Feasibility of Human-in-the-loop Minimum Error Rate Training

Omar Zaidan and Chris Callison-Burch

Department of Computer Science
Johns Hopkins University

March 24, 2010

Outline

Tuning MT Systems

- Machine translation systems use various features to score translation candidates
- Feature weights tuned through MERT using automatic metrics
- Systems evaluated by humans, metrics should reflect human judgments (adequacy, post-editing effort)
- Correlation of BLEU metric with adequacy judgments: 0.61 (MetricsMATR08)

Evaluation for Tuning

- Judgments too expensive, time consuming to collect for N-best lists in MERT
- Fast, reliable (Papineni et al., 2002) BLEU metric used as stand-in:

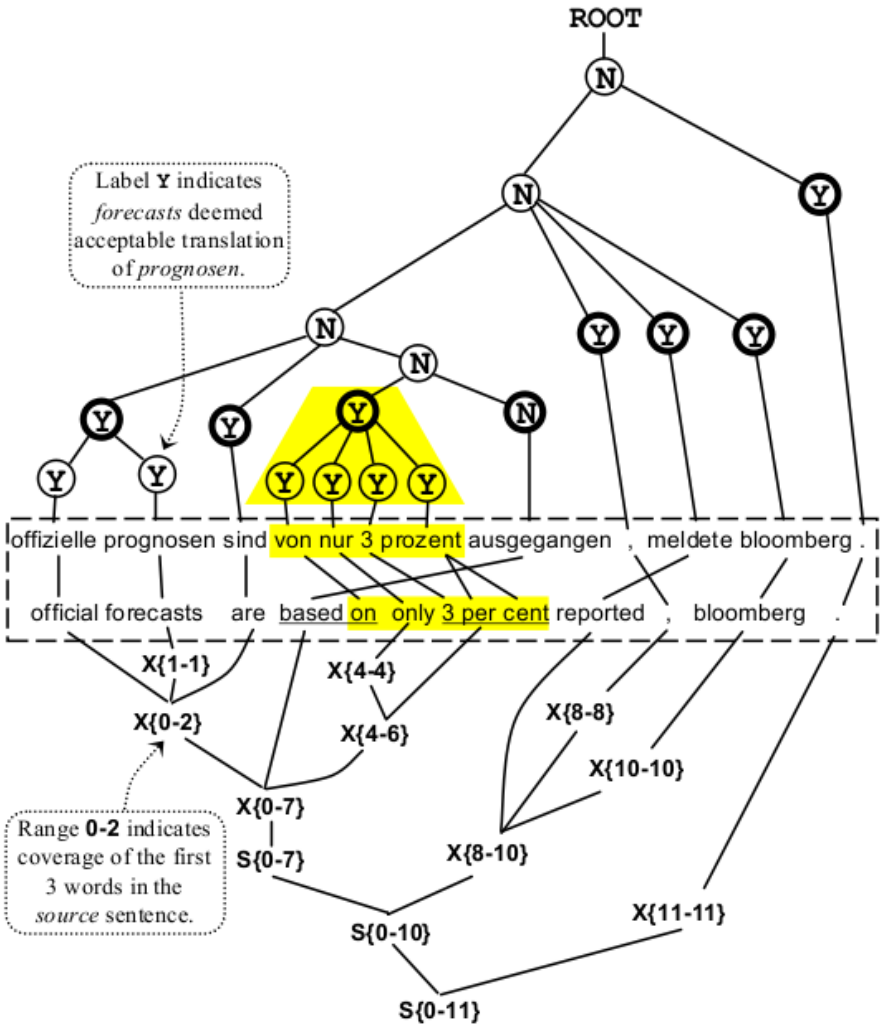
$$BLEU = BP(len(h), len(r)) \cdot \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log p_n\right)$$

- Recent work discusses weaknesses of BLEU (Banerjee and Lavie 2005; Chiang et al., 2008; Callison-Burch et al., 2008; Przybocki et al., 2008; Callison-Burch et al., 2009)

The RYPT Metric

- Goal: metric based on human judgments, also feasible for MERT
- Idea: collect database of sub-sentential human judgments, use to automatically score translation candidates
- Parse source sentence, reward constituents correctly translated, penalize constituents incorrectly translated
- RYPT: **R**atio of **Y**ES nodes in **P**arse **T**ree

RYPT Example



Collecting Human Judgments

Assumption I: substring judgments can be reused for multiple candidates with same source:

- Judgment form:
  <source substring, target substring, judgment>

- Original pair (`YES`):
  der patient wurde isoliert .
  the patient was isolated .

- Other candidates (`YES`):
  the patient isolated .
  the patient was in isolation .
  the patient has been isolated .

Collecting Human Judgments II

Assumption II: judgments can be percolated up/down parse tree:

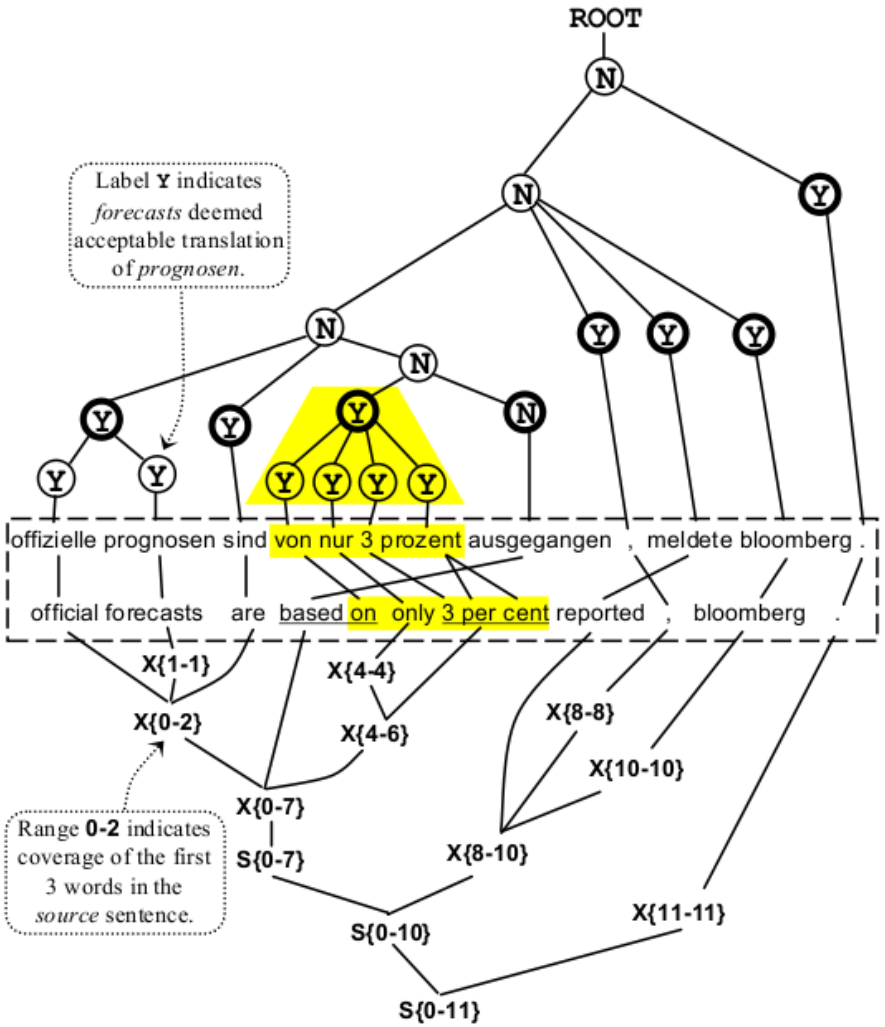
- Node labeled `NO`: ancestors likely labeled `NO`
- Node labeled `YES`: descendants likely labeled `YES`

Tradeoff: approximate actual judgments to dramatically reduce number of human judgments required

Source-Candidate Alignments

- Need to align structure of translation candidate to parse of source sentence
- Use a hiero-style decoder (Joshua): output derivations, associated spans in source sentence
- Deduce new phrase alignments by discarding source words in other phrase alignments
- For many-to-many alignments, use word alignments from training corpus

RYPT Example



Data Collection

Data set:

- 250 segments from WMT08 German-English news
- Candidate translations from last iteration of MERT

Select substrings to judge:

- Choose segments covered exactly by parse subtree
- Maximize amount of `YES`/`NO` percolation
- Idea: select substrings that fully cover source, do not overlap

Substring selection

Identify ideal "frontier" nodes:

1. Select `maxLen` for source segments
2. Starting at root node, propagate "frontier" set such that:
   a. Set of nodes fully covers sentence
   b. No nodes have overlapping subtrees
   c. No node covers more than `maxLen` words

- Allows full downward-`YES` and upward-`NO` propagation
- Greatly reduces number of judgments required

Collecting Judgments

Judgments collected using Amazon Mechanical Turk

- Task (HIT) given to users:

  You are shown a "source" German sentence with a highlighted segment, followed by several candidate translations with corresponding highlighted segments. Your task is to decide if each highlighted English segment is an acceptable translation of the highlighted German segment.

- Possible choices: `YES`, `NO`, `NOT SURE`
- Users shown up to 10 translation candidates

Evaluating RYPT

Experiment design: RYPT vs BLEU

- For each N-best list, extract BLEU 1-best and RYPT 1-best
- Present both to human, have human select more adequate translation
- Obtain 3 judgments per translation pair
- (Essentially a special case of ranking task)

Results

	References shown; unrestricted		References not shown; restricted to DE workers	
Preferred	judgments	% judgments	judgments	% judgments
RYPT	346	46.1	113	45.2
BLEU	270	36.0	73	29.2
Neither	134	17.9	64	25.6
Total	750	100.0	250	100.0

Table: Ranking comparison results

Analysis of Data Collection

Explore impact of assumptions:

- Collect *complete* set of judgments for 50 source sentences
- Vary `maxLen` from 1 to 7
- Collect 5 judgments per node
- 68.9% of nodes: at least 4 of 5 judges agree
- Use data to explore coverage and accuracy after percolation

Analysis of Data Collection II

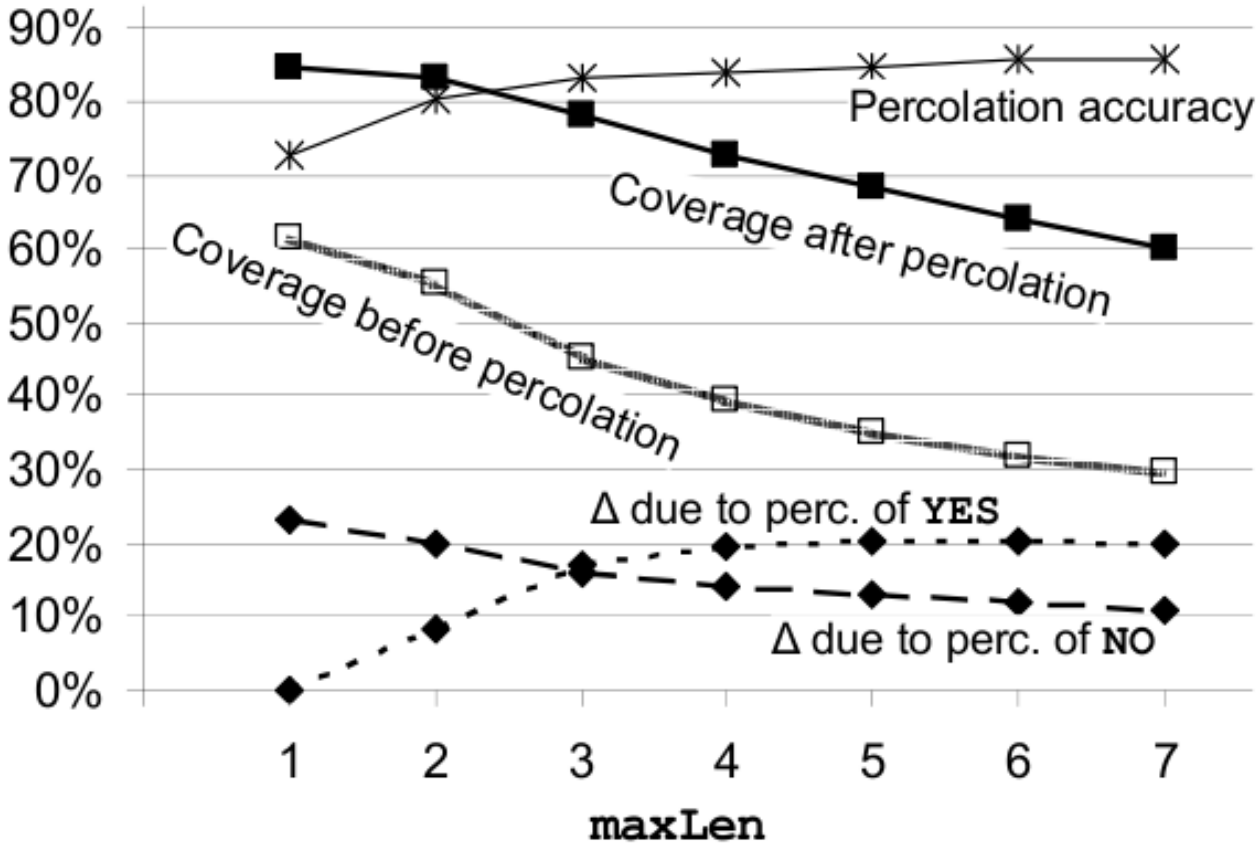


Figure: Results of label percolation for various `maxLen`

Discussion Points

- Collection of judgments for phrases using M-Turk
- Design of RYPT vs BLEU evaluation

MetricsMATR08 Comparison

MaxSim	ULCh	ULCopt	Meteor-v0.7
TERp	METEOR-v0.6	SNR	METEOR-ranking
LET	NIST-v11b	DP-Or	CDer
BLEU-1	EDPM	SEPIA2	ATEC2
ATEC1	ATEC4	SVM-Rank	SEPIA1
ATEC3	RTE-MT	BleuSP	DR-Or
SR-Or	4-GRR	RTE	BLEU-v12

Table: Metrics outperforming BLEU-4 in MetricsMATR08

- BLEU-4 ranks 29 of 39 in single-reference segment-level ranking task
- Pearson's r of BLEU-4 with human ranking judgments: 0.26