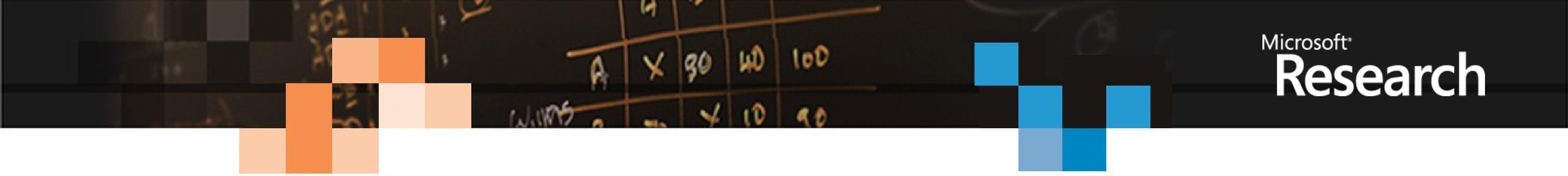


The MSRA MT System for IWSLT'10

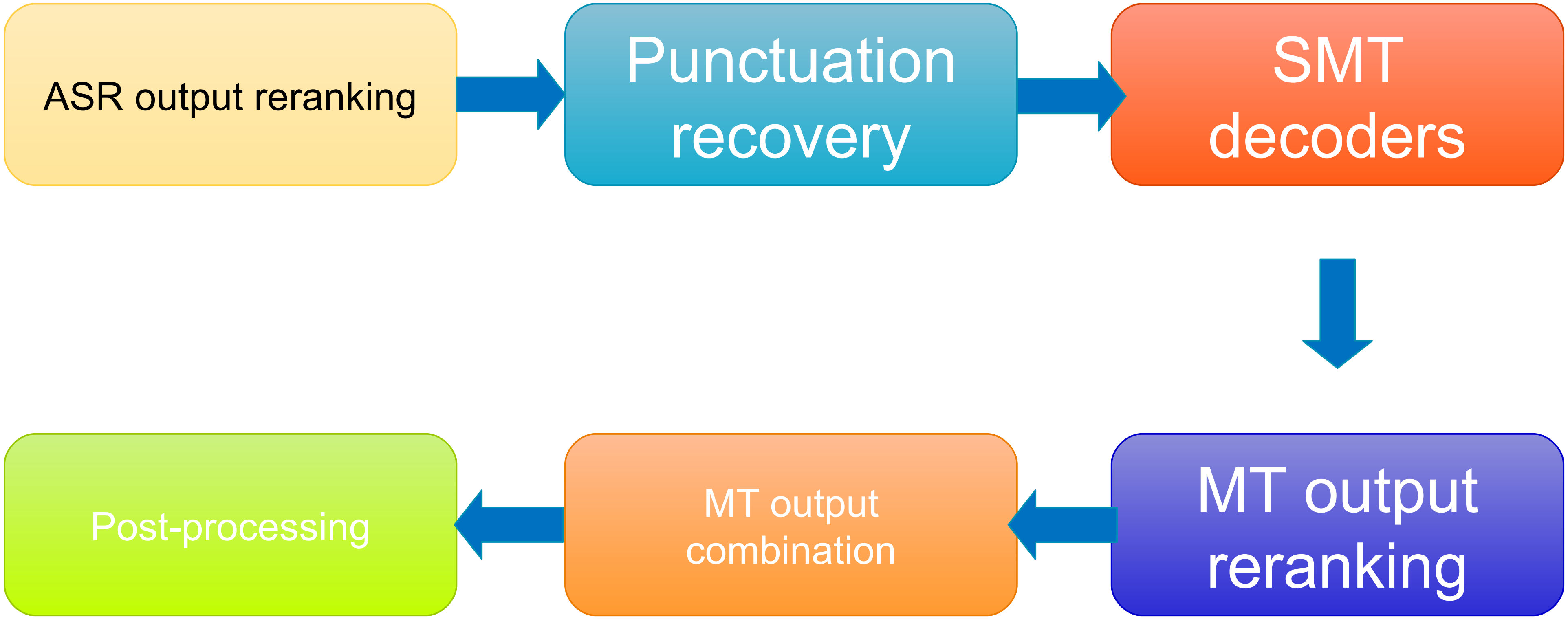
Basic MT Architecture

- Translate by many decoders, then combine MT output
- Phrase-based decoders
    - Hiero, BTG, Moses
- Syntax-augmented phrase-based decoders
    - Hiero & BTG with dependency syntax-based LM
    - Treelet
- Syntax-based decoder
- MT output combination
    - Incremental HMM alignment of translation candidates (ACL'09)

ASR-MT Interface

- How to handle N-best ASR output?
  - Rerank, then select the top one
- How to recover punctuation marks?
  - Where to add a punctuation mark?
  - Which punctuation mark to be added at a given position?
  - Both by CRF classifier

Complete Architecture



ASR Output

- Seems N-best ASR output is better than 1-best
- Indeed even worse if simply feed all N-best
    - N-best contain more errors than error fixes
    - Drop > 1.0 Bleu point
- Better if use top one after reranking

ASR Output Reranking

- Data: some devset where the Oracle (CRR) of ASR is known
- Objective function: BLEU (CRR as reference)
- Training method: Max-Bleu-training
- Model: log-linear
- Features:
  - ASR output scores
  - LM probabilities
  - Count of characters/words

ASR Output Reranking

	BTG	Hiero	Hiero+DepLM	BTG+DepLM	Syntax
Original 1-best	41.05	40.43	42.80	39.16	39.28
Reranked 1-best	41.98	41.24	43.57	39.93	39.99

- Task = CE/ASR; tuning set = devset8+c2e.DIALOG
- Test set = devset9; training set = everything else

Punctuation Recovery

- We tried three approaches
  - Implicit recovery thru translation model.
    - strip all punctuation marks on the source side of bilingual data
  - 1-stage classification
    - after each word, ask if it’s followed by ‘,’, ‘.’, ‘!’, or nothing
  - 2-stage classification
    - after each word, ask if it’s followed by a punctuation mark or not
    - at each position found in stage 1, ask if it’s ‘,’, ‘.’ or ‘!’

Punctuation Recovery

	BTG	Hiero
Implicit recovery thru translation	47.87	40.98
1-stage classification	48.63	41.51
2-stage classification	48.96	41.78

- Task = EC/ASR; tuning set = devset10+e2c.DIALOG
- Test set = devset11; training set = everything else

Useful Tricks

- Manual rules for number/date/time translation
- Combination of word alignment
  - Concatenate alignment matrices of the same data by different word aligners
- Reranking of N-best translation output

MT Output Reranking

- Data: some devset with reference translations
- Objective function: BLEU
- Training method: Max-Bleu-training
- Model: log-linear
- Features:
  - N-gram posterior probabilities
  - Sentence length posterior probabilities
  - LM probabilities
  - Length ratio between source input and translation

MT Output Reranking

	BTG	Hiero	Hiero+DepLM	BTG+DepLM	Syntax
Original 1-best	45.69	45.45	48.47	44.15	44.45
Reranked 1-best	46.07	48.42	47.80	45.74	45.65

- Task = CE/CRR; tuning set = devset8+c2e.DIALOG
- Test set = devset9; training set = everything else

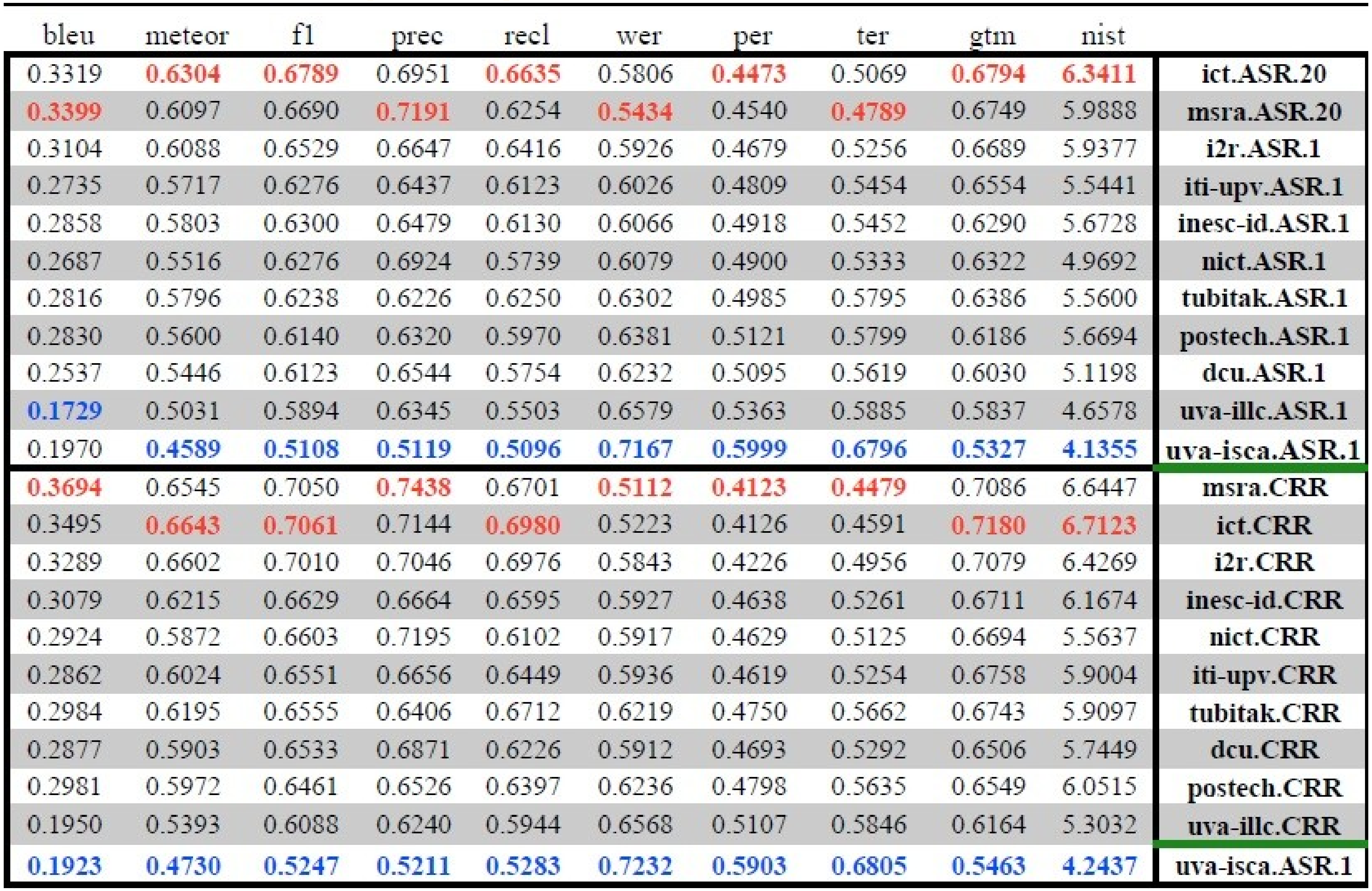
MT Output Reranking

some observations

- Raw output of different decoders are very different from each other
- Reranked output of diffferent decoders are quite similar to each other
- Bleu(combo(raw_output)) >> Bleu(raw_output)
- Bleu(combo(reranked_output)) > Bleu(reranked_output)
- Bleu(combo(raw_output)) ~= Bleu(combo(reranked_output))

IWSLT'10 Evaluation

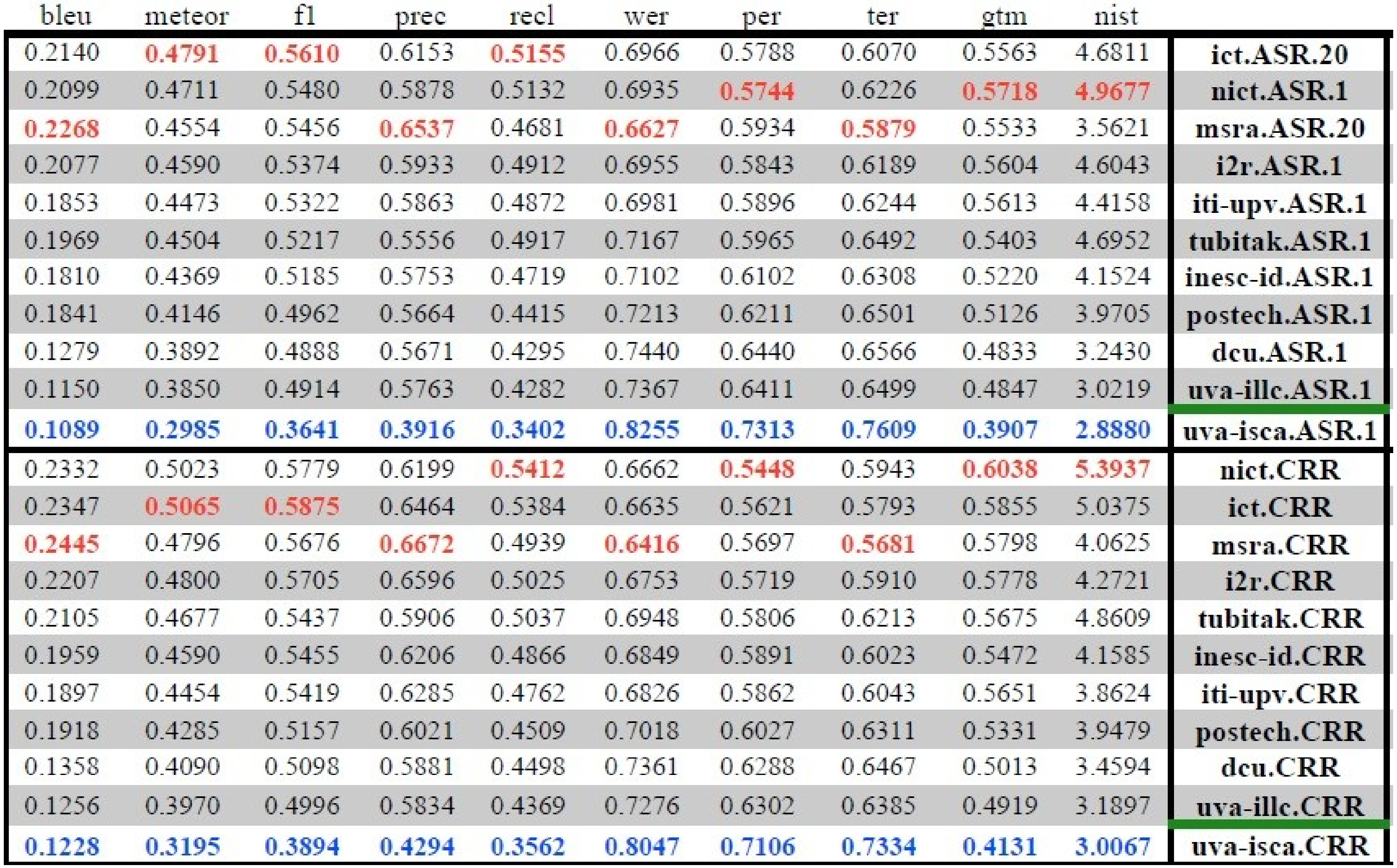
2009 testset (CE)



bleu	meteor	f1	prec	recl	wer	per	ter	gtm	nist	
0.3319	0.6304	0.6789	0.6951	0.6635	0.5806	0.4473	0.5069	0.6794	6.3411	ict.ASR.20
0.3399	0.6097	0.6690	0.7191	0.6254	0.5434	0.4540	0.4789	0.6749	5.9888	msra.ASR.20
0.3104	0.6088	0.6529	0.6647	0.6416	0.5926	0.4679	0.5256	0.6689	5.9377	i2r.ASR.1
0.2735	0.5717	0.6276	0.6437	0.6123	0.6026	0.4809	0.5454	0.6554	5.5441	iti-upv.ASR.1
0.2858	0.5803	0.6300	0.6479	0.6130	0.6066	0.4918	0.5452	0.6290	5.6728	inesc-id.ASR.1
0.2687	0.5516	0.6276	0.6924	0.5739	0.6079	0.4900	0.5333	0.6322	4.9692	nict.ASR.1
0.2816	0.5796	0.6238	0.6226	0.6250	0.6302	0.4985	0.5795	0.6386	5.5600	tubitak.ASR.1
0.2830	0.5600	0.6140	0.6320	0.5970	0.6381	0.5121	0.5799	0.6186	5.6694	postech.ASR.1
0.2537	0.5446	0.6123	0.6544	0.5754	0.6232	0.5095	0.5619	0.6030	5.1198	dcu.ASR.1
0.1729	0.5031	0.5894	0.6345	0.5503	0.6579	0.5363	0.5885	0.5837	4.6578	uva-illc.ASR.1
0.1970	0.4589	0.5108	0.5119	0.5096	0.7167	0.5999	0.6796	0.5327	4.1355	uva-isca.ASR.1
0.3694	0.6545	0.7050	0.7438	0.6701	0.5112	0.4123	0.4479	0.7086	6.6447	msra.CRR
0.3495	0.6643	0.7061	0.7144	0.6980	0.5223	0.4126	0.4591	0.7180	6.7123	ict.CRR
0.3289	0.6602	0.7010	0.7046	0.6976	0.5843	0.4226	0.4956	0.7079	6.4269	i2r.CRR
0.3079	0.6215	0.6629	0.6664	0.6595	0.5927	0.4638	0.5261	0.6711	6.1674	inesc-id.CRR
0.2924	0.5872	0.6603	0.7195	0.6102	0.5917	0.4629	0.5125	0.6694	5.5637	nict.CRR
0.2862	0.6024	0.6551	0.6656	0.6449	0.5936	0.4619	0.5254	0.6758	5.9004	iti-upv.CRR
0.2984	0.6195	0.6555	0.6406	0.6712	0.6219	0.4750	0.5662	0.6743	5.9097	tubitak.CRR
0.2877	0.5903	0.6533	0.6871	0.6226	0.5912	0.4693	0.5292	0.6506	5.7449	dcu.CRR
0.2981	0.5972	0.6461	0.6526	0.6397	0.6236	0.4798	0.5635	0.6549	6.0515	postech.CRR
0.1950	0.5393	0.6088	0.6240	0.5944	0.6568	0.5107	0.5846	0.6164	5.3032	uva-illc.CRR
0.1923	0.4730	0.5247	0.5211	0.5283	0.7232	0.5903	0.6805	0.5463	4.2437	uva-isca.ASR.1

IWSLT'10 Evaluation

2010 testset (CE)



bleu	meteor	f1	prec	recl	wer	per	ter	gtm	nist	
0.2140	0.4791	0.5610	0.6153	0.5155	0.6966	0.5788	0.6070	0.5563	4.6811	ict.ASR.20
0.2099	0.4711	0.5480	0.5878	0.5132	0.6935	0.5744	0.6226	0.5718	4.9677	nict.ASR.1
0.2268	0.4554	0.5456	0.6537	0.4681	0.6627	0.5934	0.5879	0.5533	3.5621	msra.ASR.20
0.2077	0.4590	0.5374	0.5933	0.4912	0.6955	0.5843	0.6189	0.5604	4.6043	i2r.ASR.1
0.1853	0.4473	0.5322	0.5863	0.4872	0.6981	0.5896	0.6244	0.5613	4.4158	iti-upv.ASR.1
0.1969	0.4504	0.5217	0.5556	0.4917	0.7167	0.5965	0.6492	0.5403	4.6952	tubitak.ASR.1
0.1810	0.4369	0.5185	0.5753	0.4719	0.7102	0.6102	0.6308	0.5220	4.1524	inesc-id.ASR.1
0.1841	0.4146	0.4962	0.5664	0.4415	0.7213	0.6211	0.6501	0.5126	3.9705	postech.ASR.1
0.1279	0.3892	0.4888	0.5671	0.4295	0.7440	0.6440	0.6566	0.4833	3.2430	dcu.ASR.1
0.1150	0.3850	0.4914	0.5763	0.4282	0.7367	0.6411	0.6499	0.4847	3.0219	uva-illc.ASR.1
0.1089	0.2985	0.3641	0.3916	0.3402	0.8255	0.7313	0.7609	0.3907	2.8880	uva-isca.ASR.1
0.2332	0.5023	0.5779	0.6199	0.5412	0.6662	0.5448	0.5943	0.6038	5.3937	nict.CRR
0.2347	0.5065	0.5875	0.6464	0.5384	0.6635	0.5621	0.5793	0.5855	5.0375	ict.CRR
0.2445	0.4796	0.5676	0.6672	0.4939	0.6416	0.5697	0.5681	0.5798	4.0625	msra.CRR
0.2207	0.4800	0.5705	0.6596	0.5025	0.6753	0.5719	0.5910	0.5778	4.2721	i2r.CRR
0.2105	0.4677	0.5437	0.5906	0.5037	0.6948	0.5806	0.6213	0.5675	4.8609	tubitak.CRR
0.1959	0.4590	0.5455	0.6206	0.4866	0.6849	0.5891	0.6023	0.5472	4.1585	inesc-id.CRR
0.1897	0.4454	0.5419	0.6285	0.4762	0.6826	0.5862	0.6043	0.5651	3.8624	iti-upv.CRR
0.1918	0.4285	0.5157	0.6021	0.4509	0.7018	0.6027	0.6311	0.5331	3.9479	postech.CRR
0.1358	0.4090	0.5098	0.5881	0.4498	0.7361	0.6288	0.6467	0.5013	3.4594	dcu.CRR
0.1256	0.3970	0.4996	0.5834	0.4369	0.7276	0.6302	0.6385	0.4919	3.1897	uva-illc.CRR
0.1228	0.3195	0.3894	0.4294	0.3562	0.8047	0.7106	0.7334	0.4131	3.0067	uva-isca.CRR

IWSLT'10 Evaluation

- High Bleu but low recall-oriented metrics
- Due to evil word drop property
- If parameters are tuned by α•BLEU + (1-α) •UNIGRAM_RECALL
- then for 2009 testset (CE) and Hiero

α	BLEU	NIST	METEOR	F1	WER	PER	TER	GTM
1.0	31.29	6.2845	0.6238	0.6735	0.5799	0.4544	0.5134	0.6834
0.9	30.55	6.1862	0.6208	0.6673	0.5956	0.4662	0.5292	0.6790
0.8	31.46	6.2384	0.6274	0.6834	0.5567	0.4398	0.4891	0.6826
0.7	32.24	6.3378	0.6339	0.6771	0.5862	0.4499	0.5231	0.6897
0.6	30.68	6.1366	0.6326	0.6716	0.6088	0.4668	0.5502	0.6858
0.5	30.37	6.0920	0.6319	0.6674	0.6054	0.4713	0.5522	0.6890

THANK YOU!