
Sentence-level features for MT confidence estimation

Confidence Estimation- WS 2003

Center for Language & Speech Processing

Johns Hopkins University

Feature Classification Types

- Specific to base MT model
 - May describe or include method (pruning steps, alignment template usage)
 - May describe typical generation results (average hyp length, average density)
- Feature dependencies
 - Source sentence only
 - Target sentence only
 - Combination source and target sentences

Classification Overview

- 91 sentence-level features developed
 - 24 directly from MT system base model
 - 13 alignment-template based
 - 34 model-independent
 - 23 source sentence only
 - 38 target sentence only
 - 30 source and target both (13 from AT)
 - 9 generated only for top 1000 hypotheses

Base model features

13 types, 24 features

- Rank in N-best list (*Nicola*)
- Model base score and **feature scores** (*George*)
- N-best list length (N)
- Average Hypothesis Length (N)
- Overall Hypothesis Density (# tgt words/ src length) (N)
- Ratio of hypothesis model score to best model score (N)
- Ave. # of active hypotheses post-pruning (N)
- Ave # of hypotheses pruned / cardinality (N)
- Ave best hyp score pre-pruning (N)
- Ave worst hyp score pre-pruning (N)
- Ave worst hyp score post-pruning (N)

Base model features

13 types, 24 features

- Rank in N-best list (*Nicola*)
- Model base score and **feature scores** (*George*)
- N-best list length (N)
- Average Hypothesis Length (N)
- Overall Hypothesis Density (# tgt words/ src length) (N)
- **Ratio of hypothesis model score to best model score** (N)
- Ave. # of active hypotheses post-pruning (N)
- Ave # of hypotheses pruned / cardinality (N)
- Ave best hyp score pre-pruning (N)
- Ave worst hyp score pre-pruning (N)
- Ave worst hyp score post-pruning (N)

Alignment Template-based features

2 types, 13 features

■ **Consistent alignments** (*Alberto*)

- Ave # times each AT is aligned with the same source word throughout the N-best list

Also a sub-sentential feature; more info later

■ **AT-based N-gram language model (A)**

Trained on each N-best list

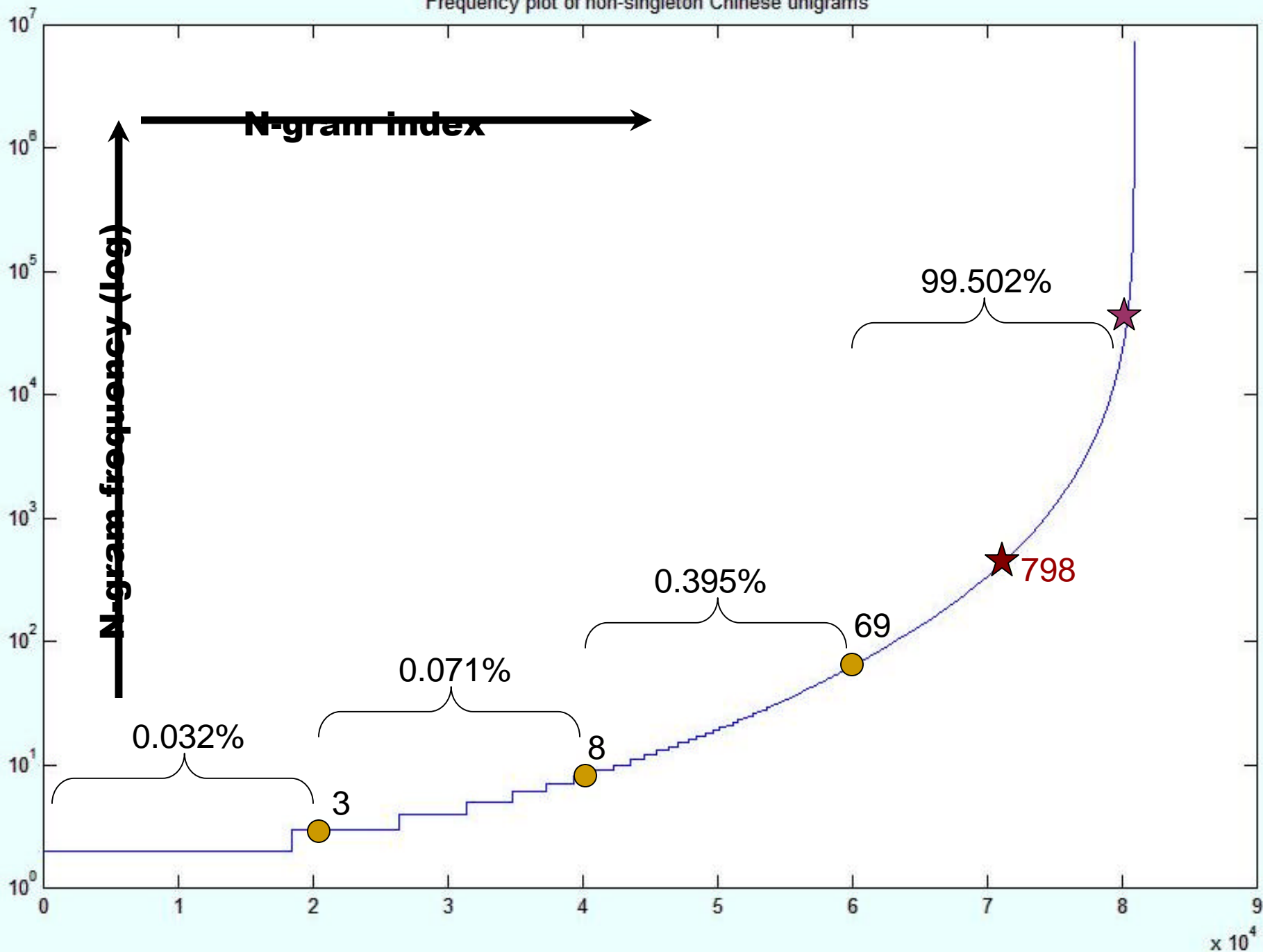
- Sentence scores and perplexities (*no smoothing*)
- “Vocab” of ATs / ave target length
- “Vocab” of ATs / ave source length

Source-based features

11 types, 23 features

- Source sentence length (*Simona*)
- **N-gram LM score** (*Erin*)
 - Trained on Chinese half of parallel corpus*
 - SRI Toolkit: trigram LM, Kneser-Ney discounting
 - Log-probabilities and perplexities determined
- **N-gram frequency ranges** (*E*)
 - Frequency quartile designation for 1- ,2- ,and 3-grams

Frequency plot of non-singleton Chinese unigrams



Target-based features

9 types, 38 features

- Ave and individual hyp length (*Nicola, Simona*)

- **Word-based N-gram LM** (*Alberto*)

Trained for each N-best list

- Sentence scores and perplexities
- N-best vocab / ave target length
- N-best vocab / ave source length

- **N-gram LM score** (*Erin*)

Trained on English half of parallel corpus

- SRI Toolkit: trigram LM, Kneser-Ney discounting
- Log-probabilities and perplexities determined

Target-based features

9 types, 38 features

- P's&Q's- parentheses and quotation mark check (*E*)
- Word features averaged over sentence (*Nicola*)
Also a sub-sentential feature; more info later
- **Center hypotheses** (*Erin*)
Unweighted Levenshtein dist bw top 1K hyps
 - Levenshtein distance from the center hypothesis
 - Percentage of hypotheses less than some threshold away from hyp

Source- & target-based features

9 types, 30 features

- Ratio of src length to target length (*Simona*)
- **Alignment monotonicity** (S)
- Number of source words not “aligned” (S)

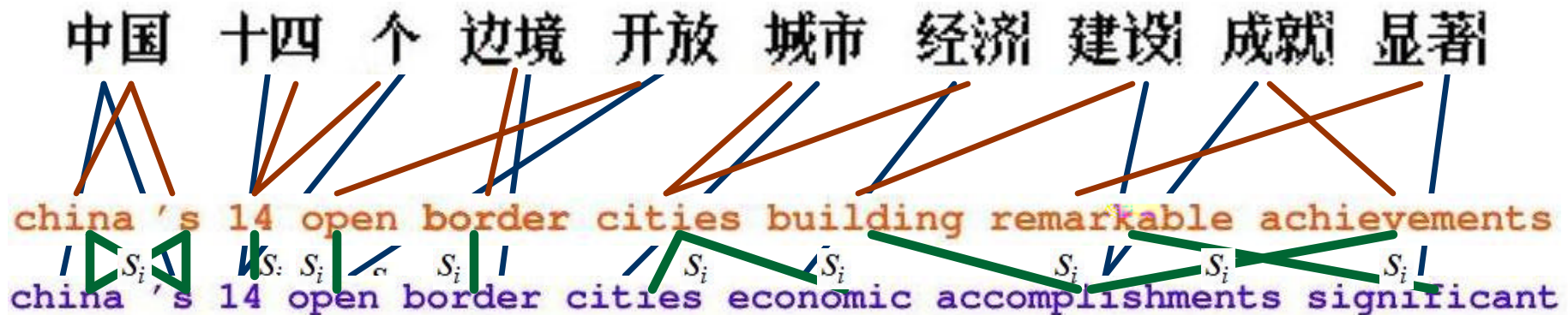
- **IBM Model 1 scores** (*Erin*)
 - Two models, trained for $C \rightarrow E$ and $E \rightarrow C$

Source- & target-based features

9 types, 30 features

■ Semantic Similarity (*John*)

- Number of distinct word-alignments in N-best list
- Used WordNet to estimate semantic similarities between words and between hyps [Banerjee & Pedersen 2002]



Next: Results
