N -gram posterior probability confidence measures for statistical machine translation: an empirical study

Adrià de Gispert, Graeme Blackwood,
Gonzalo Iglesias and William Byrne

Machine Translation Journal

Overview - I

- Purpose
  - An empirical study of confidence measures based on posterior probabilities of n-grams
- Contributions
  - An efficient and practical algorithm for fast computation of n-gram posterior probabilities
  - From large translation word lattices
  - Required for lattice Minimum Bayes-Risk (MBR) decoding and for confidence estimation

Overview – II

- Comprehensive evaluation for
  - Different language pairs, domains and conditions
  - Effect on reference precision of using single or multiple references
  - Computation from k-best lists vs. full evidence space of the lattice
  - Improved confidence by combination of multiple lattices in a multi-source translation framework

N-gram Posterior Probabilities

- Posterior probabilities for words have been used as a confidence measure for SMT
- This papers tries the same with n-grams
- From the probability distribution based on the translation model and language model:
  - “With what probability does an n-gram occur in the reference translations”
  - “What percentage of words in a hypothesis can be expected to occur in the reference translations?”
- Builds on the idea that high posterior probability n-grams in the maximum likelihood translation hypothesis are more likely to be found in human reference translations

Applications of N-gram Posterior Probabilities

- Interactive MT and Computer Aided Translation
  - Assign sentence level confidence estimates to hypotheses in interactive MT
- Rapidly identify parts that require correction or refinement
- Error-driven source sentence paraphrasing for better translation
- Address particular deficiencies in SMT hypotheses, such as the monolingual coverage constraints
  - Apply more sophisticated models in re-coding over low confidence regions
- Better harvest user corrections

Lattice MBR Decoding

- MBR decoding can be applied to any MT system that defines a posterior distribution over translation hypotheses
- For SMT, it has the general form:

$$\hat{E} = \arg\min_{E' \in \mathcal{E}} \sum_{E \in \mathcal{E}} L(E, E') P(E|F)$$

- Where Ɛ is some space of translation hypotheses
- L(E, E') is some loss between two hypotheses E and E'
- P(E|F) is the posterior probability of translating the source sentence F as the target sentence E

Posterior Probability

- For a log-linear model of translation:

$$P(E|F) = \frac{\exp(\alpha H(E, F))}{\sum_{E'} \exp(\alpha H(E', F))}$$

- Where H(E, F) is the score assigned by the model to sentence pair (E, F), e.g. dot product of feature weights and feature values
- The scaling factor α smooths the posterior distribution, flattening when $\alpha < 1$ and sharpening when $\alpha > 1$

Loss Function

- The linearized form of the lattice MBR decoder becomes the loss function in the earlier equation
  - With a conditional expected gain based on an approximation of BLEU score
- This gain is computed as a weighted sum of local n-gram gain functions and a constant multiplied by the sentence length:

$$\hat{E} = \arg\max_{E' \in \mathcal{E}} \left\{ \theta_0 |E'| + \sum_{n=1}^{4} \sum_{u \in \mathcal{N}_n} \theta_n \#_u(E') p(u|\mathcal{E}) \right\}$$

- Where N_n is the set of n-grams (of order n) in the lattice
- #u (E') is the number of times the n-gram u occurs in hypothesis E' and parameters θ are constants estimated over the data

Path Posterior Probability of N-gram

- The quantity p(u|Ɛ) is the path posterior probability of the n-gram *u :*

$$p(u|\mathcal{E}) = \sum_{E \in \mathcal{E}} \delta_u(E) P(E|F) = \sum_{E \in \mathcal{E}_u} P(E|F)$$

- That is, over the subset of paths containing the n-gram u at least once
- Note that posterior probability is different from the expected count (it is accumulated once per path)
- It is possible to extract and enumerate all these n-grams exactly
  - Whereas it is usually impossible to enumerate all paths
- While linearisation of the gain function is an approximation, it can be computed exactly even for very large lattices

Efficient posterior probability computation

- From translation lattices, having the form of a directed acyclic graph
- Word sequences and scores of translation hypotheses are encoded in the lattice as a Weighted Finite State Transducer
- It is particularly efficient in its representation of translation hypotheses, and thus for posterior probability computation
- Previous approaches using WFSA can be slow over large lattices with many n-grams
  - As they may involve separate intersection and summation over matching paths for each n-gram in the lattice

Efficient posterior probability computation (Cond.)

- The efficient algorithm presented is based on a forward procedure that allows fast and exact computation
- A lattice specialization of the hypergraph vector-indexed algorithm
- The typical forward procedure calculates forward probabilities $\alpha(q)$: The marginal probability of the partial paths which lead from the start state to state q
- The modified forward procedure calculates quantities $\alpha(q, u)$: The marginal probabilities of the paths which lead to state q and that pass through at least one arc with the input symbol u
- It can be seen as a modified form of marginalization, rather than a counting procedure

Efficient posterior probability computation (Cond.)

- The modified forward procedure can be extended to marginalize probabilities over paths which contain n-grams
- However, it is easier first to transduce word lattices to n-gram lattices and then use the modified forward procedure simply count individual n-gram tokens
- The order-n mapped lattice $\mathcal{E}_n$ is obtained by composing the word lattice $\mathcal{E}$ with the mapping transducer Φ_n

$$\mathcal{E}_n = \min(\det(\mathrm{rmeps}(\Pi_2(\mathcal{E} \circ \Phi_n))))$$

  The resulting acceptor $\mathcal{E}_n$ is a compact lattice of n-gram sequences of order-n consistent with the hypotheses and scores of the original lattice $\mathcal{E}$
- The path labeled with the words of a hypothesis has the weight P(E | F)

Algorithm

COMPUTE-NGRAM-POSTERIORS

for each state $q \in Q$ $\triangleright$ In topologically sorted order
  do for each edge $e \in E[q]$
    do $\alpha(n[e]) \leftarrow \alpha(n[e]) + (\alpha(q) \times w[e])$
      if $i[e] \notin \mathcal{N}_{n[e]}$
        then $\mathcal{N}_{n[e]} \leftarrow \mathcal{N}_{n[e]} \cup \{i[e]\}$
      $\alpha(n[e], i[e]) \leftarrow \alpha(n[e], i[e]) + (\alpha(q) \times w[e])$
      for each n-gram $u \in \mathcal{N}_q$ **where** $u \neq i[e]$
        do if $u \notin \mathcal{N}_{n[e]}$
          then $\mathcal{N}_{n[e]} \leftarrow \mathcal{N}_{n[e]} \cup \{u\}$
        $\alpha(n[e], u) \leftarrow \alpha(n[e], u) + (\alpha(q, u) \times w[e])$
  if $q \in F$
    then for each n-gram $u \in \mathcal{N}_q$
      do $p(u|\mathcal{E}) \leftarrow p(u|\mathcal{E}) + (\alpha(q, u) \times \rho[q])$
  $\mathcal{N}_q \leftarrow \emptyset$ $\triangleright$ Clean up state q

Mapping Transducer for N-grams

words	bigram
$w_1 w_1$	u_1
$w_2 w_2$	u_2
$w_1 w_2$	u_3
$w_2 w_1$	u_4

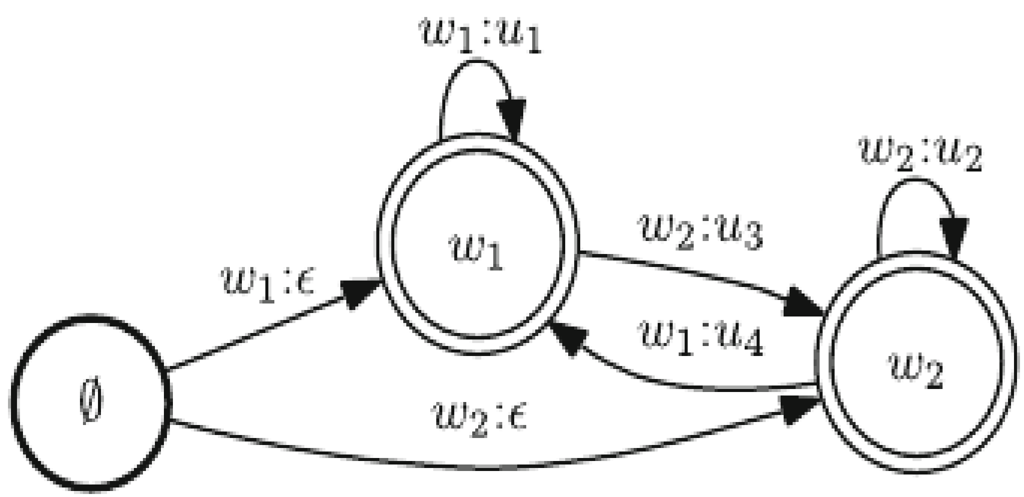


Fig. 2 Mapping transducer Φ_2 for all possible bigrams $\Sigma_2 = \{u_1, u_2, u_3, u_4\}$ formed from unigram alphabet $\Sigma_1 = \{w_1, w_2\}$. *States* and *arcs* need only be added for bigrams $u \in \mathcal{N}_2$

Predictive Power of N-gram Posterior Probabilities

- Analyze the relation between posterior probability and translation quality by computing:
  - The precision of high posterior n-grams with respect to the human reference translations available for each source sentence
  - The translation hypothesis coverage of high posterior n-grams
  - The converse precision of low posterior n-grams with respect to the human references
  - The precision of high posterior n-grams in a system combination scenario

Posterior Probability Reference Precisions

$$\mathcal{P}_{n,\beta} = \frac{|\mathcal{R}_n \cap \mathcal{N}_{n,\beta}|}{|\mathcal{N}_{n,\beta}|}$$

- The precision at order n for threshold β is the proportion of n-grams in N(n, β) also present in the references
- Rn is the set of n-grams of order n in the union of references

Posterior Probability Hypothesis Coverage

- How many words in the top hypothesis are covered by N(n,β) at each confidence threshold β
- The coverage at order n for threshold β is the proportion of hypothesised words covered by n-grams in N(n,β):

$$\mathcal{C}_{n,\beta} = \frac{100 * |\mathcal{W}_{n,\beta}|}{I - n + 1}$$

- Where I is the length of the ML translation 1-best hypothesis
- W(n,β) is the set of words in the hypothesis that belong to n-grams of order n with posterior probability greater than or equal to β
- Can be extended to k-best list or lattice

Posterior Probability Converse Reference Precisions

- The converse precision at order n for threshold γ is the proportion of n-grams in N(n,γ) that are not present in the references

$$\mathcal{Q}_{n,\gamma} = \frac{|\mathcal{N}_{n,\gamma} \setminus \mathcal{R}_n|}{|\mathcal{N}_{n,\gamma}|}$$

- Tests the ability of the posteriors to indicate how reliable the portions of translation are
- Ideally, low posteriors should be as informative as high posteriors

System Combination Reference Precisions

- The effect on reference precision of computing n-gram posterior probabilities from a combination of multiple translation lattices in the context of multi-input and multi-source translation

$$p_i(u|\mathcal{E}^{(i)}) = \sum_{E \in \mathcal{E}_u^{(i)}} P(E|F)$$

- Treating each lattice as a WFSA, the evidence space is the union of M individual lattices
- We sum over all paths in each lattice with one or more occurrence of the n-gram *u*
- We compute the n-gram confidence p(u|Ɛ) as a weighted combination (sum or product) of the probabilities from individual lattices
- Weights should reflect qualities of various systems, e.g. using grid search over parameters based on optimal BLEU score

System Development

- Arabic→English
- Chinese→English
- French→English
- Spanish→English
- English→Spanish

MBR Decoding Efficiency

Table 4 Average time (s/sentence) to compute n-gram path posterior probabilities using the sequential method, path counting transducers, and symbol-specific forward algorithm

	Arabic→English		Chinese→English	
	mt0205tune	mt0205test	tune.nw	tune.web
Sequential	1.52	1.62	4.43	4.73
Transducers	0.84	0.88	1.68	1.69
Symbol-specific	0.13	0.14	0.41	0.40

MBR Decoding Efficiency (Cond.)

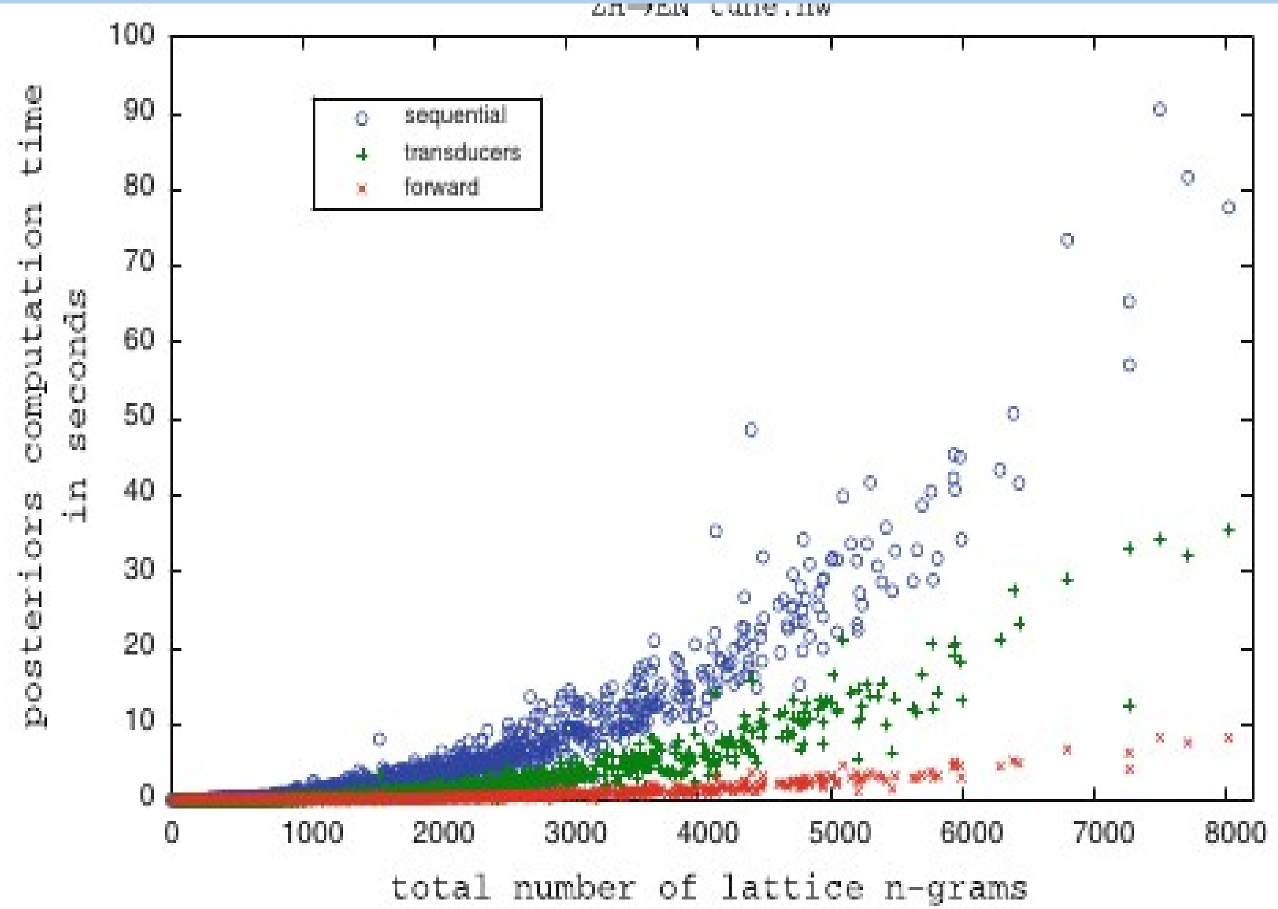
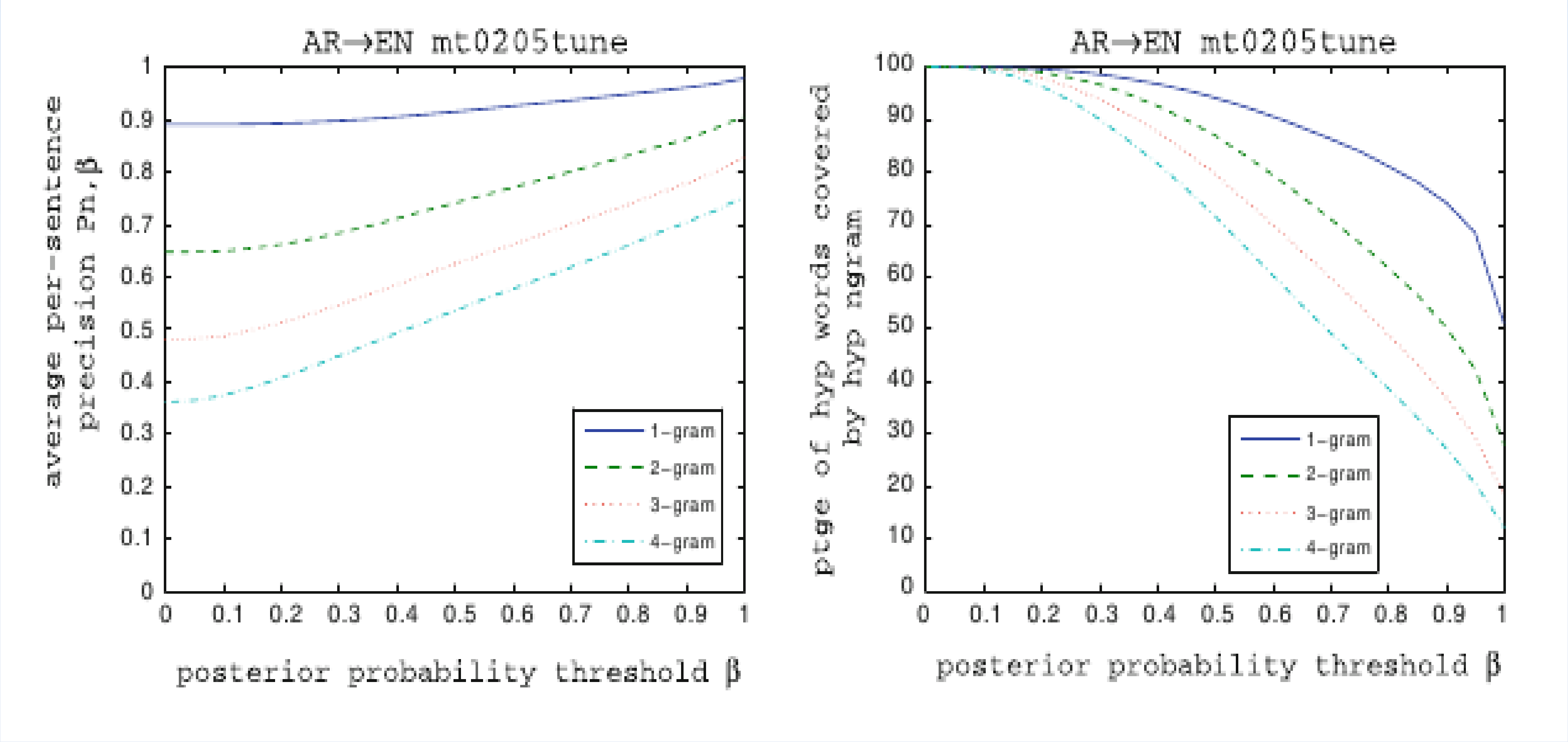
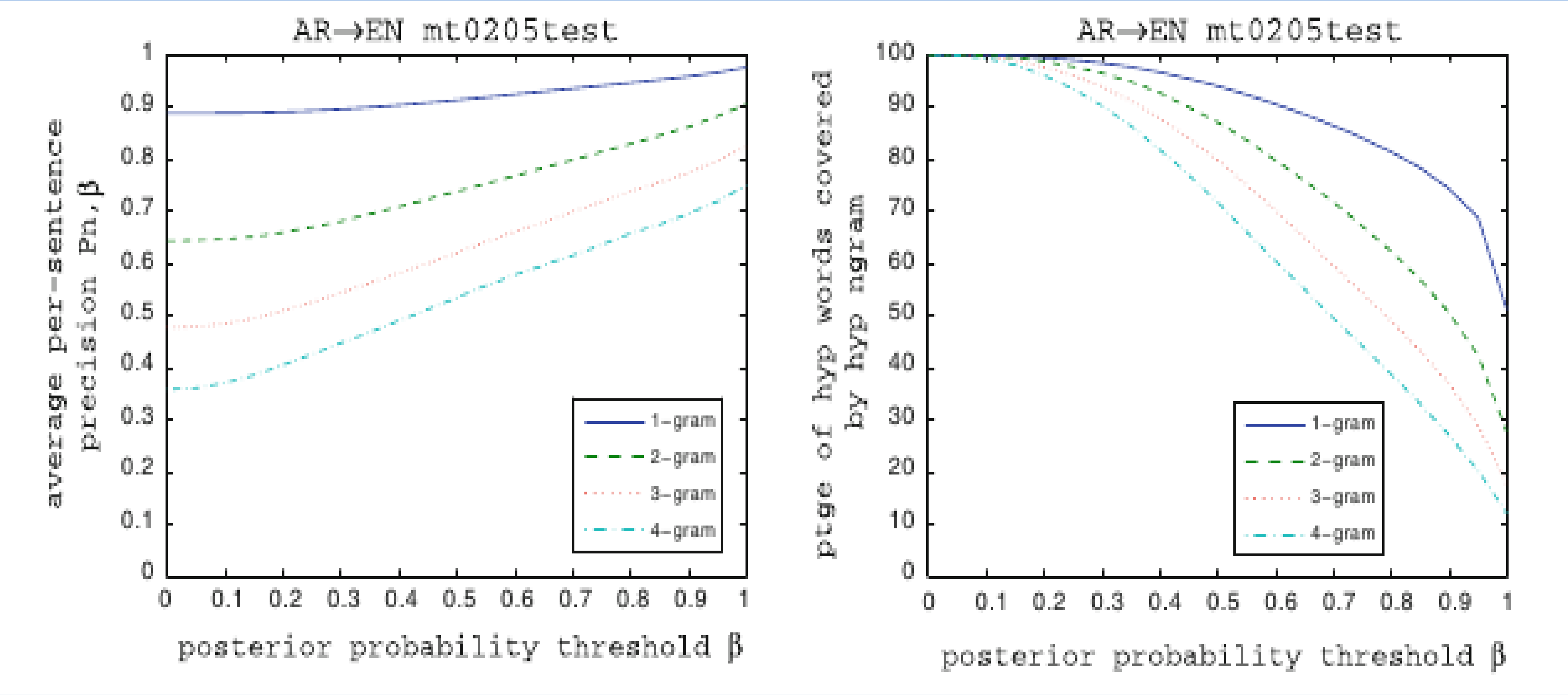


Fig. 3 Posterior probability computation time (s) versus # of lattice n-grams using the sequential method, path counting transducers, and symbol-specific forward algorithm for each sentence of the Chinese→English tune.nw testset

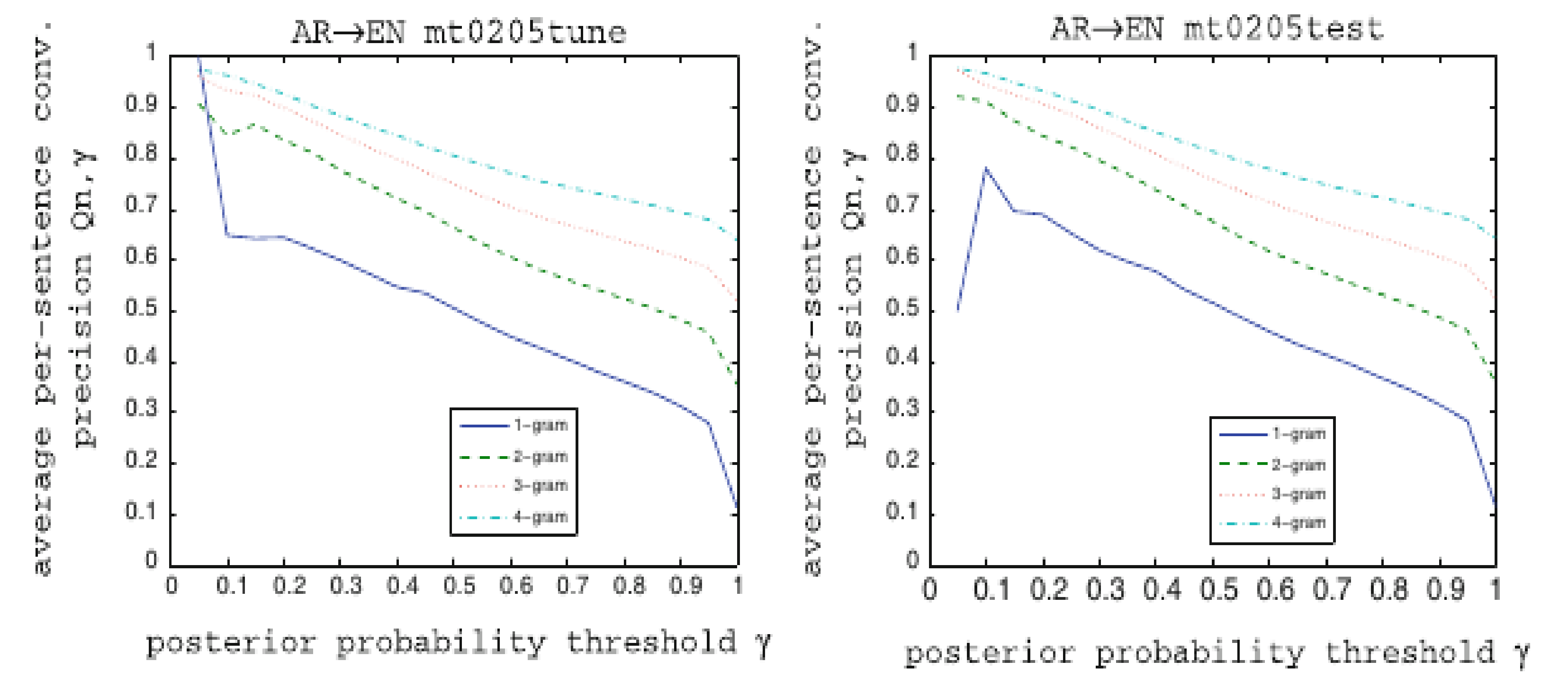
Precision and Coverage



Precision and Coverage (Contd.)

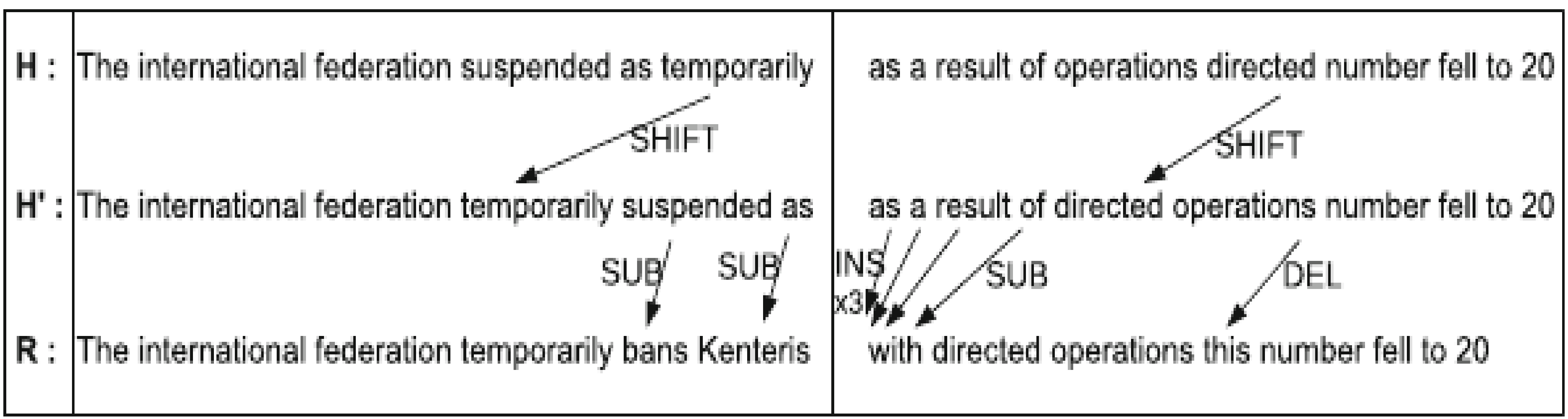


Converse Precision

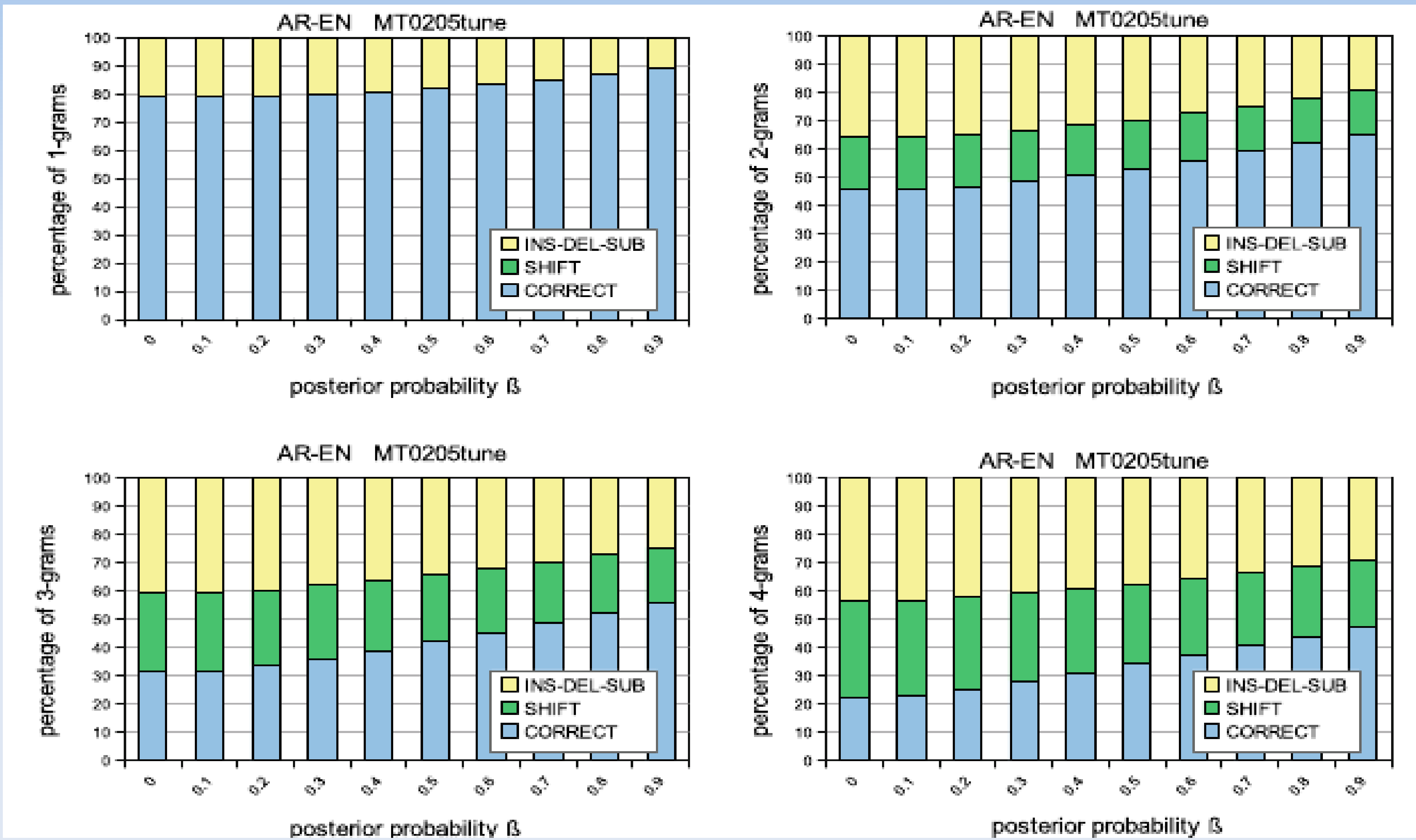


Translation Edit Rate (TER)

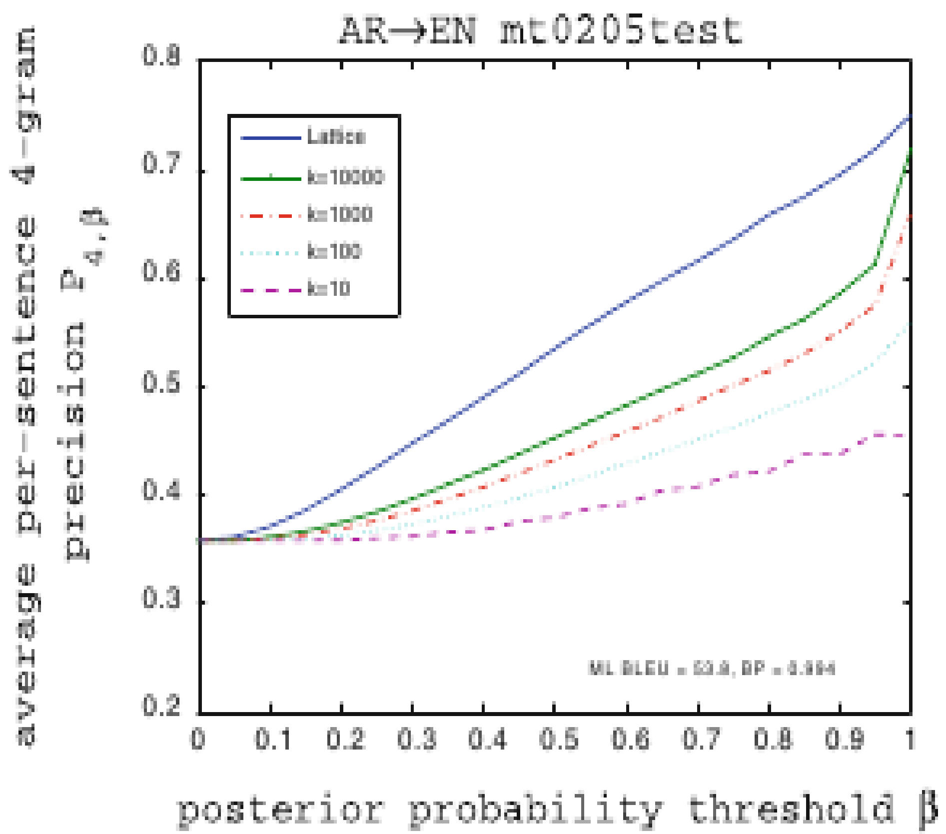
$$\mathrm{TER} = \frac{\#\mathrm{Dels} + \#\mathrm{Ins} + \#\mathrm{Subs} + \#\mathrm{Shifts}}{\#\mathrm{Words\ in\ Ref}}$$

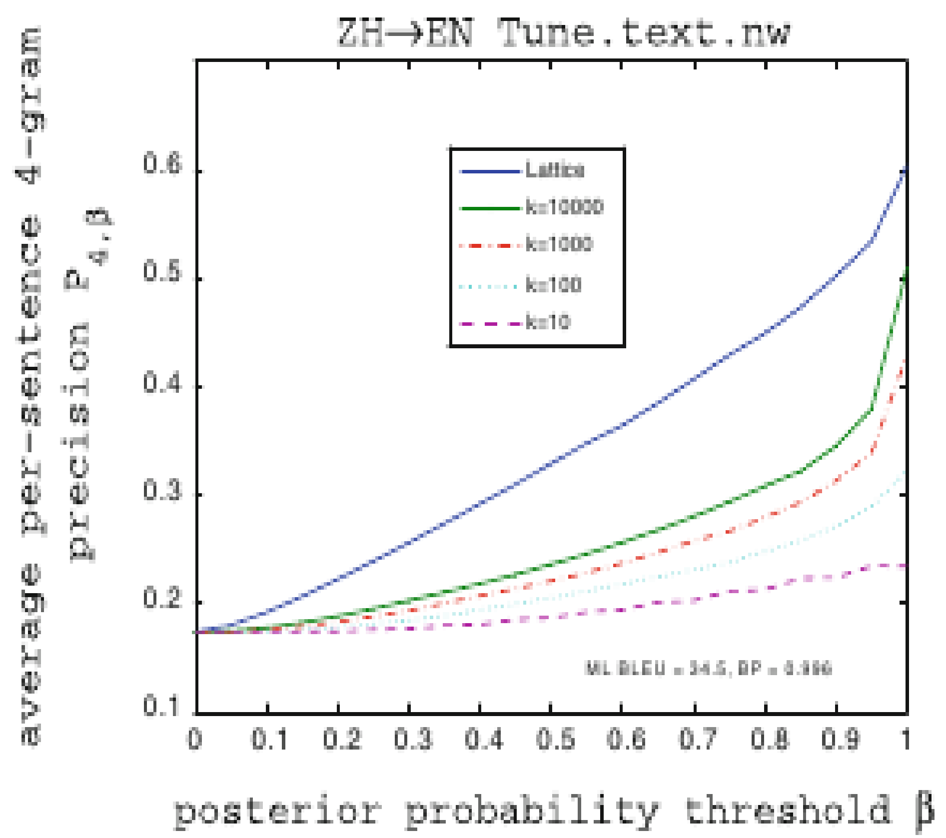


Evaluation in Terms of TER



Evidence Space Size and Reference Precisions



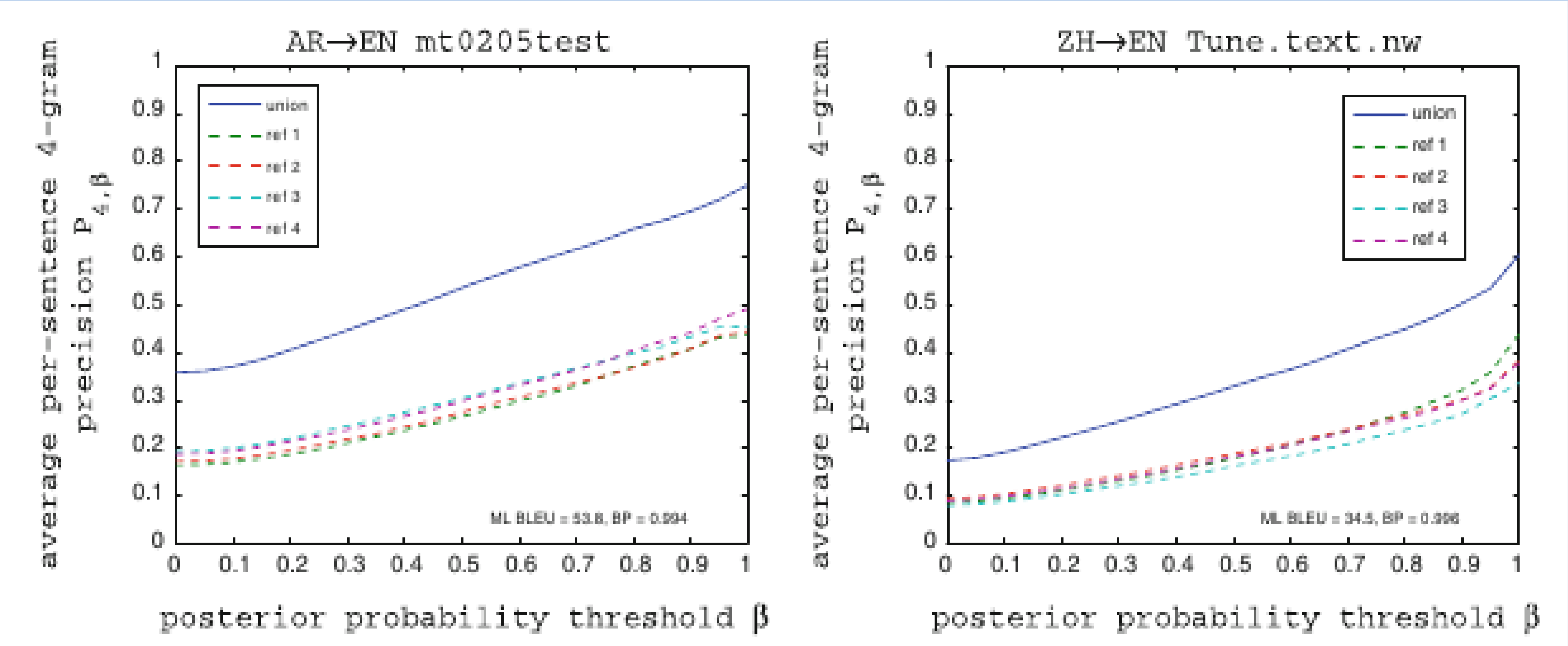


Missing Probability Mass from k-best Lists

k	mt0205tune	mt0205test
1,000	24.41	24.91
10,000	13.96	14.27
20,000	11.73	12.00
50,000	9.30	9.52
100,000	7.78	7.98

Arabic - English

Single vs. Multiple References

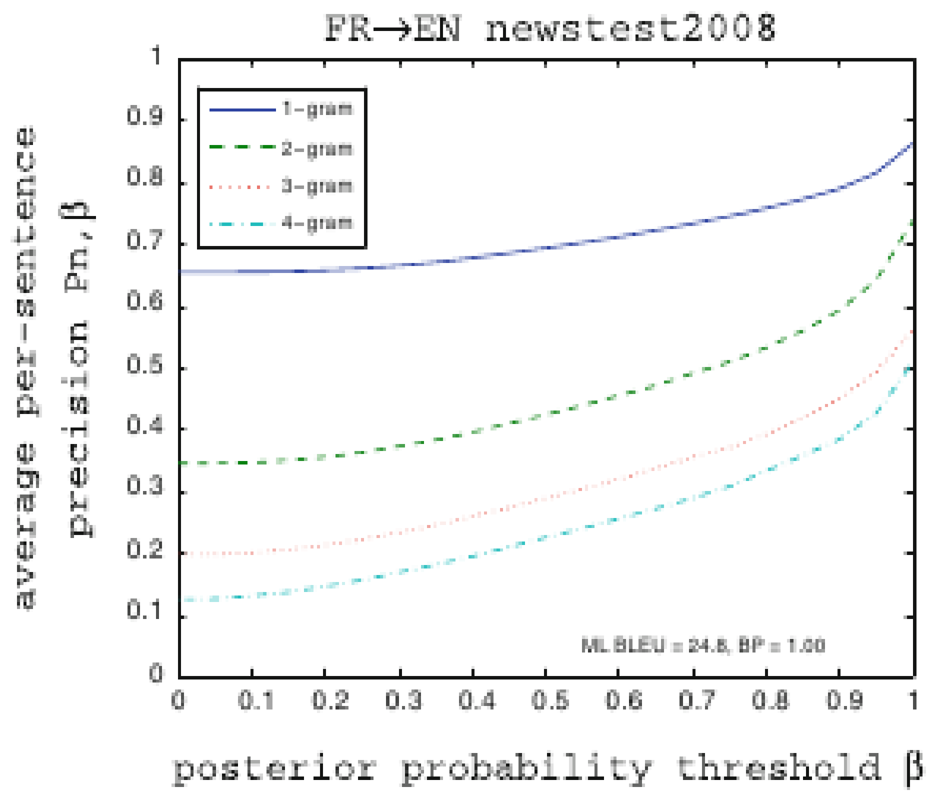


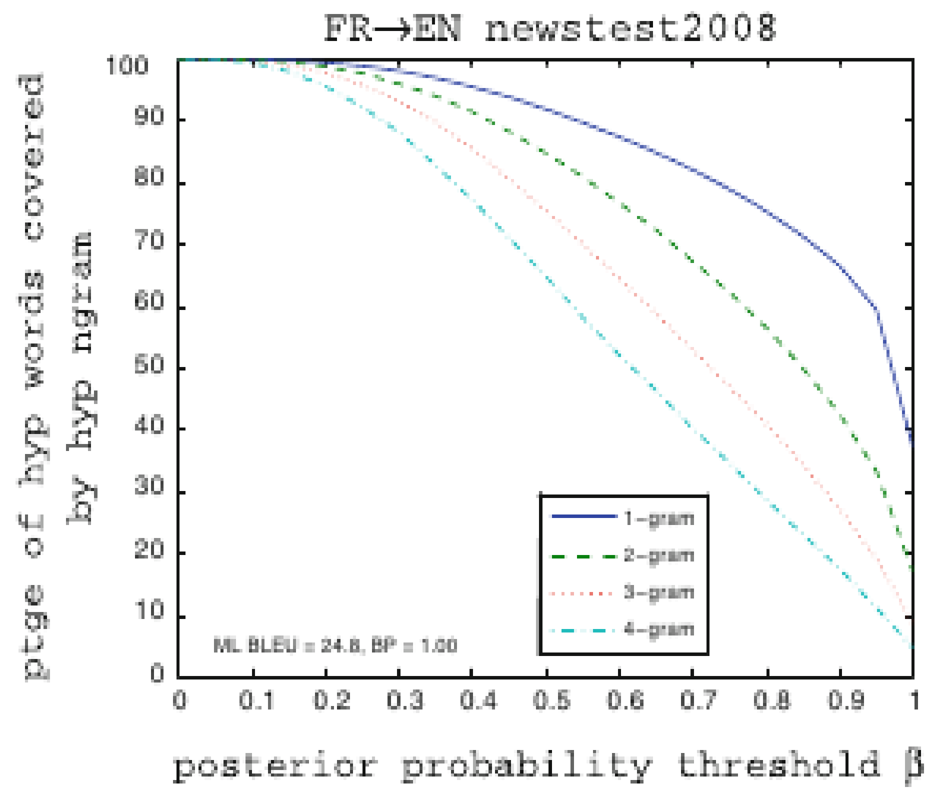
Confidence-based Hypothesis Segmentation

the newspaper “ constitution ” quoted brigadier abdullah krishan **, the chief of police in** karak governorate **(521 km south @-@ west of amman) as saying that the seizure** took place after **police received information that** there were attempts by the group to sell for more than $ 100 thousand dollars , **the police rushed to** the arrest in possession .

- High-confidence sub-sequences correspond to partial hypotheses for which there is consensus amongst the translations in the first-pass evidence space
- High-confidence subsequences are often of higher quality than low-confidence subsequences
- Shows how n-gram posterior probability confidence measures can be used to identify low-confidence portions of translation hypotheses that may benefit from re-decoding, post-processing, targeted application of specific models, or user input in an interactive translation setting

Confidence-based Hypothesis Segmentation (Contd.)





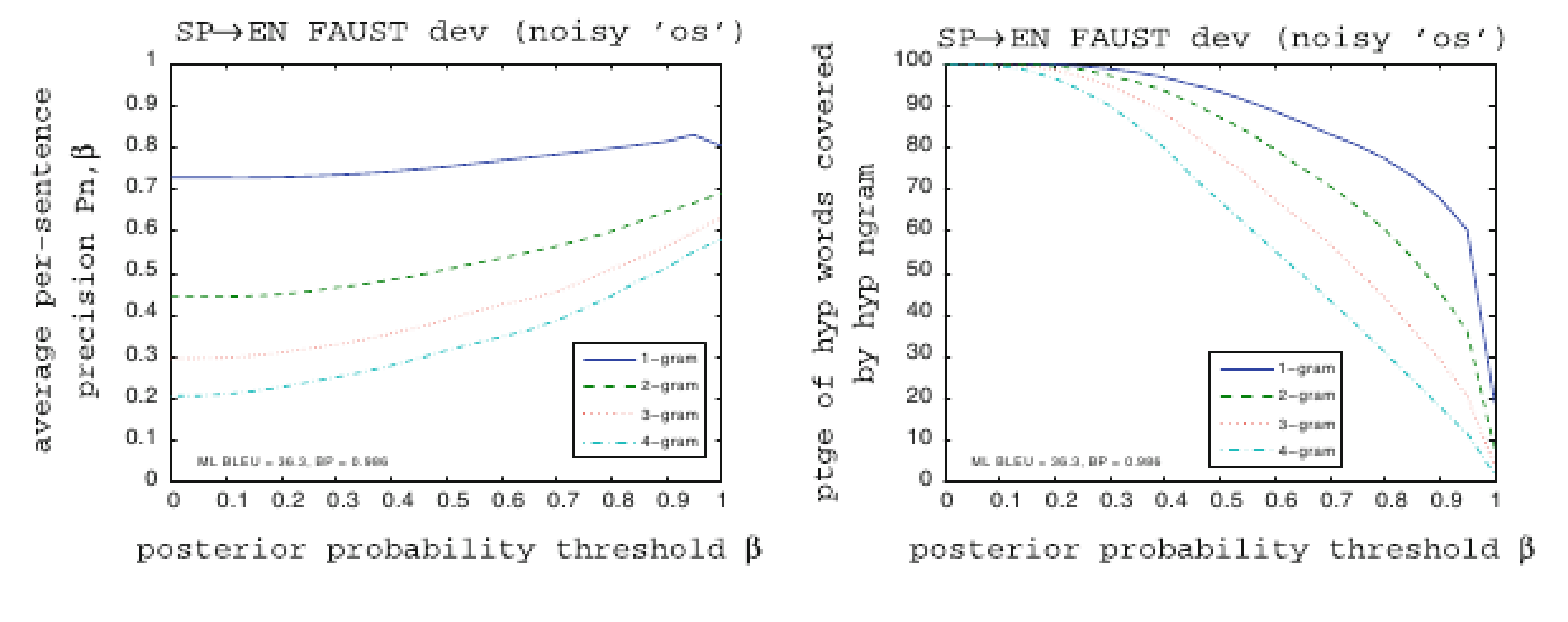
Evaluation on FAUST Data

- More like real life data and based on actual user interaction
- Shows roughly similar results
- There is a difference, however, between translating from clean data and 'noisy' data
- Precision, converse precision and coverage are good metrics for this purpose
  - As is TER, in a different way

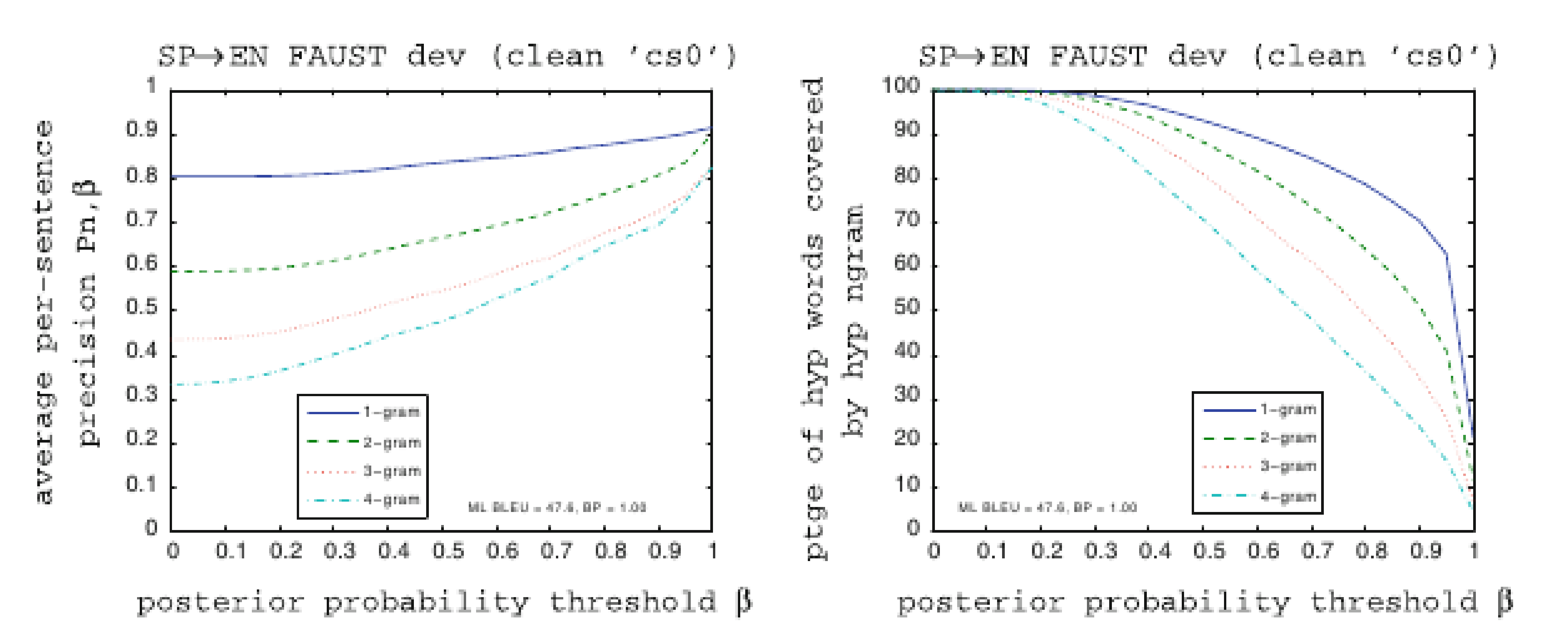
Translating from Clean vs. Noisy Data

	noisy 'os'		clean 'cs0'		clean 'cs1'	
	dev	test	dev	test	dev	test
HiFST	36.3	35.9	47.6	46.9	45.9	45.9
+LMBR	36.2	35.9	48.6	47.9	47.1	46.7

Precision and Coverage on Noisy Data



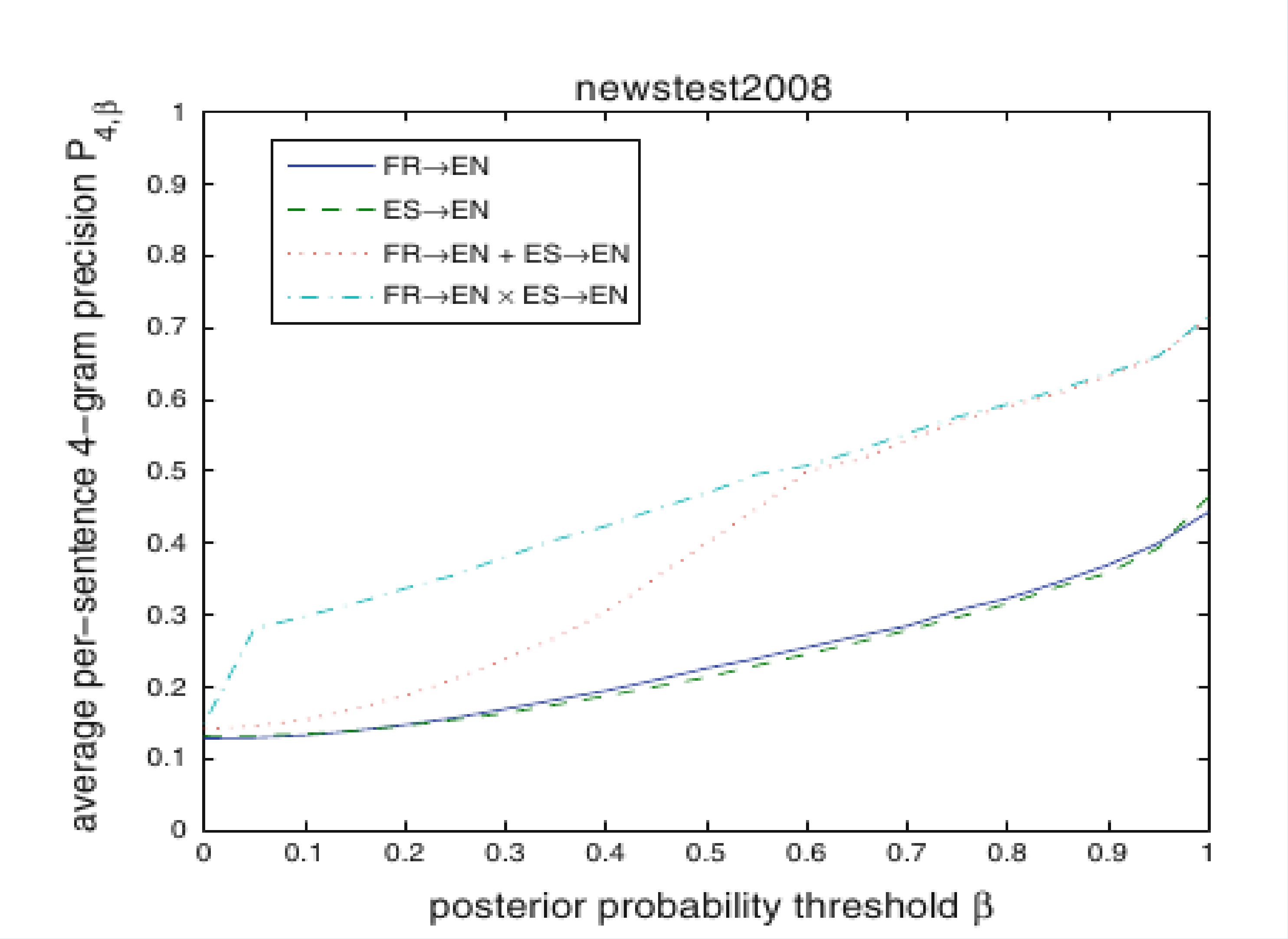
Precision and Coverage on Clean Data



Multi-source Translation

- Multi-source translation is possible whenever the source-language sentence is available in multiple languages
- The motivation is that some of the ambiguity that must be resolved in translating between one pair of languages may not be present in a different pair

Multi-source Translation Confidence



Conclusions

- N-gram posterior probabilities are good estimates of translation quality
- There is an efficient method to calculate them
- Precision, converse precision and coverage are good metrics for this purpose
  - As is TER, in a different way
- Using the full lattice space helps, rather than increasing the size of the k-best list
- More references help
- Multiple source translation helps
- Cleaning the source data helps too