
Findings of the 2009 Workshop on Statistical Machine Translation

Chris Callison-Burch, Philipp Koehn, Christof Monz and Josh Schroeder

30 March 2009



Schedule

- 9:15–9:45 Shared Task Overview
- 9:45–10:15 “Boaster Session” for Evaluation and System Combination posters
- 10:15–11:30 Poster Session for Evaluation and System Combination shared tasks
- 11:30–12:30 Invited Talk by Martin Kay
- 12:30–14:30 Lunch
- 14:30–15:30 Boaster Session for Translation posters
- 15:30–17:30 Poster Session for Translation task

Tomorrow: Full paper presentations (9:00–18:00)

3 Shared Task

- Translation Task
 - English ↔ Czech, French, German, Hungarian, and Spanish
 - 87 machine translation systems
- System Combination Task
 - Create better quality translation by combining output of multiple systems
 - 22 system combination entries
- Evaluation Task
 - Automatically score the system outputs and try to predict the human judgments of translation quality
 - 20+ metrics and variants

Goals of WMT Shared Tasks

- Forum for European language translation
 - provide incentive to work on European language translation
 - low barrier of entry → attract one-grad-student teams
 - showcase state of the art
- Openness
 - open to non-statistical systems
 - full release of submissions, results, judgments
- Large scale manual evaluation
 - judge quality of individual systems and system combinations
 - circumvent bias in automatic metrics
 - use human judgments to validate automatic metrics

New This Year

- Larger Training Sets
 - French-English parallel corpus with 800,000,000 words
 - Monolingual news data for better language modeling
- Reduced Number of Conditions
 - Eliminated in-domain Europarl test set
 - Defined one type of human evaluation as primary
- New Evaluation
 - Edited system outputs without seeing reference
 - Judged whether edited output was correct

Test Set

- Traditionally
 - held-out portion of Europarl training data used as test set
 - does not accurately reflect translation quality for other domains
- Better
 - news stories from European media
 - but no existing translations available
(even multilingual BBC and EuroNews sites do not translate)
- EUROMATRIX funding allowed creation of news test set
 - 136 stories from seven different languages
 - each translated into the other six languages (42 language pairs)
 - co-ordinated by CELCT, with contributions from all EUROMATRIX partners
 - total cost: 31,700 euros

Sources for Test Set

Hungarian: hvg.hu, Napi, MNO, Népszabadság

Czech: iHNed.cz, iDNES.cz, Lidovky.cz, aktuálně.cz, Novinky

French: dernieresnouvelles, Le Figaro, Les Echos, Liberation, Le Devoir

Spanish: ABC.es, El Mundo

English: BBC, New York Times, Times of London

German: Süddeutsche Zeitung, Frankfurter Allgemeine Zeitung, Spiegel, Welt

Italian: ADN Kronos, Affari Italiani, ASCA, Corriere della Sera, Il Sole 24 ORE, Il Quotidiano, La Repubblica

Italian translation was not one of this year's official tasks, but the test set is available for anyone who is interested.

Translation Task Participants

Carnegie Mellon University
Columbia University
Charles University (2 groups)
Dublin City University
University of Geneva
Google (online system)
Johns Hopkins University (2 groups)
LIMSI
Linköping University
University of Le Mans / Systran
Morphologic
NICT, Japan
National University of Singapore
RWTH Aachen

RBMT: Eurotranxp, L&H, Lingenio,
Lucy, pctrans, PROMT, SDL
University of Stuttgart
Systran
Universitat Politecnica de Catalunya
University of Edinburgh
University of Karlsruhe
University of Maryland
University of Saarland

BBN
Carnegie Mellon University (2)
Dublin City University
RWTH Aachen
University of Saarland

Manual Evaluation

- 87 system submissions were manually evaluated
- Recruited 150+ judges, who contributed 480 hours for 70,000+ judgments
- Scoring
 - judges were asked to rank the same sentence output from up to 5 systems
 - convert to pairwise judgment : how many times was A ranked \geq B?
 - overall score: how often was A ranked \geq any other system?

Translation Task Results

- Ranking of each system according to how often it was judged \geq any other system
- “Best systems” are those for which no other system was statistically significantly better
- Best overall systems are marked with ●
- Best constrained system are marked with ★
- Constrained systems are highlighted in yellow

French Systems

French-English		English-French	
System	\geq others	System	\geq others
GOOGLE ●	.76	LIUM-SYSTRAN ●	.73
DCU ★	.66	GOOGLE ●	.68
LIMSI ●	.65	UKA ●★	.66
JHU ★	.62	SYSTRAN ●	.65
UEDIN ★	.61	RBMT3 ●	.65
UKA	.61	DCU ●★	.65
LIUM-SYSTRAN	.60	LIMSI ●	.64
RBMT5	.59	UEDIN ★	.60
CMU-STATXFER ★	.58	RBMT4	.59
RBMT1	.56	RWTH	.58
USAAR	.55	RBMT5	.57
RBMT3	.54	RBMT1	.54
RWTH ★	.52	USAAR	.48
COLUMBIA	.50	GENEVA	.38
RBMT4	.47		
GENEVA	.34		

German Systems

German-English		English-German	
System	\geq others	System	\geq others
RBMT5	.66	RBMT2 ●	.66
USAAR ●	.65	RBMT3 ●	.64
GOOGLE ●	.65	RBMT5 ●	.64
RBMT2 ●	.64	USAAR	.58
RBMT3	.64	RBMT4	.58
RBMT4	.62	RBMT1	.57
STUTTGART ●★	.61	GOOGLE	.54
SYSTRAN ●	.60	UKA ★	.54
UEDIN ★	.59	UEDIN ★	.51
UKA ★	.58	LIU ★	.49
UMD ★	.56	RWTH ★	.48
RBMT1	.54	STUTTGART	.43
LIU ★	.50		
RWTH	.50		
GENEVA	.33		
JHU-TROMBLE	.13		

Spanish Systems

Spanish–English		English–Spanish	
System	\geq others	System	\geq others
GOOGLE ●	.70	RBMT3 ●	.66
TALP-UPC ●★	.59	UEDIN ●★	.66
UEDIN ★	.56	GOOGLE ●	.65
RBMT1 ●	.55	RBMT5 ●	.64
RBMT3 ●	.55	RBMT4	.61
RBMT5 ●	.55	NUS ★	.59
RBMT4 ●	.53	TALP-UPC	.58
RWTH ★	.51	RWTH	.51
USAAR	.51	RBMT1	.25
NICT	.37	USAAR	.48

Czech & Hungarian Systems

Czech–English		English–Czech	
System	\geq others	System	\geq others
GOOGLE ●	.75	RBMT6 ●	.67
UEDIN ★	.57	RBMT7 ●	.67
CU-BOJAR ★	.51	GOOGLE	.66
		CU-BOJAR ★	.61
		UEDIN	.53
		CU-TECTOMT	.48

Hungarian–English	
System	\geq others
MORPHO ●	.75
UMD ★	.66
UEDIN	.45



System Combination Shared Task

- An official task this year (last year was by invitation only)
- Combine the output of all individual systems to get better translation quality:
- Provided materials:
 - 87 primary system submissions, 42 secondary submissions
 - Created system combination dev set by translating 25 articles / 500 sentences
 - Solicited 100-best lists from participants, received 30
- Evaluated system combinations alongside individual systems
- Question: statistically significantly better than best individual systems?

System Combination Task Results

French–English

System	\geq
GOOGLE ●	.76
BBN-COMBO	.73
CMU-HYPSEL ●	.71
DCU-COMBO ●	.67
CMU-COMBO ●	.66
LIMSI	.65
USAAR-COMBO	.57

English–French

System	\geq
USAAR-COMBO ●	.77
DCU-COMBO ●	.74
LIUM-SYSTRAN ●	.73
GOOGLE ●	.68
UKA	.66
SYSTRAN	.65
RBMT3 ●	.65
DCU ●	.65
LIMSI	.64

German–English

System	\geq
RWTH-COMBO ●	.70
BBN-COMBO ●	.68
USAAR ●	.65
GOOGLE	.65
RBMT2 ●	.64
CMU-COMBO ●	.63
USAAR-COMBO	.62
CMU-HYPSEL	.62
STUTT GART ●	.61
SYSTRAN	.60

English–German

System	\geq
RBMT2 ●	.66
RBMT3 ●	.64
RBMT5 ●	.64
USAAR-COMBO	.52

Spanish–English

System	\geq
GOOGLE ●	.70
CMU-COMBO ●	.70
USSAR-COMBO ●	.69
BBN-COMBO ●	.64
TALP-UPC ●	.59
RBMT1	.55
RBMT3	.55
RBMT5	.55
RBMT4	.53

English–Spanish

System	\geq
RBMT3 ●	.66
UEDIN ●	.66
GOOGLE ●	.65
RBMT5 ●	.64
USSAR-COMBO ●	.61

Czech–English

System	\geq
GOOGLE ●	.75
CMU-COMBO ●	.73
BBN-COMBO	.65

Hungarian–English

System	\geq
MORPHO ●	.75
CMU-HYPSEL	.68
CMU-COMBO	.62
BBN-COMBO	.54

Multisource–English

System	\geq
RWTH-COMBO ●	.67
BBN-COMBO	.62
CMU-COMBO	.58



Evaluation Task

- How well do the automatic metrics correspond to human judgments?
- Compare the metric scores and human rankings (at the system level)
- Measure Spearman's rank correlation



Automatic Metrics

- Over 20 automatic metrics and their variants were submitted
- These metrics employed different tactics

N-gram matching:

Bleu, Meteor, M-TER, M-Bleu

Non-exact matching:

Similarity scores, Paraphrases

Linguistic info:

Dependency overlap, Semantic role overlap, Bleu over POS tags

Aggregate measures:

Combination of many of the above

Evaluation Task Results

into English			out of English		
Metric	ρ	Highest for	Metric	ρ	Highest for
ULC	.83	de-en	TERP	.46	en-cz
MAXSIM	.8	es-en	TER	.35	
RTE	.79	hu-en	BLEUSP	.30	
METEOR	.75	fr-en	BLEU	.27	
TERP	.72		BLEU-TER/2	.26	
NIST	.56		WCD6P4ER	.22	
WPF	.56		NIST	.20	
TER	.54		out of English, no Czech		
BLEU	.51		Metric	ρ	Highest for
BLUESP	.48		WPBLEU	.54	en-de, en-fr, en-es
WPBLEU	.46		TERP	.48	
WCD6P4ER	.45		WPF	.47	
			BLEU	.33	



Evaluation Task Results (sentence-level)

- Only 2/3 of the automatic metrics performed better than chance on predicting human judgments at the sentence level
- None of the metrics on translation out-of-English performed better than chance overall
- When above chance not much more so
- See paper for details

However: when metrics' system-level scores are treated as their sentence level scores, they perform significantly better



Refinements to Manual Evaluation

- One goal of this workshop is to improve the manual evaluation
- We gather annotator agreement and timing information to verify that people can do the task reliably
- We defined sentence ranking to be our official metric
- We also introduced an experimental type of manual evaluation based on editing

Annotator Agreement

- We measured agreement among annotators using the kappa coefficient:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where

- $P(A)$ is the proportion of times that the annotators agree
 - $P(E)$ is the proportion of time that they would agree by chance.
- Interpretation of K scores varies, but:
 - .6 – .8 is good agreement
 - .4 – .6 is moderate agreement
 - .2 – .4 is fair agreement
 - $< .2$ is slight agreement

Inter-Annotator Agreement

Evaluation type	$P(A)$	$P(E)$	K	agreement
Fluency*	.40	.2	.25	fair
Adequacy*	.38	.2	.23	fair
Sentence ranking	.55	.333	.32	fair
Constituent ranking [†]	.68	.333	.52	moderate
Constituent yes/no [†]	.83	.5	.65	good

* From WMT07

† From WMT08

Intra-Annotator Agreement

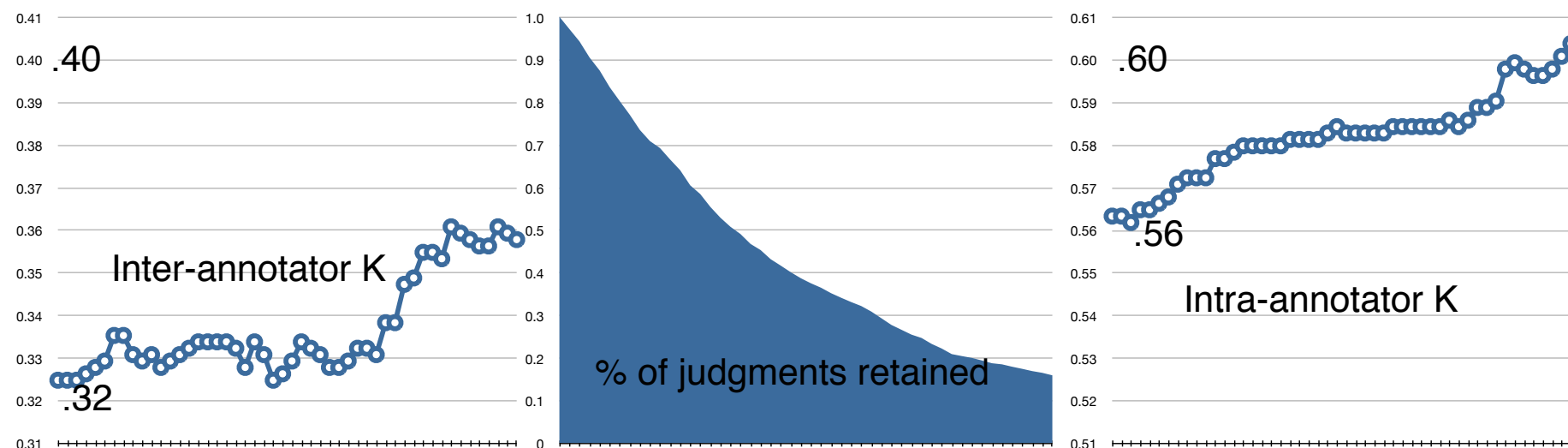
Evaluation type	$P(A)$	$P(E)$	K	agreement
Fluency*	.63	.2	.54	moderate
Adequacy*	.57	.2	.47	moderate
Sentence ranking	.71	.333	.56	moderate
Constituent ranking [†]	.83	.333	.75	good
Constituent yes/no [†]	.93	.5	.86	very good

* From WMT07

† From WMT08

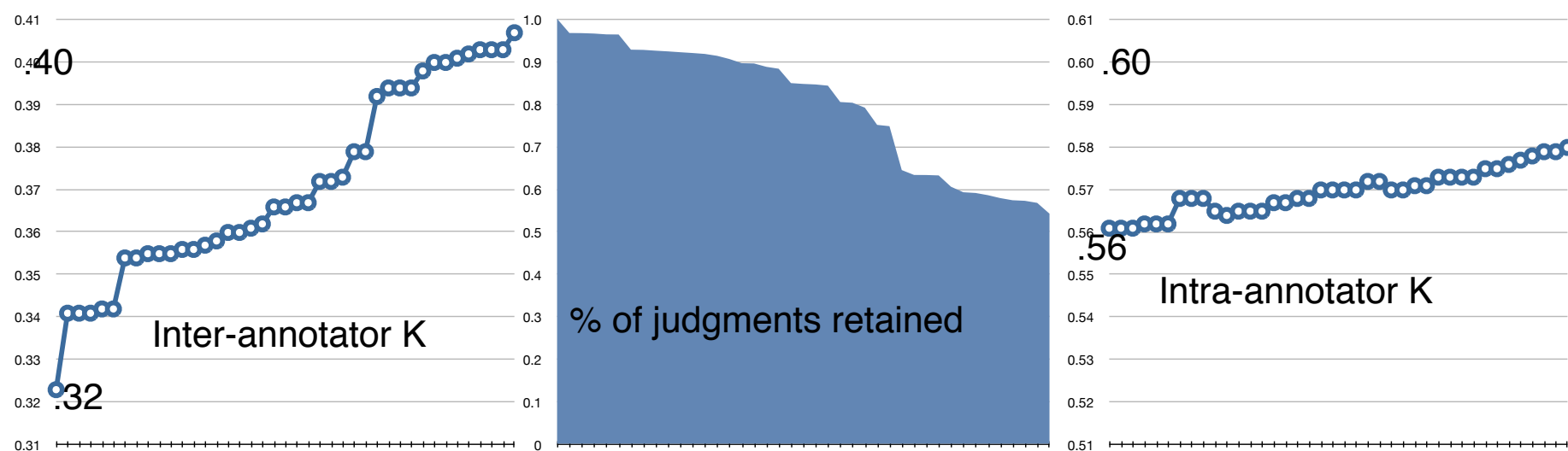
Improving Agreement on Sentence Ranking

- Tried removing up to 50 initial judgments to allow for a “learning period”
- Intra-annotator agreement moved from moderate to good
- However, 80% of the judgments were discarded



Improving Agreement on Sentence Ranking

- Tried removing judges with the lowest agreement with others
- Inter-annotator agreement moved from fair to moderate while retaining 60% of judgments





Evaluation through Editing

- New type of evaluation where people edit the output of machine translation systems (without seeing the source or reference)
- The edited output is later displayed alongside the source and reference and it is judged to be correct or not
- The hope is that this would somehow reflect the understandability of the translations

Editing MT output

Original: They are often linked to other alterations sleep as nightmares, night terrors, the nocturnal enuresis (pee in bed) or the sleepwalking, but it is not always the case.

Edit:

They are often linked to other sleep disorders, such as nightmares, night terrors, the nocturnal enuresis (bedwetting) or sleepwalking, but this is not always the case.

[Reset Edit](#)

- Edited.
- No corrections needed.
- Unable to correct.

- Shown without source or reference, also without context
- Careful not to have an annotator correct the output of multiple systems for a single sentence, since understanding could be influenced
- Ensured sentence-ranking items did not overlap with editing for same reason

Judging Correctness

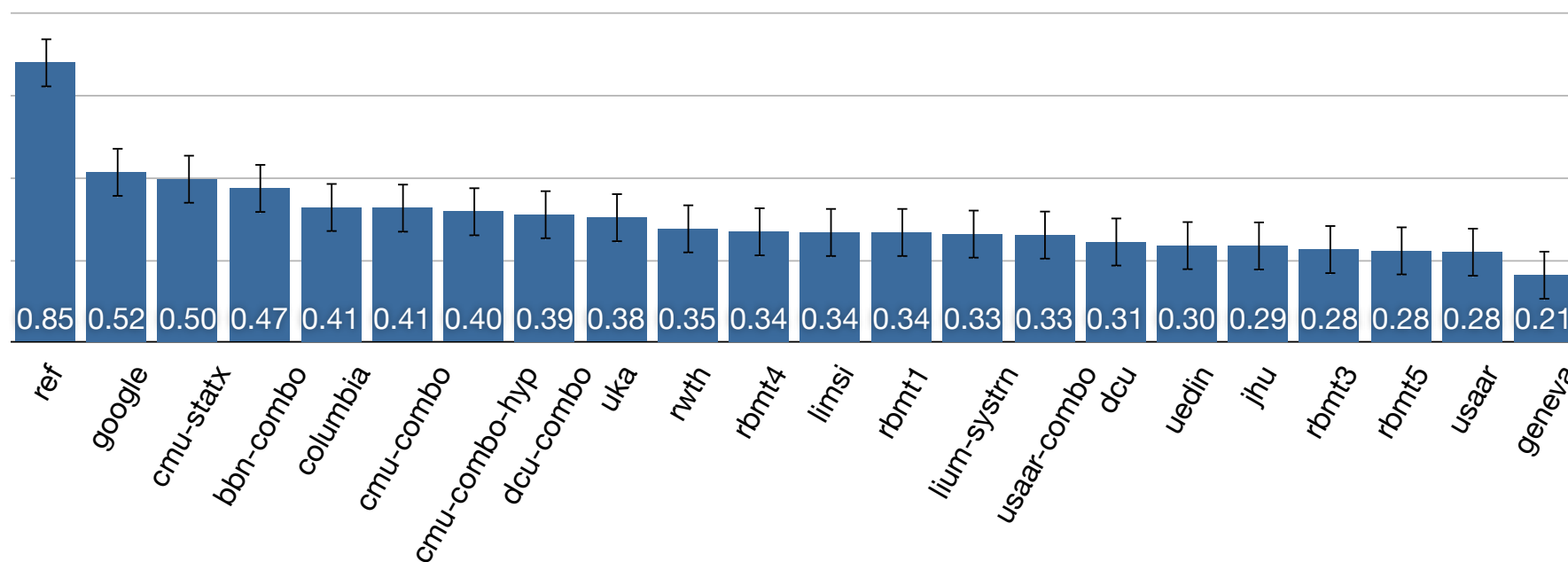
Reference: Meanwhile, the Belgian, Dutch and Luxembourg governments partially nationalized the European financial conglomerate Fort Analysts at Barclays Capital said the frantic weekend negotiations that led to the bailout agreement "appear to have failed to revive market **As the economic situation deteriorates, the demand for commodities, including oil, is expected to slow down.**

"The outlook for global equity, interest rate and exchange rate markets has become increasingly uncertain," analysts at Deutsche Bank wrote investors.

"We believe commodities will be unable to escape the contagion.

Translation	Verdict
While the economic situation is deteriorating, demand for commodities, including oil, should decrease.	<input checked="" type="radio"/> Yes <input type="radio"/> No
While the economic situation is deteriorating, the demand for raw materials, including oil, should slow down.	<input checked="" type="radio"/> Yes <input type="radio"/> No
Alors que the economic situation deteriorated, the request in rawmaterial enclosed, oil, would have to slow down.	<input type="radio"/> Yes <input checked="" type="radio"/> No
While the financial situation damaged itself, the first matters affected, oil included, should slow down themselves.	<input type="radio"/> Yes <input checked="" type="radio"/> No
While the economic situation is depressed, demand for raw materials, including oil, will be slow.	<input type="radio"/> Yes <input checked="" type="radio"/> No

Acceptability of French-English Translations



- Highest acceptability is around 50%, not strongly correlated with sentence-level ranking for all languages



Summary

- Large scale evaluation of machine translation for European languages
- Some research systems doing as well as Google for many languages
- Jury still out on whether system combination is significantly better than best individual systems
- Bleu not the best automatic metric
- Judgments will be downloadable <http://www.statmt.org/wmt09/results.html>
 - Further analysis
 - Creating new metrics
 - Tuning systems with human judgments



Plans for WMT10

- The European Union has funded EuroMatrixPlus for 3 years
- We will host workshops in 2010–2012
- Testing period for WMT10 around January 2010
- Workshop at ACL Uppsala, Sweden (July 2010)
- Also maintaining ongoing online evaluation
<http://matrix.statmt.org/>



Discussion

How should we manage the system combination task?

How should we refine the process for manual evaluation through editing?

How can we better deal with sentence-level metrics?

Feedback on the evaluation campaign?

Interest in other languages?