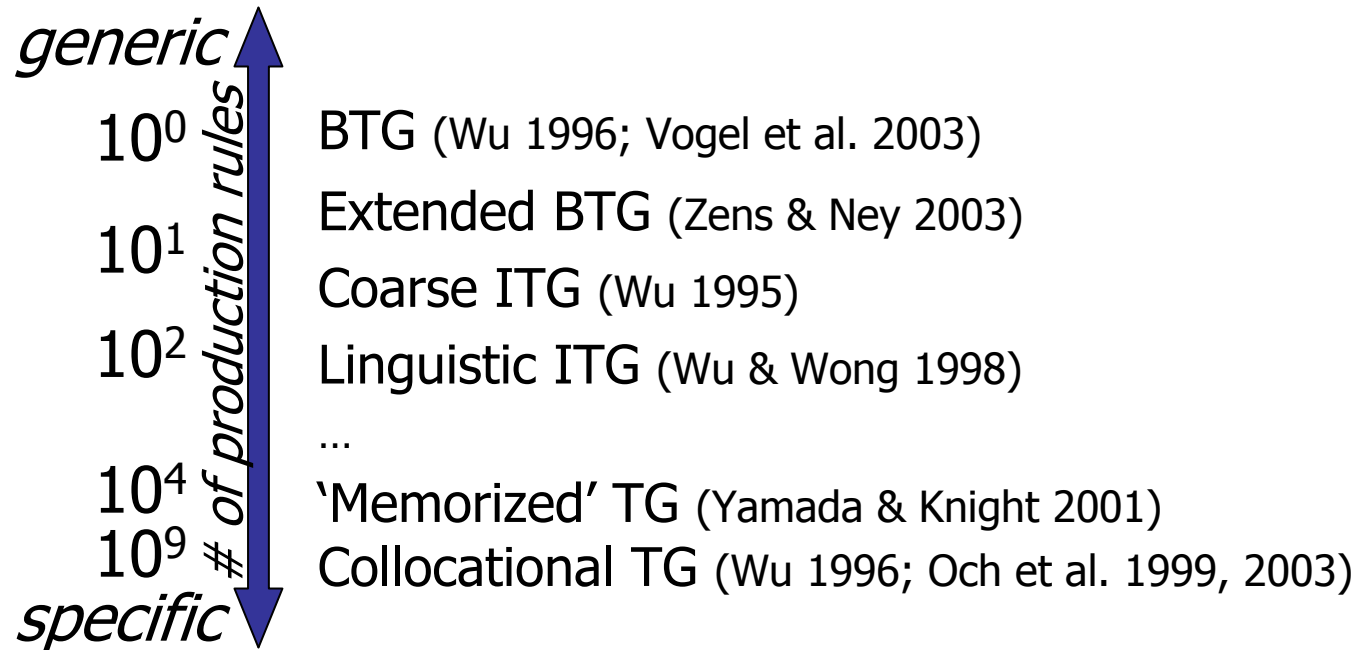
## Trees in statistical transduction models

- Beware...
  - Adding syntax to language models...?
- Still... MT is more inherently tree-ish than speech
  - Constituent order does vary between languages
  - Need freedom to generate legitimate paraphrases
  - Need efficient but (nearly) optimal decoding for both training and runtime
- But a number of fundamental questions need to be answered...



# Trees in Statistical Transduction

- More generic or more specific?
- Linguistic interpretation or not?

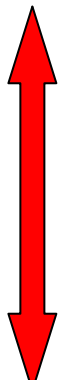




# Trees in Statistical Transduction

---

- Degree of coupling?



Completely independent source, target trees

- How to link?

Transduction grammars

- (aka bigrammars, synchronous grammars)

- Variants

- Heads identified or not?

- (aka dependency models) (Alshawi et al. 2001; Melamed 2003)
- Just notation, or real mathematical distinctions?



# Trees in Statistical Transduction

---

- Bias toward input or output language?
  - Input: parse input sentence
  - Output: Coerce input language observables into output language hidden
    - Improves fluency of output
    - Also seems to improve adequacy!



# Trees in Statistical Transduction

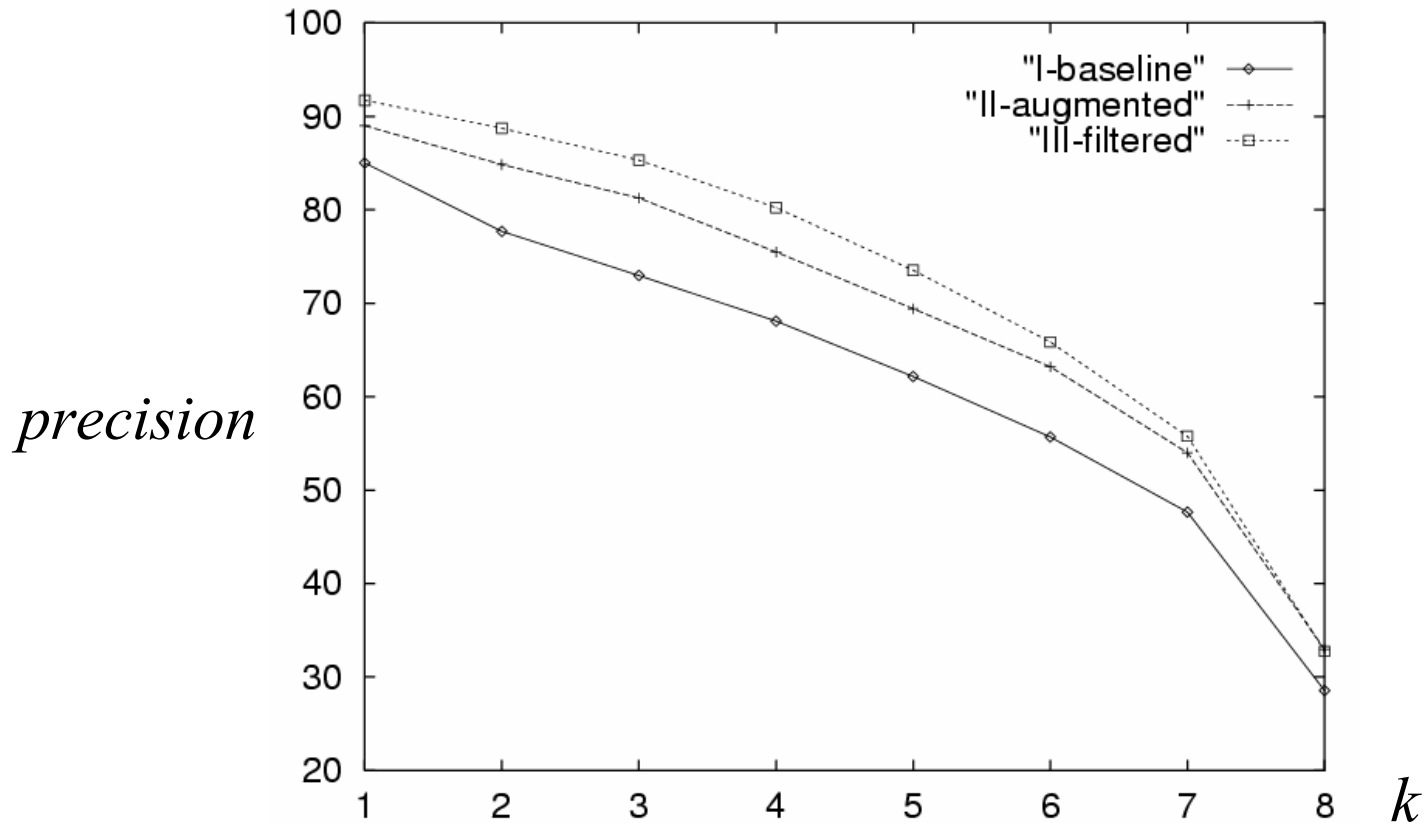
---

- Collocations matter
  - Everything is collocations (esp in Chinese)
  - Collocation segmentation greatly affects accuracy (aka segment/phrase chunking/tokenization)
- How to find `correct' segmentation?
- How to evaluate?





There's no *a priori* 'correct' segmentation...  
Modifying the performance measure so that it rewards 'fixed points' can impact scores heavily.



$nk$ -blind precision comparisons for  $n = 8$  judges (Wu & Fung 1994)



# “Soft Segmentation”

---

## ■ **Soft segmentation**

- Accuracy improved by *integrating* segmentation with other translation decisions
- Avoids premature commitments  
(Wu 1996; Zens & Ney 2003)

## ■ **Nested brackets**

- Better coverage, generalization, explanatory power, compared to flat 1-level bracketing
- Fast, when done right  
(Wu 1996, 1998; Alshawi et al. 1998; Melamed 2003)



# Finding the Holy Grail

---

## **Trees in statistical transduction... but smarter.**

Some fundamental questions need to be answered.

- Trees with what characteristics?
  - Generic or specific? Linguistic or not?
  - Degree of coupling?
  - Dependency and other variants? Just notation, or real distinctions?
  - Bias toward input or output language?
- Will performance measures pick up on grammaticality improvements? How?
  - No empirical verification yet
  - Interannotator agreement reward?



# So have we found the Holy Grail?

---

- Definitely moving fast, in the right direction.
- Where did the myth that statistical models don't have structure come from?
  - (In particular: the structure may be tree-like!)
- On evaluation – why would we want to stop making best guesses as to how well our systems are doing?
- Regardless – why would we want to discard any of the power of statistical modeling from our toolbox?