

Confidence Estimation for Statistical Machine Translation

John Blatz–Princeton, Erin Fitzgerald–JHU,
George Foster–U de M, Simona Gandrabur–U de M,
Cyril Goutte–XRCE, Alex Kulesza–Harvard
Alberto Sanhchis–UPV, Nicola Ueffing–RWTH,

Confidence Estimation for SMT

- goal: attempt to determine whether *outputs* of the base SMT system(s) are *correct*
- output: various granularities (sentence, sub-sentence)
+ context
- correctness: difficult to evaluate MT correctness → adopt a more flexible application-specific definition of *usefulness*

CE for SMT Workshop Overview

- sentence level CE
 - goal: estimate “correctness” probabilities for each alternative translation in nbest list
 - time-span: weeks 1-6
- sub-sentence level CE
 - goal: estimate “correctness” probabilities for partial translations: words and low-order ngrams
 - time-span: weeks 3-6

Two Week Report – Goals

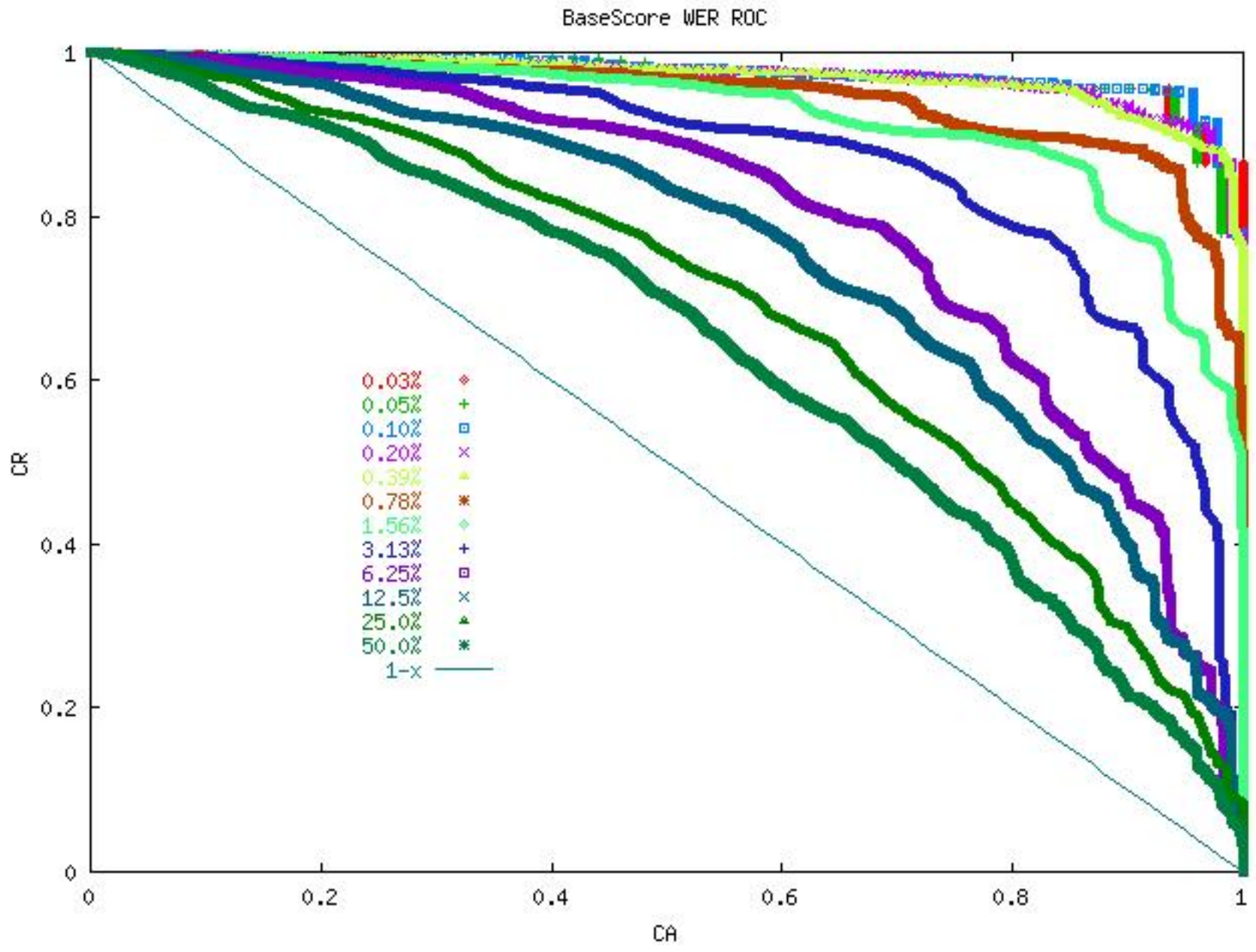
- set-up
 - CE-framework
 - nbest preprocessing
- sentence level confidence feature generation
- ML experiments
- human evaluation experiment

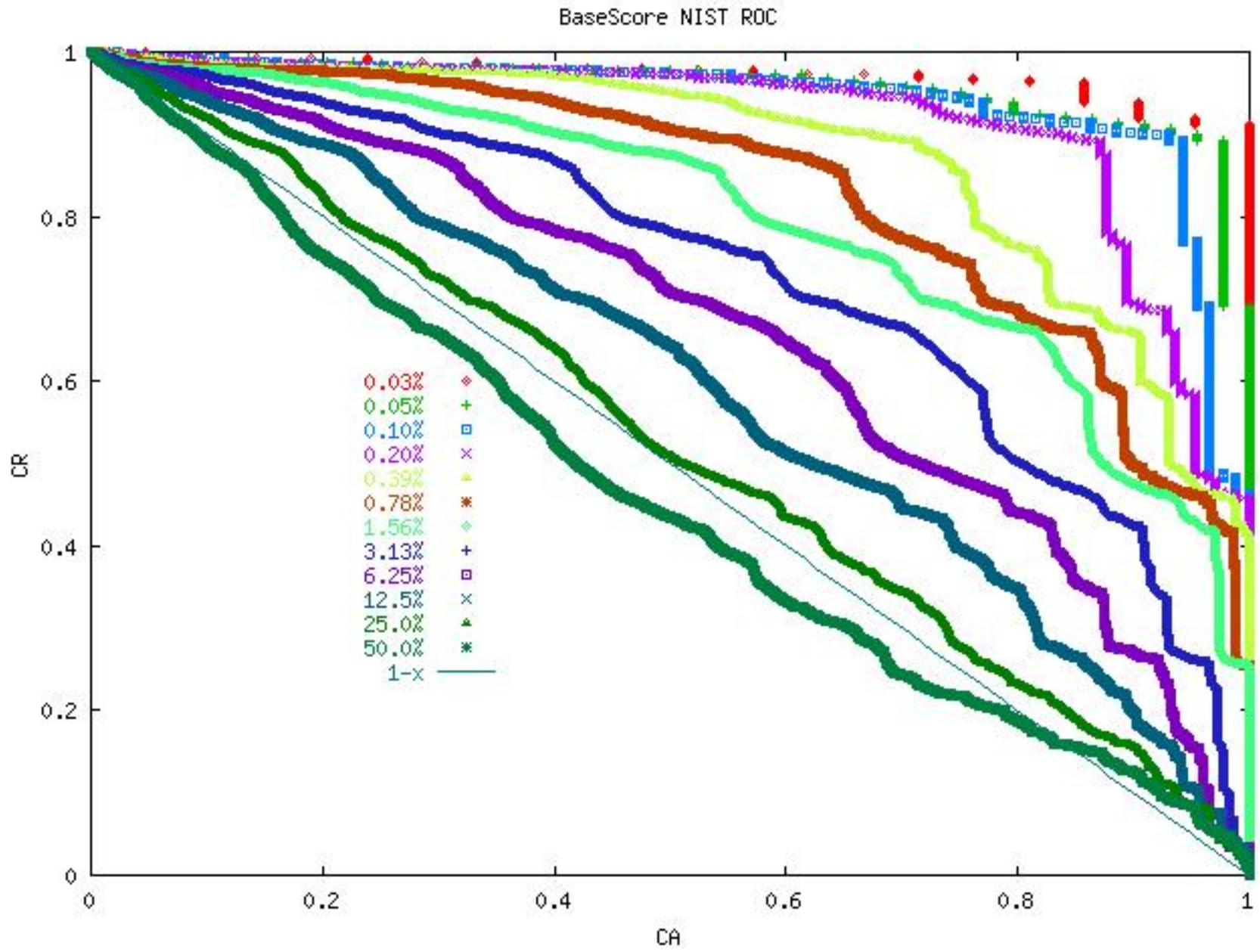
Confidence Features

- base model features: sentence score, feature functions
- LM-based features for source and target
- length-based features
- semantic features based on WordNet (*)
- alignment-based features (*)
- search-based features
- nbest-based features (*)

Ongoing Work

- sub-sentence CE features
- sentence CE features
- ML experiments
- applications



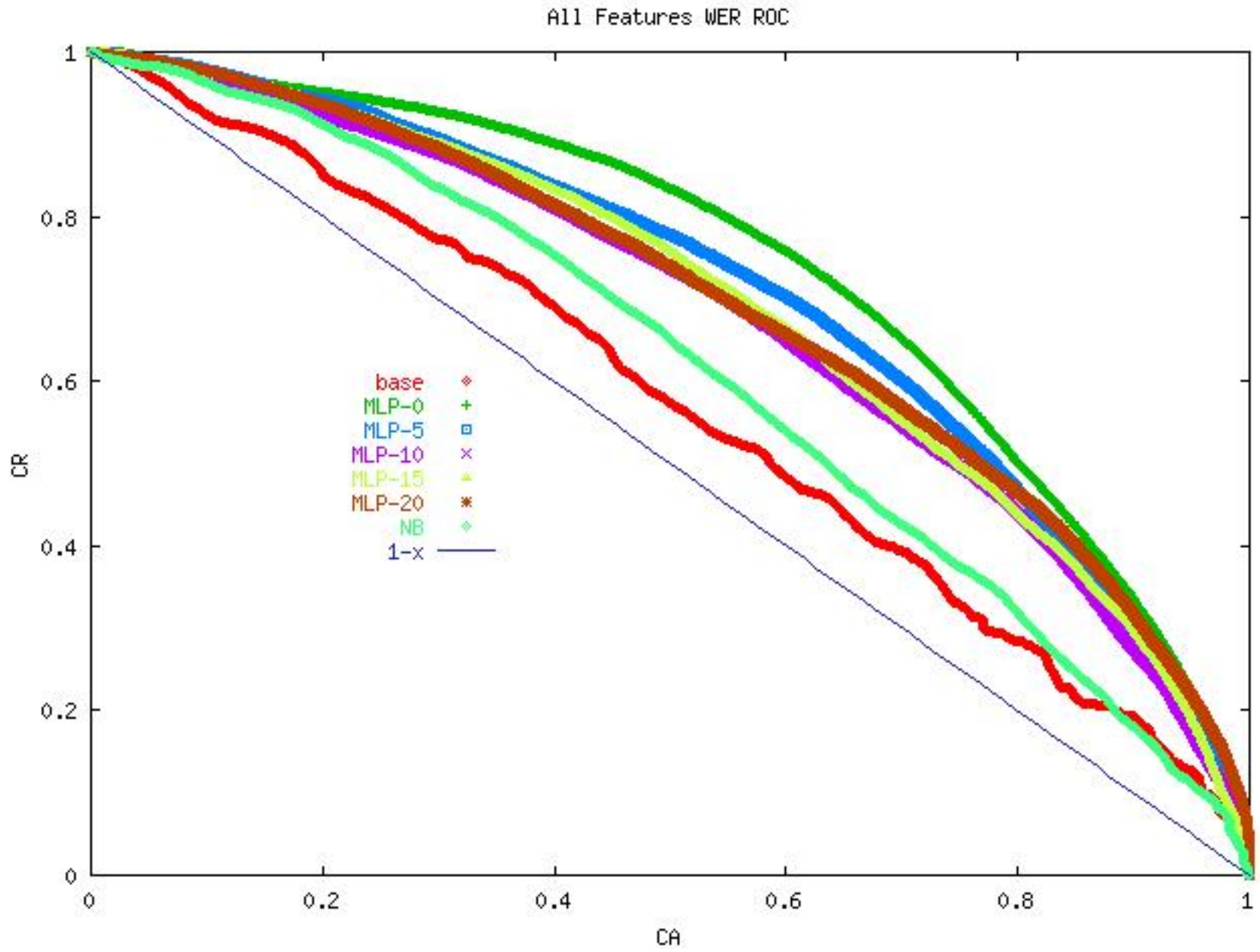


Performance Versus Threshold

split	WER		NIST	
	ACC	IROC	ACC	IROC
50	60.30	0.644354	51.86	0.467489
75	75.61	0.693197	75.01	0.529515
87.5	87.23	0.757764	87.50	0.599596
93.75	93.77	0.799305	93.76	0.663607
96.88	96.93	0.869965	96.88	0.731391
98.44	98.43	0.922409	98.44	0.797121
99.22	99.14	0.945544	99.22	0.834450
99.61	99.60	0.970117	99.61	0.877164
99.80	99.79	0.971641	99.81	0.918668
99.90	99.90	0.975645	99.90	0.941727

Feature Comparison

Feature	ACC	Feature	ACC
BaseFeature-11	74.23	searchfeat-2	73.93
To1k-Align-Comp-3	74.20	length-average	73.85
top1K-quartile-8	74.19	to1K-align-comp-1	73.85
top1K-quartile-5	74.16	top1K-quartile-ok-12	73.85
BaseScore	74.11	BaseFeature-1	73.83
top1K-quartile-7	74.10	prob-3gram	73.80
searchfeat-4	74.00	BaseFeature-8	73.79
searchfeat-5	74.00	BaseFeature-5	73.79
searchfeat-3	73.98	BaseFeature-6	73.76
top1K-lmscore-3	73.97	2gram-perplex	73.76
BaseFeature-2	73.97	to1K-align-comp-ok-2	73.75



Model Comparison

model	ACC	IROC	CE (bits)
BaseScore	76.37	0.56910	—
Naive Bayes	76.58	0.61043	—
MLP 00	78.03	0.73816	0.863575
MLP 05	77.52	0.70336	0.904420
MLP 10	77.26	0.67681	0.967838
MLP 15	77.17	0.68898	0.980132
MLP 20	78.21	0.69211	0.989122

Baseline is 76.36%

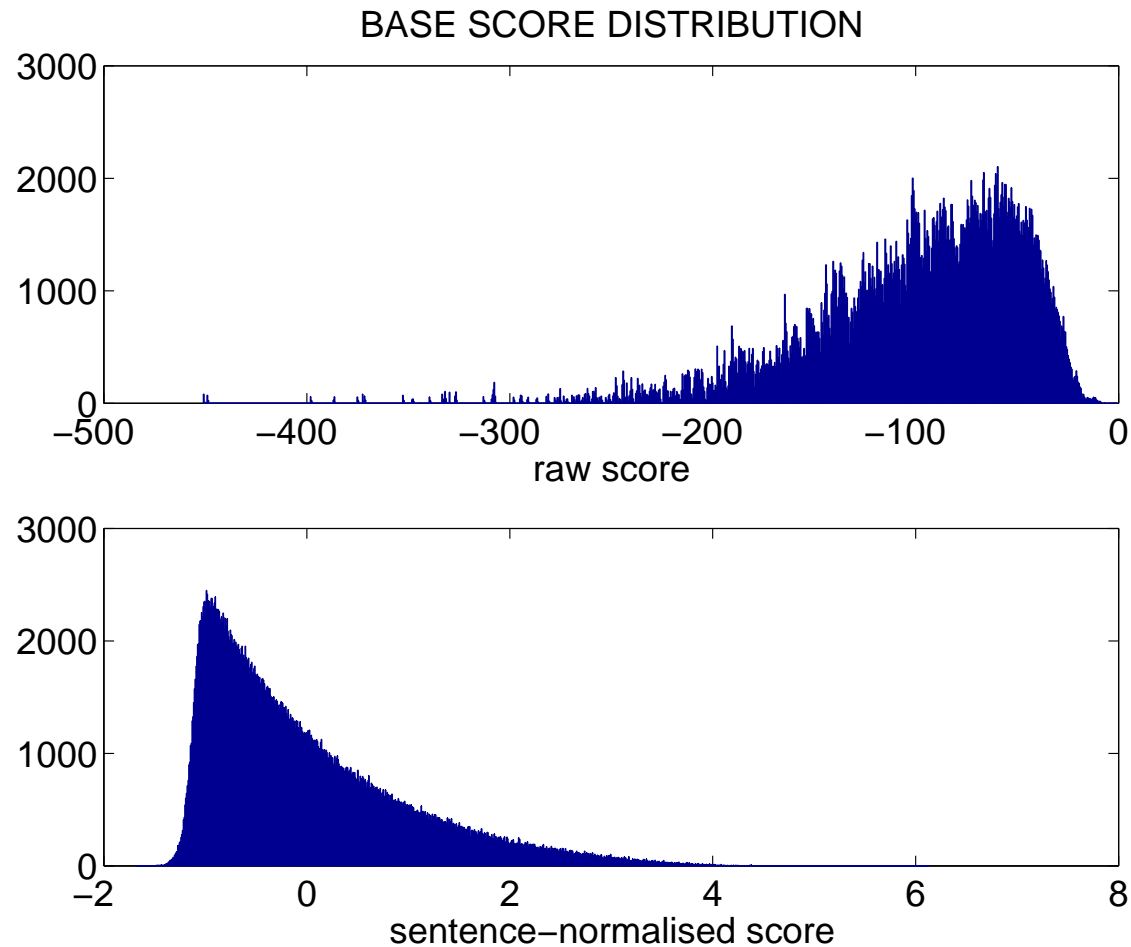
Further Experiments

Parameters to vary:

- ML tuning
- features
- error metric and threshold
- nbest length
- classification versus regression

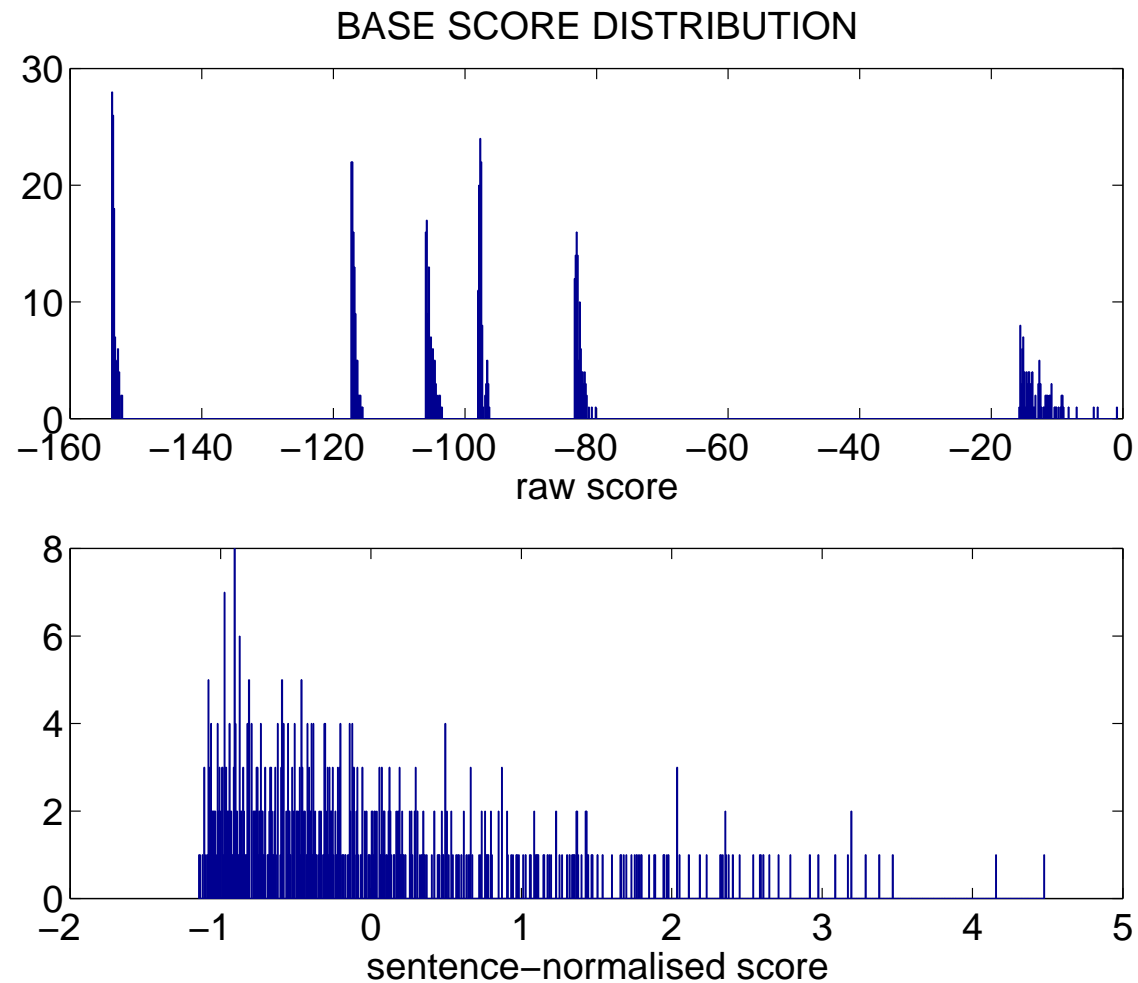
Sentence-based normalisation

Base score = \log posterior probability (used for ranking)



“base score” feature contd.

Standardise feature ($\mu = 0, \sigma = 1$) within each sentence



Evaluation of hypothesis translations

Evaluation = the SMT equivalent to religious wars

Why go through (yet another) evaluation?

- ✓ use as human reference for translation correctness
- ✓ check that WER/NIST emulate translation correctness reasonably well

We do *not* claim this evaluation is uniformly better or more useful than any other MT evaluation scheme ever proposed.

Example

Hypothesis:

→ last year , 583 people surrendered to the korean record .

Reference:

→ last year , a record number of 583 persons fled to rok .

1: Useless ? 2: Poor ? 3: Mediocre ? 4: Acceptable ? 5: Good ?

→ what is *your* standard for MT quality?

Preliminary results

Less than 1000 votes so far. Top voters:

User	Nb. votes
ueffing	152
kulesza	134
jblatz	100
simona	99
erin	67

User	Nb. votes
och	60
fraser	47
cyril	46
foster	39
libin	36

Remember: voting is a civic duty!

Distributions of vote:

1: 102

2: 214

3: 352

4: 210

5: 92

With large deviations depending on users.

Vote consensus

Among sentences that have been judged by two different evaluators:

[1]	31				
[2]	19	41			
[3]	15	90	76		
[4]	5	19	81	40	
[5]	1	3	14	24	25
	[1]	[2]	[3]	[4]	[5]

Good consensus overall, some variability within 1 point.

(More results to be presented when more data have been collected and analysed)