NRC·CNRC

NRC.CANADA.CA

YiSi – an open-source evaluation metric and its application

羅致翹 Chi-kiu (Jackie) Lo

Research Officer, Multilingual Text Processing team,
Digital Technologies Research Centre

A useful translation is one from which
human readers accurately understand

“**who** *did* **what** to **whom**, **when**, **where**, **why** and **how**”

from the original input sentence.

However…

Most commonly used MT evaluation metrics (aka BLEU) fail to accurately reflect translation utility

MT1	So far , the sale in the mainland of China for nearly two months of SK - II line of products .	Worst	[sentBLEU: 0.124, best]
MT2	So far , nearly two months sk - ii the sale of products in the mainland of China to resume sales .	Better	[sentBLEU: 0.012, worst]
MT3	So far , in the mainland of China to stop selling nearly two months of SK - 2 products sales resumed .	Best	[sentBLEU: 0.013, better]
REF	Until after their sales had ceased in mainland China for almost two months , sales of the complete range of SK - II products have now been resumed .		
SRC	至此， 在 中国 内地 停售 了 近 两 个 月 的 ＳＫ－Ⅱ 全线 产品 恢复 销 。		

How can we solve the problem?

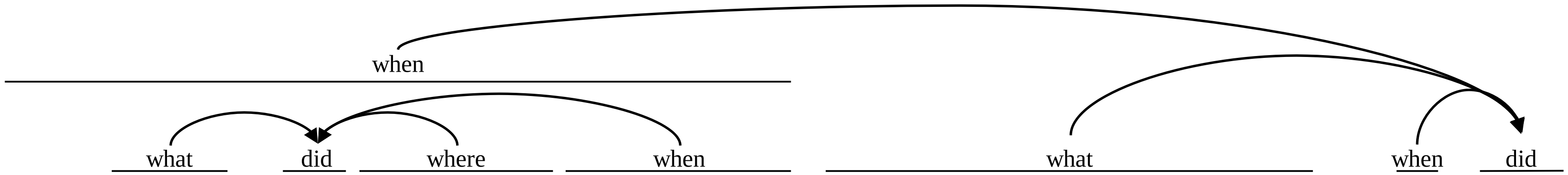
Semantic frames / semantic role labels captures the event structure of a sentence

Represents in a predicate-argument structure

- Predicate verb
- A number of semantic arguments labeled with functional roles

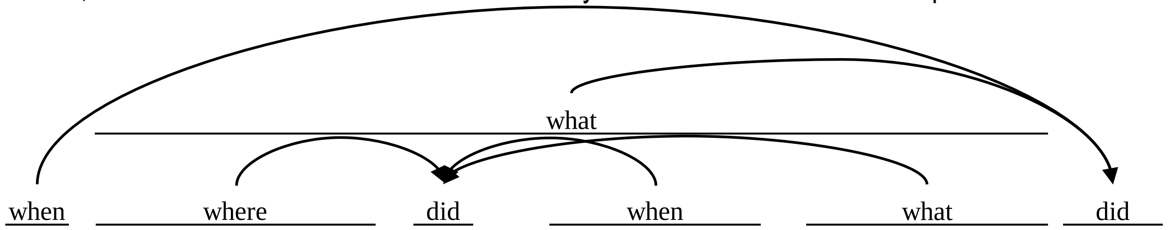
Provides high level understanding of a sentence

Example: Higher-utility (more adequate) translation More SRL matches, but fewer N-gram and syntax-subtree matches!



[REF] Until after their sales had ceased in mainland China for almost two months , sales of the complete range of SK – II products have now be resumed .

[MT1] So far , the sale in the mainland of China for nearly two months of SK - II line of products .



[MT3] So far , in the mainland of China to stop selling nearly two months of SK - 2 products sales resumed .

N-gram	MT1	MT3	Syntax-subtree	MT1	MT3	SRL	MT1	MT3
1-gram matches	15	15	1-level subtree matches	34	**35**	Predicate matches	0	**2**
2-gram matches	4	4	2-level subtree matches	8	**6**	Argument matches	0	**2**
3-gram matches	3	**1**	3-level subtree matches	2	**1**			
4-gram matches	1	**0**	4-level subtree matches	0	0			

Why would it work?

Accuracy of automatic SRL, over 75%

English

ASSERT (Pradhan et al., 2004); SENNA (Collobert et al., 2009);
MATE (Bjorkelund et al., 2009); MATEPLUS (Roth and Woodsend, 2014) and lots more!

Chinese

C-ASSERT (Wu et al., 2006); MATE

Spanish

MATE

German

MATEPLUS

Using SRL significantly improved other NLP tasks

YiSi

romanization of the Cantonese word “意思” (meaning)

Design principles of YiSi

Desirable characteristics of MT quality metrics:

- Accurately reflect translation adequacy
- Inexpensive (minimal data resources)
- Representational transparency for scientific error analysis

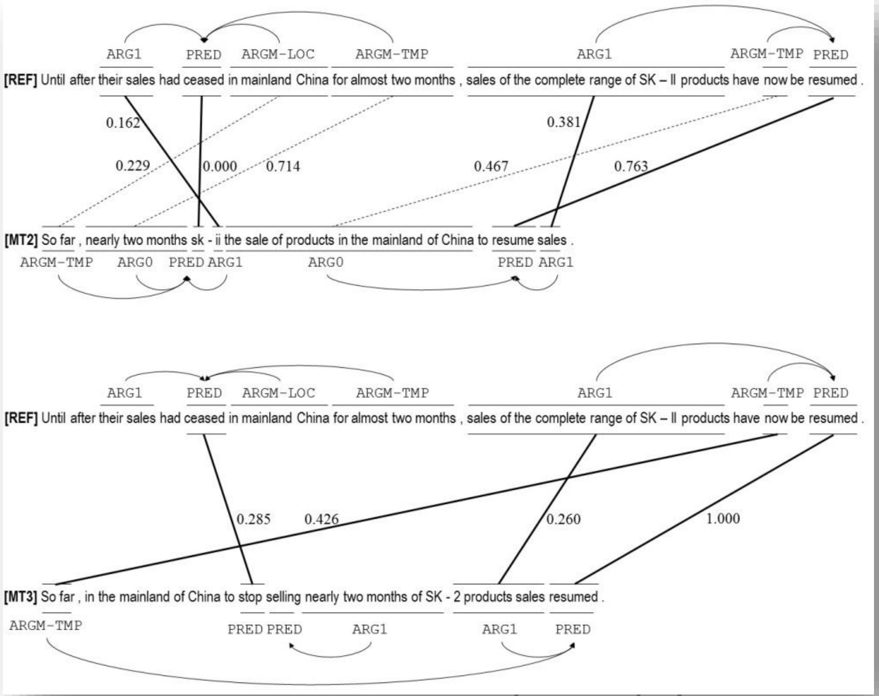
Overview of YiSi algorithm

1. Automatic shallow semantic parser extracts the event structure of both sentences

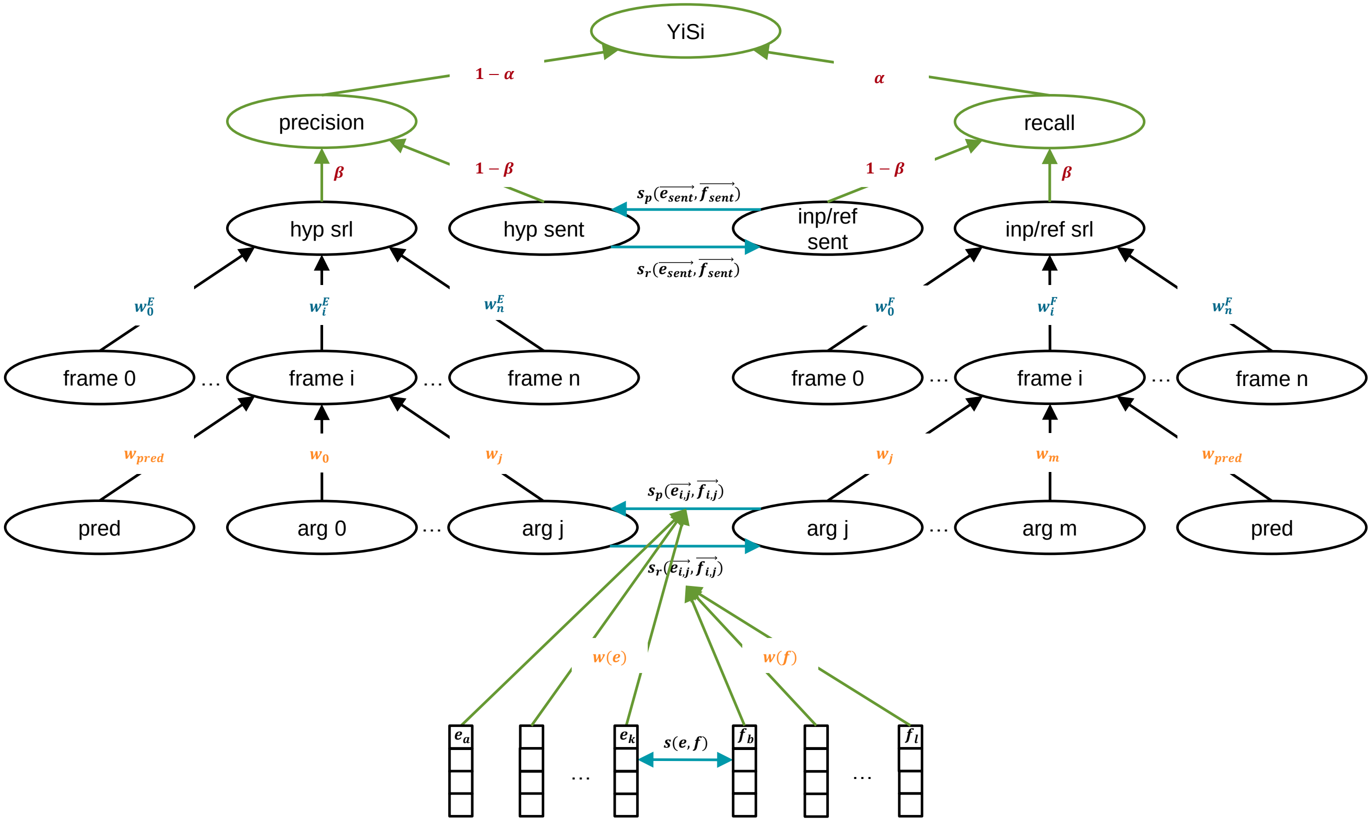
2. Maximum weighted bipartite matching algorithm aligns the semantic frames between the two sentences according to the token similarity of the predicates

3. For each pair of the aligned frames, maximum weighted bipartite matching algorithm aligns the arguments between the two sentences according to the segmental similarity of the role fillers

4. Compute the weighted f-score over the aligned predicates and arguments with matching role labels



YiSi score computation



$w(u)$ = weight of unit u

$s(e,f)$ = similarity of unit e and f

$$s_p\left(\vec{e}, \vec{f}\right) = \frac{\sum_a max_b \sum_{k=0}^{n-1} w(e_{a+k}) \cdot s(e_{a+k}, f_{b+k})}{\sum_a \sum_{k=0}^{n-1} w(e_{a+k})}$$

$$s_r\left(\vec{e}, \vec{f}\right) = \frac{\sum_b max_a \sum_{k=0}^{n-1} w(f_{b+k}) \cdot s(e_{a+k}, f_{b+k})}{\sum_b \sum_{k=0}^{n-1} w(f_{b+k})}$$

$q_{i,j}^E$ = set of ARG j of aligned frame i in E

$q_{i,j}^F$ = set of ARG j of aligned frame i in F

$$w_i^E = \frac{\text{\#tokens filled in aligned frame } i \text{ of E}}{\text{total \# tokens in E}}$$

$$w_i^F = \frac{\text{\#tokens filled in aligned frame } i \text{ of F}}{\text{total \# tokens in F}}$$

$$w_{pred} = 0.25 \cdot \text{count(pred in } \mathcal{F})$$

$$w_j = \text{count(ARG } j \text{ in } \mathcal{F})$$

$$srl_p = \frac{\sum_i w_i^E \frac{w_{pred} + \sum_j w_j s_p\left(\overrightarrow{e_{i,j}}, \overrightarrow{f_{i,j}}\right)}{w_{pred} + \sum_j w_j \left|q_{i,j}^E\right|}}{\sum_i w_i^E}$$

$$srl_r = \frac{\sum_i w_i^F \frac{w_{pred} + \sum_j w_j s_r\left(\overrightarrow{e_{i,j}}, \overrightarrow{f_{i,j}}\right)}{w_{pred} + \sum_j w_j \left|q_{i,j}^F\right|}}{\sum_i w_i^F}$$

$$\text{precision} = \beta srl_p + (1-\beta) s_p\left(\overrightarrow{e_{sent}}, \overrightarrow{f_{sent}}\right)$$

$$\text{recall} = \beta srl_r + (1-\beta) s_r\left(\overrightarrow{e_{sent}}, \overrightarrow{f_{sent}}\right)$$

$$\text{YiSi} = \frac{\text{precision} \cdot \text{recall}}{\alpha \cdot \text{precision} + (1-\alpha) \cdot \text{recall}}$$

YiSi-1
MT evaluation metric

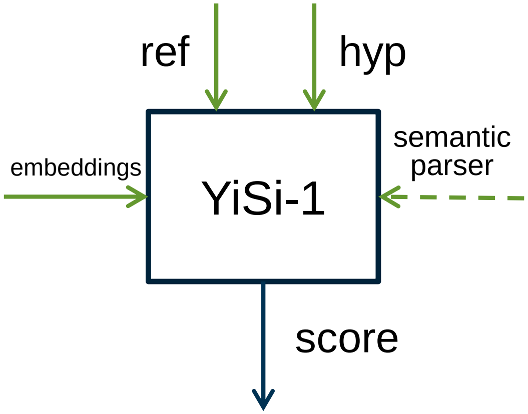
Requirements:

- Human reference translation for evaluating the lexical semantic similarity and the importance weighting
  - Inverse document frequency in the reference document

$$w(u) = idf(u) = \log(1 + \frac{|\mathbb{F}| + 1}{|\mathbb{F}_{\ni u}| + 1})$$

- Large monolingual corpus for building the word embeddings
  - Word embedding models, e.g. `word2vec` (Mikolov et al., 2013)
  - Contextual embeddings, e.g. BERT (Devlin et al., 2018)

$$v(u) = \text{embedding of unit } u$$
$$s(e,f) = \cos\big(v(e), v(f)\big)$$

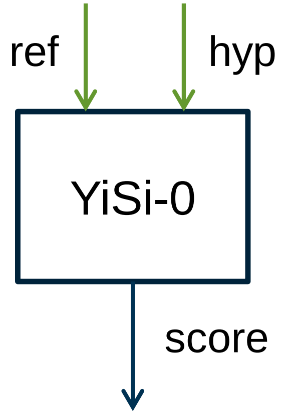


YiSi-0
MT evaluation metric for low-resource languages

What if we do not have access to a large monolingual corpus?

- Using longest common substring to determine the word similarity

$$l(e,f) = \text{longest common substring of } e \text{ and } f$$
$$s(e,f) = \frac{2 \cdot l(e,f)}{|e| + |f|}$$



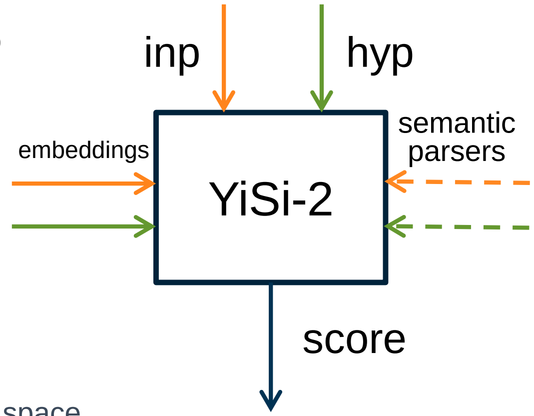
YiSi-2
MT quality estimation metric

What if we don’t have access to a human reference translation?

- Word weights are estimated in the input and output language respectively

$$w(u) = idf(u) = \log(1 + \frac{|\mathbb{F}| + 1}{|\mathbb{F}_{\ni u}| + 1}) \Rightarrow \begin{array}{l} w(e) = idf(e) = \log(1 + \frac{|\mathbb{E}| + 1}{|\mathbb{E}_{\ni e}| + 1}) \\ w(f) = idf(f) = \log(1 + \frac{|\mathbb{F}| + 1}{|\mathbb{F}_{\ni f}| + 1}) \end{array}$$

- Using bilingual/multilingual embeddings
  - `bivec` (Luong et al., 2015)
    - using parallel sentences
  - `vecmap` (Artetxe et al., 2016)
    - using dictionary pairs to transform two monolingual embedding models into the same space



WMT18
Metrics task

Performance in MT evaluation

Correlation with human at segment level

	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en
Human Evaluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR
n	5,110	77,811	56,721	15,648	10,404	8,525	33,357
Correlation	τ	τ	τ	τ	τ	τ	τ
BEER	0.295	0.481	0.341	0.232	**0.288**	**0.229**	**0.214**
BLEND	**0.322**	**0.492**	0.354	0.226	**0.290**	**0.232**	**0.217**
CHARACTER	0.256	0.450	0.286	0.185	0.244	0.172	**0.202**
CHRF	0.288	0.479	0.328	0.229	0.269	0.210	**0.208**
CHRF+	0.288	0.479	0.332	0.234	0.279	0.218	**0.207**
ITER	0.198	0.396	0.235	0.128	0.139	-0.029	0.144
METEOR++	0.270	0.457	0.329	0.207	0.253	0.204	0.179
RUSE	**0.347**	**0.498**	**0.368**	**0.273**	**0.311**	**0.259**	**0.218**
SENTBLEU	0.233	0.415	0.285	0.154	0.228	0.145	0.178
UHH_TSKM	0.274	0.436	0.300	0.168	0.235	0.154	0.151
YISI-0	0.301	0.474	0.330	0.225	**0.294**	0.215	**0.205**
YISI-1	**0.319**	0.488	0.351	0.231	**0.300**	**0.234**	**0.211**
YISI-1_SRL	**0.317**	0.483	0.345	0.237	**0.306**	**0.233**	**0.209**

	en-cs	en-de	en-et	en-fi	en-ru	en-tr	en-zh
Human Evaluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR
n	5,413	19,711	32,202	9,809	22,181	1,358	28,602
Correlation	τ	τ	τ	τ	τ	τ	τ
BEER	**0.518**	**0.686**	0.558	**0.511**	**0.403**	**0.374**	0.302
BLEND	–	–	–	–	**0.394**	–	–
CHARACTER	0.414	0.604	0.464	0.403	0.352	**0.404**	**0.313**
CHRF	**0.516**	0.677	**0.572**	**0.520**	0.383	**0.409**	**0.328**
CHRF+	**0.513**	**0.680**	**0.573**	**0.525**	**0.392**	**0.405**	**0.328**
ITER	0.333	0.610	0.392	0.311	0.291	0.236	–
SENTBLEU	0.389	0.620	0.414	0.355	0.330	0.261	**0.311**
YISI-0	0.471	0.661	0.531	0.464	**0.394**	**0.376**	**0.318**
YISI-1	**0.496**	**0.691**	0.546	**0.504**	**0.407**	**0.418**	**0.323**
YISI-1_SRL	–	**0.696**	–	–	–	–	**0.310**

Correlation with human at system level

	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en
n	5	16	14	9	8	5	14
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
BEER	**0.958**	0.994	**0.985**	**0.991**	0.982	0.870	**0.976**
BLEND	**0.973**	0.991	0.985	**0.994**	**0.993**	**0.801**	**0.976**
BLEU	**0.970**	0.971	**0.986**	0.973	0.979	**0.657**	**0.978**
CDER	**0.972**	0.980	**0.990**	0.984	0.980	**0.664**	**0.982**
CharacTER	**0.970**	**0.993**	0.979	0.989	**0.991**	**0.782**	0.950
chrF	0.966	0.994	0.981	0.987	**0.990**	0.452	0.960
chrF+	0.966	0.993	0.981	0.989	**0.990**	0.174	0.964
ITER	**0.975**	0.990	0.975	**0.996**	0.937	**0.861**	**0.980**
meteor++	**0.945**	0.991	0.978	0.971	**0.995**	0.864	0.962
NIST	**0.954**	0.984	0.983	0.975	0.973	**0.970**	0.968
PER	**0.970**	0.985	**0.983**	**0.993**	0.967	0.159	0.931
RUSE	**0.981**	**0.997**	**0.990**	**0.991**	**0.988**	**0.853**	**0.981**
TER	**0.950**	0.970	**0.990**	0.968	0.970	0.533	**0.975**
UHH_TSKM	0.952	0.980	**0.989**	0.982	0.980	0.547	**0.981**
WER	**0.951**	0.961	**0.991**	0.961	0.968	0.041	**0.975**
YiSi-0	0.956	**0.994**	0.975	0.978	**0.988**	**0.954**	0.957
YiSi-1	**0.950**	0.992	0.979	0.973	**0.991**	**0.958**	0.951
YiSi-1 srl	**0.965**	**0.995**	0.981	0.977	**0.992**	**0.869**	0.962

	en-cs	en-de	en-et	en-fi	en-ru	en-tr	en-zh
n	5	16	14	12	9	8	14
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
BEER	**0.992**	**0.991**	**0.980**	**0.961**	**0.988**	**0.965**	0.928
BLEND	–	–	–	–	**0.988**	–	–
BLEU	0.995	0.981	0.975	**0.962**	0.983	0.826	0.947
CDER	0.997	0.986	**0.984**	**0.964**	**0.984**	0.861	0.961
CharacTER	**0.993**	**0.989**	0.956	**0.974**	0.983	0.833	**0.983**
chrF	**0.990**	**0.990**	**0.981**	**0.969**	**0.989**	0.948	0.944
chrF+	**0.990**	0.989	**0.982**	**0.970**	0.989	0.943	0.943
ITER	0.915	**0.984**	**0.981**	**0.973**	0.975	0.865	–
NIST	**0.999**	0.986	**0.983**	0.949	**0.990**	0.902	0.950
PER	0.991	0.981	0.958	0.906	**0.988**	0.859	0.964
TER	**0.997**	**0.988**	**0.981**	**0.942**	**0.987**	0.867	**0.963**
WER	**0.997**	**0.986**	**0.981**	**0.945**	0.985	0.853	0.957
YiSi-0	0.973	0.985	0.968	0.944	**0.990**	**0.990**	0.957
YiSi-1	**0.987**	0.985	**0.979**	0.940	**0.992**	**0.976**	0.963
YiSi-1 srl	–	**0.990**	–	–	–	–	0.952

WMT18
Parallel Corpus Filtering task

Application in identifying parallel sentences
(Joint work with Michel Simard, Darlene Stewart, Samuel Larkin, Cyril Goutte and Patrick Littell)

WMT18 parallel corpus filtering task

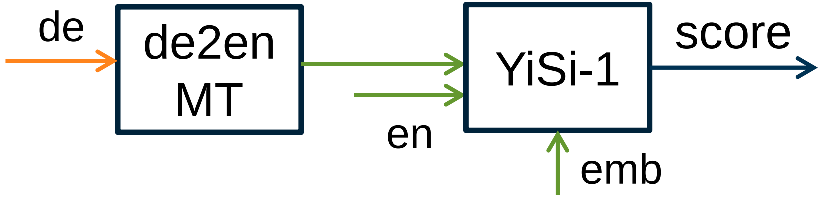
Ranking sentence pairs in web crawled noisy German-English corpora according to usefulness to train SMT/NMT system

Usefulness to train MT system

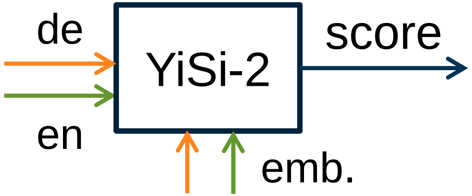
- Parallelism
- Fluency/grammaticality on both sides
- Coverage of source and target vocabularies

Internal experiments:
Precision on hand annotated subset

- YiSi-1: using SMT to translate German side into English



- YiSi-2: using `bivec` to build bilingual word representation works almost as well



features	precision
random	0.312
hunalign	0.624
YiSi-1 precision | recall	**0.796** | 0.763
YiSi-1_srl precision | recall	0.559 | 0.559
YiSi-2 precision | recall	0.753 | 0.731
YiSi-2_srl precision | recall	0.441 | 0.452
HMM $p(f\|e)$ | $p(e\|f)$	0.753 | 0.753
s2v d100 cosine	0.435
s2v Mahalanobis	0.634
perplexity ratio	0.538
POS perplexity ratio	0.441
perplexity de. | en.	0.419 | 0.355
POS perplexity de. | en.	0.376 | 0.462
regression	0.763

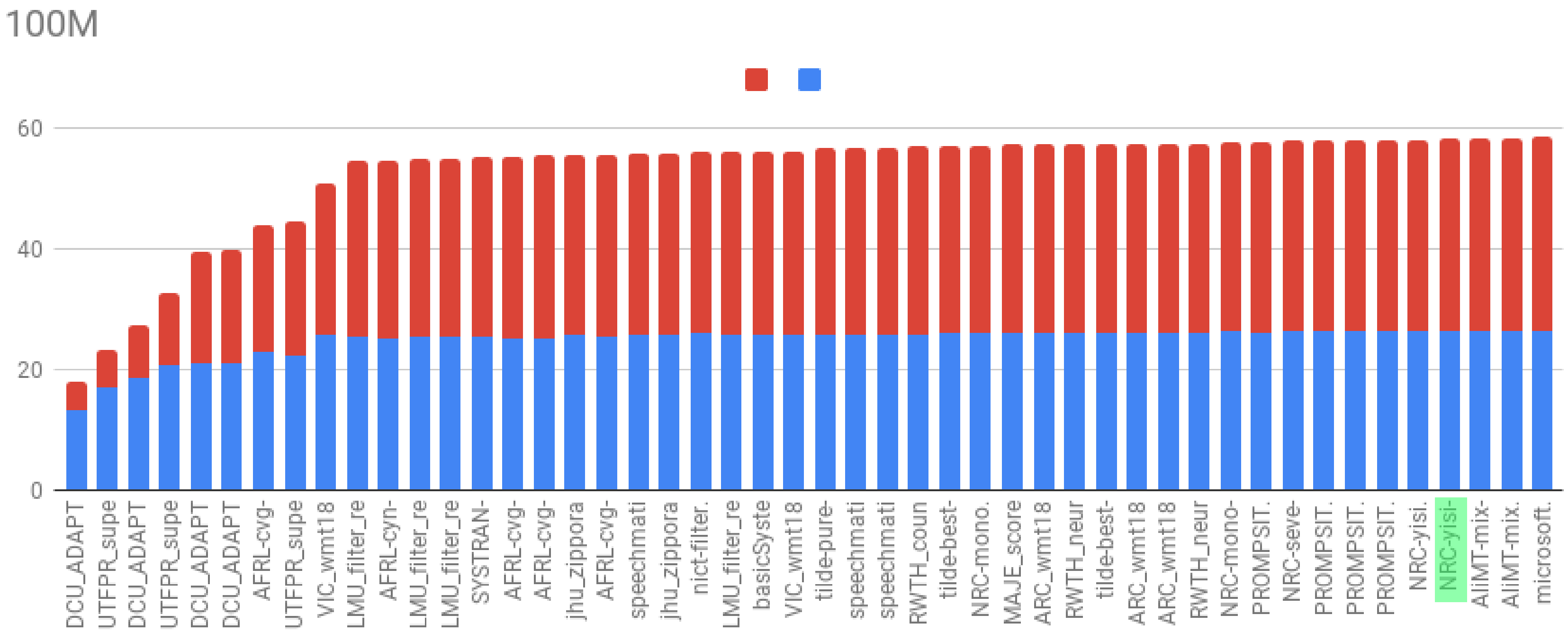
Internal experiments: MT quality check

Precision on hand annotated dev set positively correlated to the resulted MT quality

system	SMT				NMT			
	10M-word		100M-word		10M-word		100M-word	
	dev.	test	dev.	test	dev.	test	dev.	test
random	17.52	20.28	22.06	26.88	19.58	24.06	27.27	34.63
HMM $p(e\|f)$	19.09	23.55	24.42	29.73	21.16	26.59	31.53	39.52
HMM $p(e\|f)$ bicov	20.42	25.31	24.68	29.98	23.17	29.08	31.98	39.66
YiSi-1 precision (NRC-yisi)	21.56	24.68	24.47	30.10	24.24	30.75	32.49	40.27
YiSi-1 precision bicov (NRC-yisi-bicov)	**22.19**	**27.41**	**24.84**	**30.46**	**26.69**	**33.56**	**33.20**	**40.98**
regression bicov	21.86	26.97	**24.84**	30.27	25.28	31.94	31.30	39.34

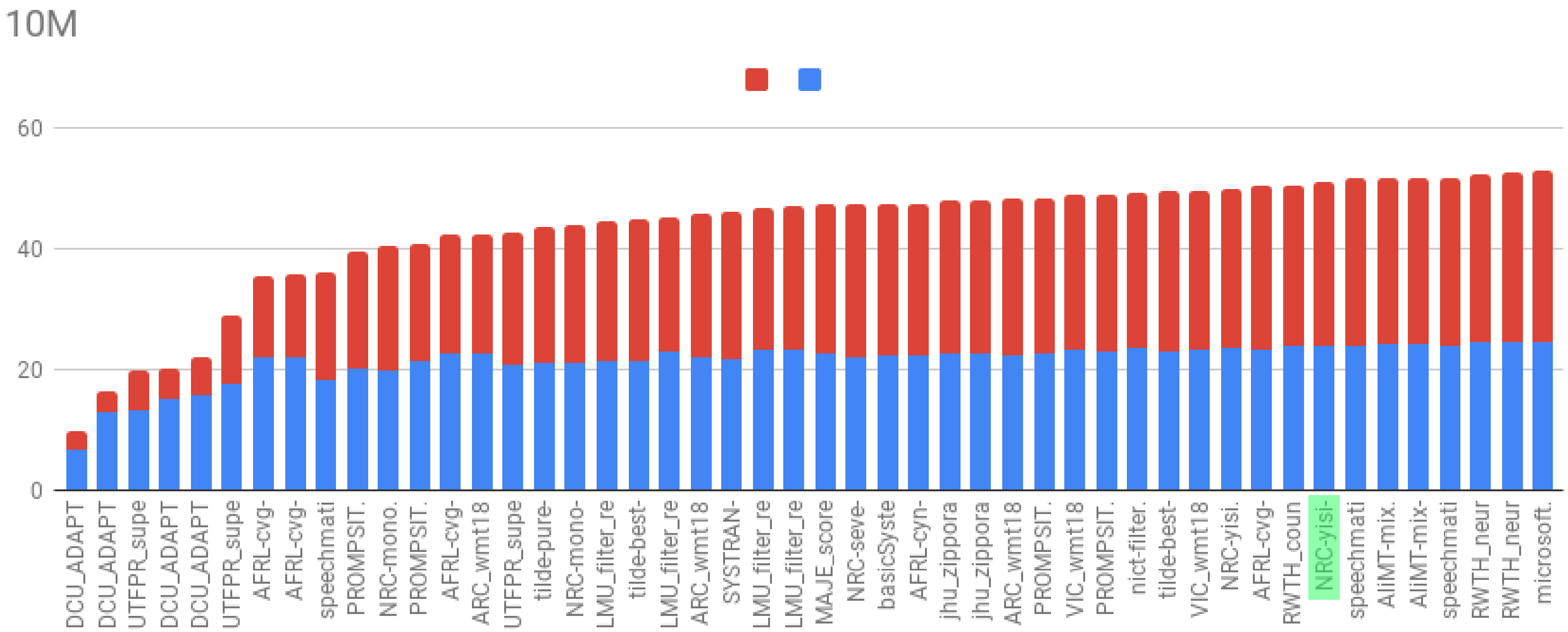
Table 2: BLEU scores of SMT and NMT systems trained on the 10M- and 100M-word corpora subselected by the scoring systems. "bicov" indicates that the final bigram coverage step (§2.4) was performed. The development set is newstest2017 and the test set is newstest2018.

100M performance



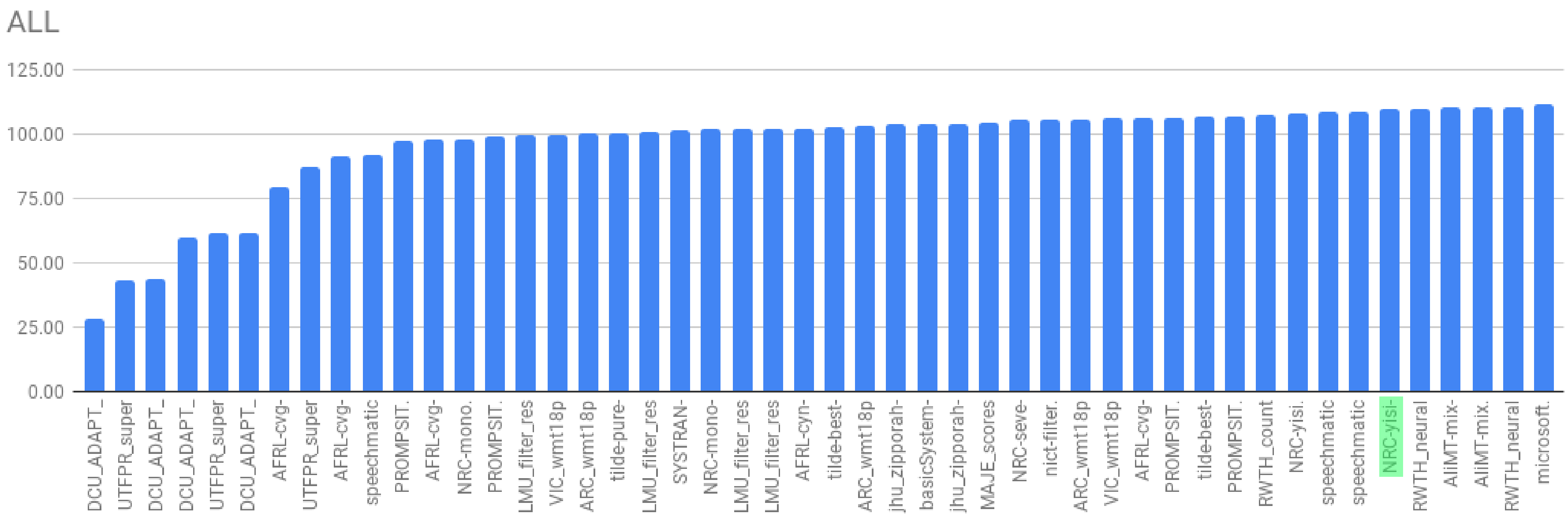
[Credit: Philipp Koehn]

10M performance



[Credit: Philipp Koehn]

Overall performance



[Credit: Philipp Koehn]

What’s new in YiSi this year?

YiSi-2 in WMT19 parallel corpus filtering task
(Joint work with Gabriel Bernier-Colborne)

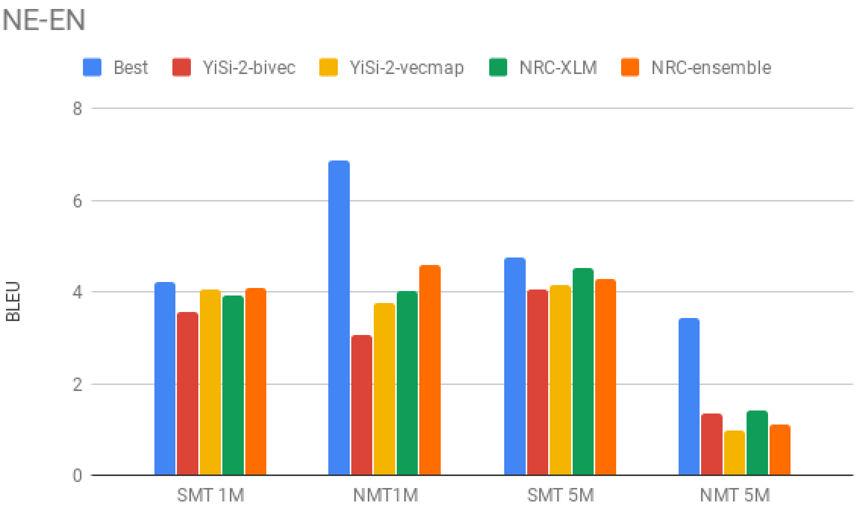
WMT19 parallel corpus filtering task

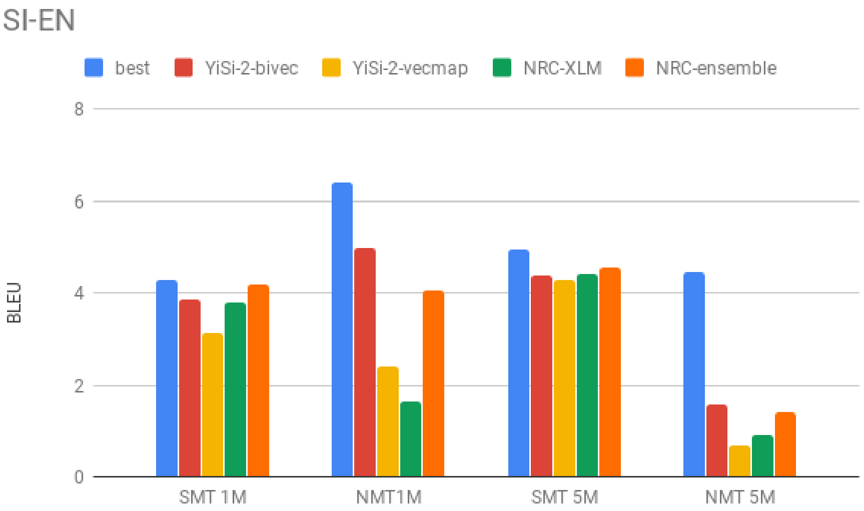
Ranking sentence pairs in web crawled noisy Nepali-English and Sinhala-English corpora according to usefulness to train SMT/NMT system

Challenges

- Low volume of “clean” parallel data: 500-600k pairs
- Domain mismatch
  - “Clean” data: IT/religious/subtitles
  - “Dirty” data: Anything from the web <- e.g. song lists, TV program…
  - MT system: Wikipedia

Burning hot results!





Rank	SMT 1M	NMT1M	SMT 5M	NMT 5M
NE-EN	10 **4** 7 **3**	12 8 7 **5**	14 12 **4** 8	10 16 9 14
SI-EN	6 13 8 **3**	**3** 12 15 **4**	9 10 7 **4**	9 15 13 11

What’s new in YiSi this year?

Using contextual embeddings from BERT

Using contextual embeddings from BERT

Word embeddings are static representation

- Not able to distinguish between different senses

Contextual embeddings are dynamic representation

- Captures the surrounding context of a word (subword unit)
  - Obtain different embeddings for the same subword unit appearing in different sentence

Burning hot results: Significantly improved correlation with human!

	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en
Human Evaluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR
n	5,110	77,811	56,721	15,648	10,404	8,525	33,357
Correlation	τ	τ	τ	τ	τ	τ	τ
BEER	0.295	0.481	0.341	0.232	**0.288**	**0.229**	**0.214**
BLEND	**0.322**	**0.492**	0.354	0.226	**0.290**	**0.232**	**0.217**
CHARACTER	0.256	0.450	0.286	0.185	0.244	0.172	**0.202**
CHRF	0.288	0.479	0.328	0.229	0.269	0.210	**0.208**
CHRF+	0.288	0.479	0.332	0.234	0.279	0.218	**0.207**
ITER	0.198	0.396	0.235	0.128	0.139	-0.029	0.144
METEOR++	0.270	0.457	0.329	0.207	0.253	0.204	0.179
RUSE	**0.347**	**0.498**	**0.368**	**0.273**	**0.311**	**0.259**	**0.218**
SENTBLEU	0.233	0.415	0.285	0.154	0.228	0.145	0.178
UHH_TSKM	0.274	0.436	0.300	0.168	0.235	0.154	0.151
YISI-0	0.301	0.474	0.330	0.225	**0.294**	0.215	**0.205**
YISI-1	**0.319**	0.488	0.351	0.231	**0.300**	**0.234**	**0.211**
YISI-1_SRL	**0.317**	0.483	0.345	0.237	**0.306**	**0.233**	**0.209**
YiSi-1-bert	0.391	0.544	0.397	0.299	0.352	0.301	0.254
YiSi-1-bert_srl	0.396	0.543	0.390	0.303	0.351	0.297	0.253

	en-cs	en-de	en-et	en-fi	en-ru	en-tr	en-zh
Human Evaluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR
n	5,413	19,711	32,202	9,809	22,181	1,358	28,602
Correlation	τ	τ	τ	τ	τ	τ	τ
BEER	**0.518**	**0.686**	0.558	**0.511**	**0.403**	**0.374**	0.302
BLEND	–	–	–	–	**0.394**	–	–
CHARACTER	0.414	0.604	0.464	0.403	0.352	**0.404**	**0.313**
CHRF	**0.516**	0.677	**0.572**	**0.520**	0.383	**0.409**	**0.328**
CHRF+	**0.513**	**0.680**	**0.573**	**0.525**	**0.392**	**0.405**	**0.328**
ITER	0.333	0.610	0.392	0.311	0.291	0.236	–
SENTBLEU	0.389	0.620	0.414	0.355	0.330	0.261	**0.311**
YISI-0	0.471	0.661	0.531	0.464	**0.394**	**0.376**	**0.318**
YISI-1	**0.496**	**0.691**	0.546	**0.504**	**0.407**	**0.418**	**0.323**
YISI-1_SRL	–	**0.696**	–	–	–	–	**0.310**
YiSi-1-bert	0.548	0.734	0.599	0.549	0.427	0.402	0.371
YiSi-1-bert_srl	--	0.719	--	--	--	--	0.368

At system level as well!

	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en
n	5	16	14	9	8	5	14
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
BEER	**0.958**	0.994	**0.985**	**0.991**	0.982	0.870	**0.976**
BLEND	**0.973**	0.991	0.985	**0.994**	**0.993**	**0.801**	**0.976**
BLEU	**0.970**	0.971	**0.986**	0.973	0.979	**0.657**	**0.978**
CDER	**0.972**	0.980	**0.990**	0.984	0.980	**0.664**	**0.982**
CharacTER	**0.970**	**0.993**	0.979	0.989	**0.991**	**0.782**	0.950
chrF	0.966	0.994	0.981	0.987	**0.990**	0.452	0.960
chrF+	0.966	0.993	0.981	0.989	**0.990**	0.174	0.964
ITER	**0.975**	0.990	0.975	**0.996**	0.937	**0.861**	**0.980**
meteor++	**0.945**	0.991	0.978	0.971	**0.995**	0.864	0.962
NIST	**0.954**	0.984	0.983	0.975	0.973	**0.970**	0.968
PER	**0.970**	0.985	**0.983**	**0.993**	0.967	0.159	0.931
RUSE	**0.981**	**0.997**	**0.990**	**0.991**	**0.988**	**0.853**	**0.981**
TER	**0.950**	0.970	**0.990**	0.968	0.970	0.533	**0.975**
UHH_TSKM	0.952	0.980	**0.989**	0.982	0.980	0.547	**0.981**
WER	**0.951**	0.961	**0.991**	0.961	0.968	0.041	**0.975**
YiSi-0	0.956	**0.994**	0.975	0.978	**0.988**	**0.954**	0.957
YiSi-1	**0.950**	0.992	0.979	0.973	**0.991**	**0.958**	0.951
YiSi-1_srl	**0.965**	**0.995**	0.981	0.977	**0.992**	**0.869**	0.962
YiSi-1-bert	0.990	0.998	0.986	0.994	0.993	0.830	0.988
YiSi-1-bert_srl	0.989	0.999	0.987	0.993	0.993	0.793	0.989

	en-cs	en-de	en-et	en-fi	en-ru	en-tr	en-zh
n	5	16	14	12	9	8	14
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
BEER	**0.992**	**0.991**	**0.980**	**0.961**	**0.988**	**0.965**	0.928
BLEND	–	–	–	–	**0.988**	–	–
BLEU	0.995	0.981	0.975	**0.962**	0.983	0.826	0.947
CDER	0.997	0.986	**0.984**	**0.964**	**0.984**	0.861	0.961
CharacTER	**0.993**	**0.989**	0.956	**0.974**	0.983	0.833	**0.983**
chrF	**0.990**	**0.990**	**0.981**	**0.969**	**0.989**	0.948	0.944
chrF+	**0.990**	0.989	**0.982**	**0.970**	0.989	0.943	0.943
ITER	0.915	**0.984**	**0.981**	**0.973**	0.975	0.865	–
NIST	**0.999**	0.986	**0.983**	0.949	**0.990**	0.902	0.950
PER	0.991	0.981	0.958	0.906	**0.988**	0.859	0.964
TER	**0.997**	**0.988**	**0.981**	**0.942**	**0.987**	0.867	**0.963**
WER	**0.997**	**0.986**	**0.981**	**0.945**	0.985	0.853	0.957
YiSi-0	0.973	0.985	0.968	0.944	**0.990**	**0.990**	0.957
YiSi-1	**0.987**	0.985	**0.979**	0.940	**0.992**	**0.976**	0.963
YiSi-1_srl	–	**0.990**	–	–	–	–	0.952
YiSi-1-bert	0.993	0.995	0.988	0.979	0.993	0.929	0.977
YiSi-1-bert_srl	--	0.995	--	--	--	--	0.976

What else is in progress?

Server-client mode of YiSi-2 for users to validate human translations

Extensive studies on contextual embeddings in YiSi

Integrating YiSi into popular NMT toolkits

Final words…

STOP reporting BLEU only!!!   щ(ºДºщ)

NRC·CNRC

NRC.CANADA.CA

YiSi – an open-source evaluation metric and its application

https://github.com/chikiulo/YiSi

羅致翹 Chi-kiu (Jackie) Lo

Multilingual Text Processing team