Search for Neural Translation and Dialogue Systems – L6

Bill Byrne

Lent 2022

Neural Machine Translation and Dialogue Systems – MLMI8

MPhil in Machine Learning and Machine Intelligence

UNIVERSITY OF CAMBRIDGE

Search Strategies for Sequence-to-Sequence Models

- Search
  - Greedy Decoding & Beam Search
    - How search errors arise
  - NMT Search Spaces
    - Constrained decoding -- lattice rescoring
  - Exact Decoding: Depth First Search
  - NMT Model consistency
    - Search errors and modelling errors
  - Alternatives to MAP decoding
    - The likelihood trap
    - Sampling
    - Nucleus Sampling
    - Minimum Bayes Risk Decoding
    - Bayesian Interpolation

N.B. Much of the experimentation into search procedures is in NMT, but limitations of MAP decoding are reported in a variety of text generation tasks

Search Strategies

Transformers define a left-to-right conditional distribution over target language sentences:

- Source sentence: **x** , Target sentence: **y** drawn from a target language vocabulary Σ

$$P_T(\mathbf{y}|\mathbf{x},\theta) = \prod_i P_T(y_i|y_{<i},\mathbf{x},\theta)$$

- For a source sentence, **x**, the maximum a posteriori decoding rule suggests generating a translation **y** :

$$\underset{\mathbf{y}\in\Sigma^*}{\operatorname{argmax}}\, P_T(\mathbf{y}|\mathbf{x},\theta) = \underset{\mathbf{y}\in\Sigma^*}{\operatorname{argmax}} \prod_i P_T(y_i|y_{<i},\mathbf{x},\theta)$$

  - sometimes referred to as the **mode** of the distribution

Today's lecture

- Can we carry out this MAP search over Σ^* exactly?
- If exact MAP search is not possible, what approximate search procedures are available?
- Is the MAP decoding strategy appropriate for the quality metrics we care about?
- Do sequence-to-sequence define a valid posterior distribution?

Greedy Decoding

Construct a single hypothesis word-by-word, simply by picking the best next word:

- For j = 1, ...
  - Partial hypothesis $y_{<j} = y_1 \ldots y_{j-1}$
  - $y_j = \underset{y \in \Sigma^*}{\operatorname{argmax}} P_{NMT}(y | y_{<j}, \mathbf{x}, \boldsymbol{\theta})$
  - Stop if y_j = </s> is the end of sentence marker

Advantages:

- Simple and fast
- Can avoid computation of the softmax in the output layer

Disadvantages:

- Clearly not a good approximation to the MAP criterion $\underset{\mathbf{y} \in \Sigma^*}{\operatorname{argmax}} P_{NMT}(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})$

Beam Search

Beam search of Width *N*

- Beam search is synchronous in the target length
- For target hypothesis length j = 1, ...
  - N partial hypotheses $y_{<j}^{k}$, k = 1, ..., N
  - Expand each partial hypothesis not ending in </s> by the *N* most likely words to follow
    - This yields up to N^2 hypotheses
  - Pruning: Keep the top scoring *N* hypothesis $y^{k}_{<j+1}$, k = 1, ..., N
  - Stop when all hypotheses end with </s>

- Aim is to avoid the search errors made in greedy decoding
  - Greedy decoding is Beam Search of size 1

How Search Errors Arise – Good Paths are Discarded Too Soon

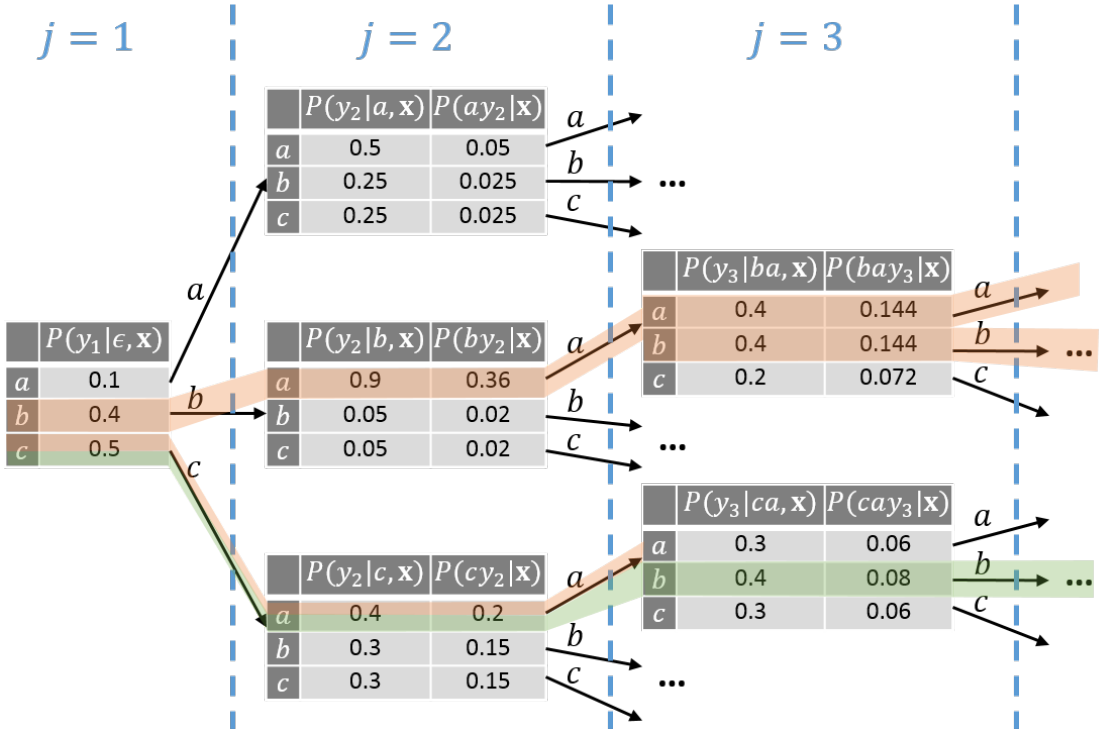
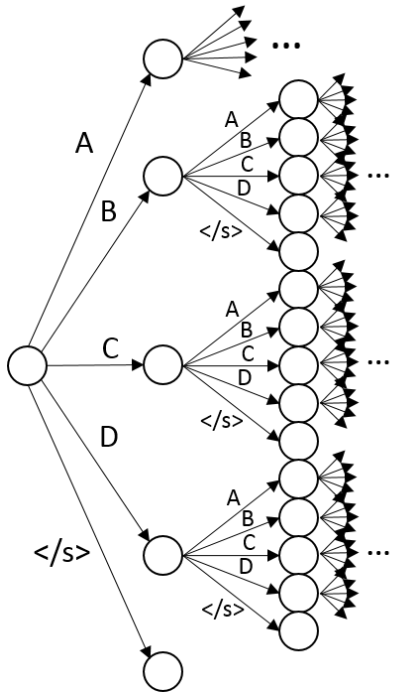


Fig. 3.16 Comparison between greedy (highlighted in green) and beam search (highlighted in orange) with beam size 2.

Search Space is Tree-Structured

Searching on trees often greatly simplifies search algorithms as we do not have to keep track of already visited nodes.

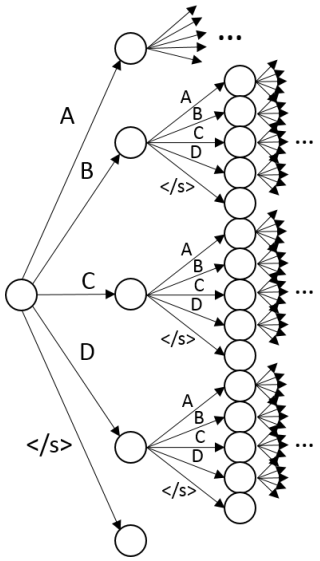
- Search procedures vary by how they explore and discard paths through the tree
- **Greedy search** and **beam search** are shortest path search procedures in an (in)finite state machine
- Beam Search is a form of breadth first search, with breadth N



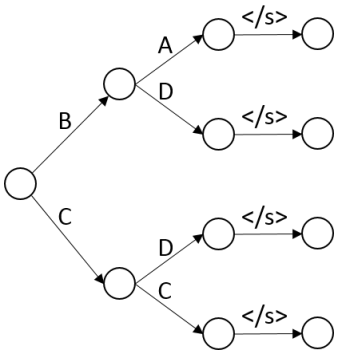
NMT Decoding Over a Lattice

Performing NMT decoding over a lattice effectively yields a search space that is the intersection of the NMT infinite tree with a tree derived from the lattice.

- lattice paths are fully expanded into a tree
- potentially still large, but typically much smaller than the unconstrained NMT search space
- A* and Depth-First Search are possible for both finite and infinite trees



(a) Unconstrained NMT search space.



(b) Search space constrained with the Hiero lattice in Fig. 4.2.

Meister et al. **Best-First Beam Search.** TACL'2020
https://www.mitpressjournals.org/doi/full/10.1162/tacl_a_00346

Depth-First Search (DFS) in NMT

A DFS strategy searches “ ‘deeper’ in the graph whenever possible"

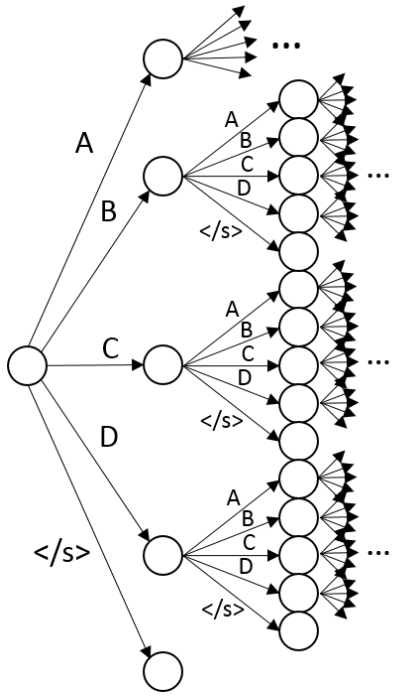
In contrast to time synchronous search like beam or breadth-first search, it explores a branch of the search tree exhaustively before “backtracking" into another branch.

Therefore, it always explores the state in the current frontier with maximum path length to the initial state

- Suppose we find a complete hypothesis **y** with $\gamma = P(\mathbf{y}|\mathbf{x};\theta)$
- It follows that $\gamma \leq \underset{\mathbf{y}\in\Sigma^*}{\operatorname{argmax}} P(\mathbf{y}|\mathbf{x};\theta)$

- Therefore any prefix $y_{<j}$ can be discarded if $P(\mathbf{y}_{<\mathbf{j}}|\mathbf{x};\theta) < \gamma$
- γ is updated for an complete hypothesis with a better score
- The algorithm stops after all partial hypotheses are discarded

<u>If the algorithm halts, it has found the global maximum</u>



UNIVERSITY OF CAMBRIDGE

Stahlberg, Byrne *On NMT Search Errors and Model Errors: Cat Got Your Tongue?* EMNLP’19
https://aclanthology.org/D19-1331/

Search Errors vs Modelling Errors

Suppose we have a metric Q(y)

- Q could be BLEU, Semantic Accuracy or some other metric
- Q(y) > Q(y') should indicate that y is better than y', in particular wrt human judgement

Modelling errors:

- Suppose we have a probabilistic model P(y).
- P should rank hypotheses in agreement with the quality metric Q
  - If P(y) > P(y') but Q(y') > Q(y) , we have a modelling error

There is evidence that *search errors* are saving current systems *from their modelling errors*

- Decoding schemes are developed that explicitly avoid trying to solve for the MAP solution
- Search errors are introduced `on purpose', to alleviate modelling errors

NMT Models Need Not Be Consistent

Probabilistic models of language should be *consistent:* $\sum_{\mathbf{y} \in \Sigma^*} P(\mathbf{y}; \theta) = 1$

Chen et al. construct Recurrent Neural Network Language Models that are not consistent:

- The model conditional distributions are consistent: $\sum_{y \in \Sigma} P(y | y_{<j}; \theta) = 1$
- But that's not enough to show overall consistency
  - Simple RNN language model over a language consists of strings $a^n = \underbrace{a a \ldots a a}_{n}$
  - But they show (very cleverly) that $\sum_{\mathbf{y} \in \{a\}^*} P(\mathbf{y}; \theta) \approx 0.14$
- Cannot assume that NMT models are normalised probability distributions

  Without a determination of the overall probability mass assigned to all finite strings, a fair comparison of language models with regard to perplexity is simply impossible.
- N.B. this problem cannot arise with models with a finite history (e.g. finite state Markov Models)

Fun Fact: Exact Search / Depth First Search applied to this model will never halt!

Chen et al. Recurrent Neural Networks as Weighted Language Recognizers NAACL 2018
https://www.aclweb.org/anthology/N18-1205/

With DFS / Exact Search We Can Exactly Assess NMT Search and Modeling Errors

Search	BLEU	Ratio	#Search errors	#Empty
Greedy	29.3	1.02	73.6%	0.0%
Beam-10	30.3	1.00	57.7%	0.0%
Exact	2.1	0.06	0.0%	51.8%

Table 1: NMT with exact inference. In the absence of search errors, NMT often prefers the empty translation, causing a dramatic drop in length ratio and BLEU.

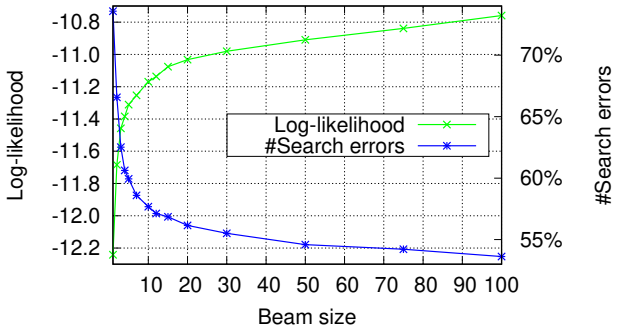


Figure 2: Even large beam sizes produce a large number of search errors.

Model	Beam-10 BLEU	Beam-10 #Search err.	Exact #Empty
LSTM*	28.6	58.4%	47.7%
SliceNet*	28.8	46.0%	41.2%
Transformer-Base	30.3	57.7%	51.8%
Transformer-Big*	31.7	32.1%	25.8%

Table 2: *: The recurrent LSTM, the convolutional SliceNet (Kaiser et al., 2017), and the Transformer-Big systems are strong baselines from a WMT'18 shared task submission (Stahlberg et al., 2018a).

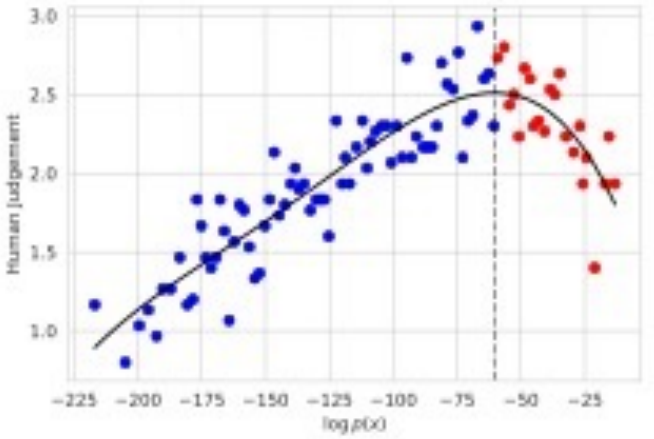
- Surprisingly, beam search fails to find these global best model scores in most cases, even with a very large beam size of 100.
- For more than 50% of the sentences, the model assigns its global best score to the empty translation, revealing a massive failure of neural models in properly accounting for adequacy.
  - verified by Shi et al. with very deep beam search (beam=500)

Stahlberg, Byrne *On NMT Search Errors and Model Errors: Cat Got Your Tongue?* EMNLP'19
Shi et al. *Why Neural Machine Translation Prefers Empty Outputs* https://arxiv.org/abs/2012.13454

The Likelihood Trap

.... Our experiments also confirm the existence of the ‘likelihood trap’, the counter-intuitive observation that high likelihood sequences are often surprisingly low quality. ...

We asked 146 crowd-workers to rate the quality of 100 sentences across a variety of model likelihoods. While model log likelihoods are generally positively correlated with average human quality judgments, we notice an inflection point after which they become negatively correlated.



Zhang et al. *Trading Off Diversity and Quality in Natural Language Generation.* https://arxiv.org/abs/2004.10450

Alternatives to MAP Decoding

Sampling

Greedy decoding constructs a single hypothesis word-by-word, by picking the best next word.

Sampling is similar: the next word is sampled from the output distribution

- For j = 1, ...
  - Partial hypothesis $y_{<j} = y_1 ... y_{j-1}$
  - Generate next word by sampling: $y_j \sim P(y|y_{<j}, \mathbf{x}, \theta)$
  - Stop if y_j = </s> , i.e. at the end of sentence marker

Can be easily extended to N hypotheses

Useful for estimating the expectation of some function of the hypotheses

- Generate *N* hypotheses $\mathbf{y}^k \sim P(\mathbf{y}|\mathbf{x}, \theta)$, $k = 1, \ldots, N$
- $E(g(Y)|X;\theta) \approx \frac{1}{N} \sum_k g(\mathbf{y}^k)$

Often set a temperature for sampling – greedy decoding is sampling at temperature 0

Nucleus Sampling for Natural Language Generation

The counter-intuitive empirical observation is that even though the use of likelihood as training objective leads to high quality models for a broad range of language understanding tasks, maximization-based decoding methods such as beam search lead to degeneration— output text that is bland, incoherent, or gets stuck in repetitive loops.

Nucleus Sampling:

- Set a threshold *p*
- Let V_j be the smallest set of target words such that $\sum_{y \in V_j} P(y|y_{<j}, \mathbf{x}, \theta) \geq p$
- Sample the next word from V_j , proportionally to *P*

NUCLEUS Sampling Reduces Hallucinations and Neurobabble in Generation

Prompt: An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search (N=16): The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

NUCLEUS (p=0.95): There has been an unprecedented number of calves caught in the nets of whaling stations that operate in WA. Pilot whales continue to migrate to feeding grounds to feed their calves. They are now vulnerable due to the decline of wild populations; they are restricted to one breeding site each year. Image copyright Yoon Bo Kim But, with sharp decline in wild populations the size of the Petrels are shrinking and dwindling population means there will only be room for a few new fowl.

NMT models do capture important aspects of translation well in expectation

Inadequacy of the mode: observations from the literature:

- translations produced via beam search typically under-estimate sequence length
- *beam search curse :* translation quality deteriorates with better approximate search
  - deeper beam search leads to quality degradation
- the true most likely translation (i.e., the mode of the distribution) is empty in many cases
- negative correlation between likelihood and quality, beyond a certain likelihood value

Assertions:

i. MAP decoding is not well-suited as a decision rule for MLE-trained NMT

ii. pathologies and biases observed in NMT are not necessarily inherent to NMT as a model or its training objective, rather, MAP decoding is at least partially responsible for many of these pathologies and biases

iii. a straight-forward approximation to a sampling-based decision rule known as minimum Bayes risk decoding gives good results, showing promise for research into decision rules that take into account the distribution holistically.

Eikma, Aziz. *Is MAP Decoding All You Need? The Inadequacy of the Mode in Neural Machine Translation.* COLING'2020

Minimum Bayes Risk Decoding

Risk combines likelihood and loss

- Loss function *L(**y**, **y'**)* - `distance' between ***y*** and ***y'***
- Loss can be (negative) BLEU, Word Error Rate – any string-to-string distance

The risk of a hypothesis ***y'*** is its expected loss relative to all other hypotheses:

$$R(\mathbf{y}'|\mathbf{x},\theta) = E_{P_\theta}[L(\mathbf{y}',\mathbf{y})\,|\,\mathbf{x}] = \sum_{\mathbf{y}} L(\mathbf{y}',\mathbf{y})P(\mathbf{y}|\mathbf{x};\theta)$$

- A hypothesis has low risk if it is similar to other highly likely hypotheses under the loss

Minimum Risk Decoding aims to find the hypothesis with minimum risk:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}'}{\operatorname{argmin}} \underbrace{\sum_{\mathbf{y}} L(\mathbf{y}',\mathbf{y})P(\mathbf{y}|\mathbf{x};\theta)}_{R(\mathbf{y}'|\mathbf{x};\theta)}$$

MBR is a form of consensus decoding

- The most likely hypotheses under the model 'vote' for alternatives similar to themselves

MBR: Sampling or Beam Search

MBR is easy to formulate as a criterion.

- Challenge is to develop efficient search procedures.

Sampling:

- generate a list of hypotheses by sampling: $H = \{\mathbf{y} \sim P(\mathbf{y}|\mathbf{x}, \theta)\}$
- Find MBR hypothesis as $\hat{\mathbf{y}} = \underset{\mathbf{y}' \in H}{\operatorname{argmin}} \sum_{\mathbf{y} \in H} L(\mathbf{y}', \mathbf{y})$

Beam Search:

- Generate a list of *N* hypotheses *H*
- Find MBR hypothesis as $\hat{\mathbf{y}} = \underset{\mathbf{y}' \in H}{\operatorname{argmin}} \sum_{\mathbf{y} \in H} L(\mathbf{y}', \mathbf{y}) P(\mathbf{y}|\mathbf{x}, \theta)$

To work well, MBR should be based on a diverse sample of good quality hypotheses

- Sampling favors diversity
- Beam search favors the MAP hypothesis

Bayesian Interpolation

In multi-domain scenarios, we can train a model for each domain

- *T* domains, train a model P(y | x, θ_t) for each domain *t*

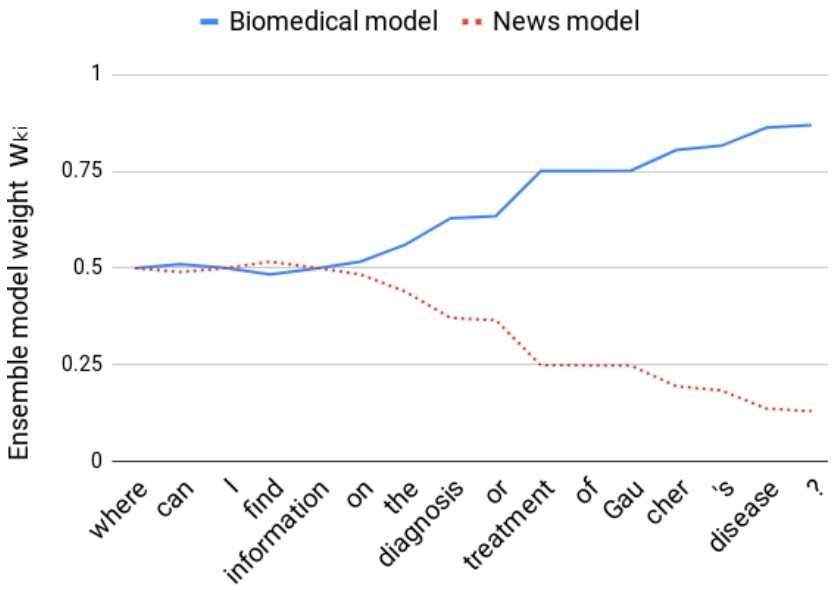
If the domain is unknown at translation time, Bayesian Interpolation can be used

- the posterior probability of each task is updated as decoding proceeds

$$p(y_j|y_{<j},\mathbf{x}) = \sum_{t=1}^{T} p(y_j,t|y_{<j},\mathbf{x})$$

$$= \sum_{t=1}^{T} p(t|y_{<j},\mathbf{x})\,P(y_j|y_{<j},\mathbf{x},\theta_t)$$

$$p(t|y_{<j}) = \frac{P(h_{<j}|\mathbf{x},\theta_t)P(t|\mathbf{x})}{\sum_{t'} P(h_{<j}|\mathbf{x},\theta_{t'})\,P(t'|\mathbf{x})}$$



Saunders et al. *Domain Adaptive Inference for Neural Machine Translation*. ACL'19 https://arxiv.org/abs/1906.00408

Conclusion

Standard search procedures are available and work well in many situations

- Greedy search, beam search, sampling

With complex models, search is nearly always sub-optimal

- Search errors are unavoidable and search interacts with the model
- With respect to quality, search errors are the `usual suspect'
- But new search procedures (DFS) allow us to assess exactly the effect of search errors

➢ Modelling errors are a bigger problem than search errors
  - ➢ Models are shown to be inconsistent – theoretically
  - ➢ DFS shows that some of the best available models are also inconsistent – in practice

- Alternative search procedures (relative to MAP) such as Nucleus Sampling are introduced to avoid the worst modelling errors

Bayesian approaches such as Minimum Bayes Risk decoding may play to the strengths

Plenty of opportunities for fundamental modelling improvements