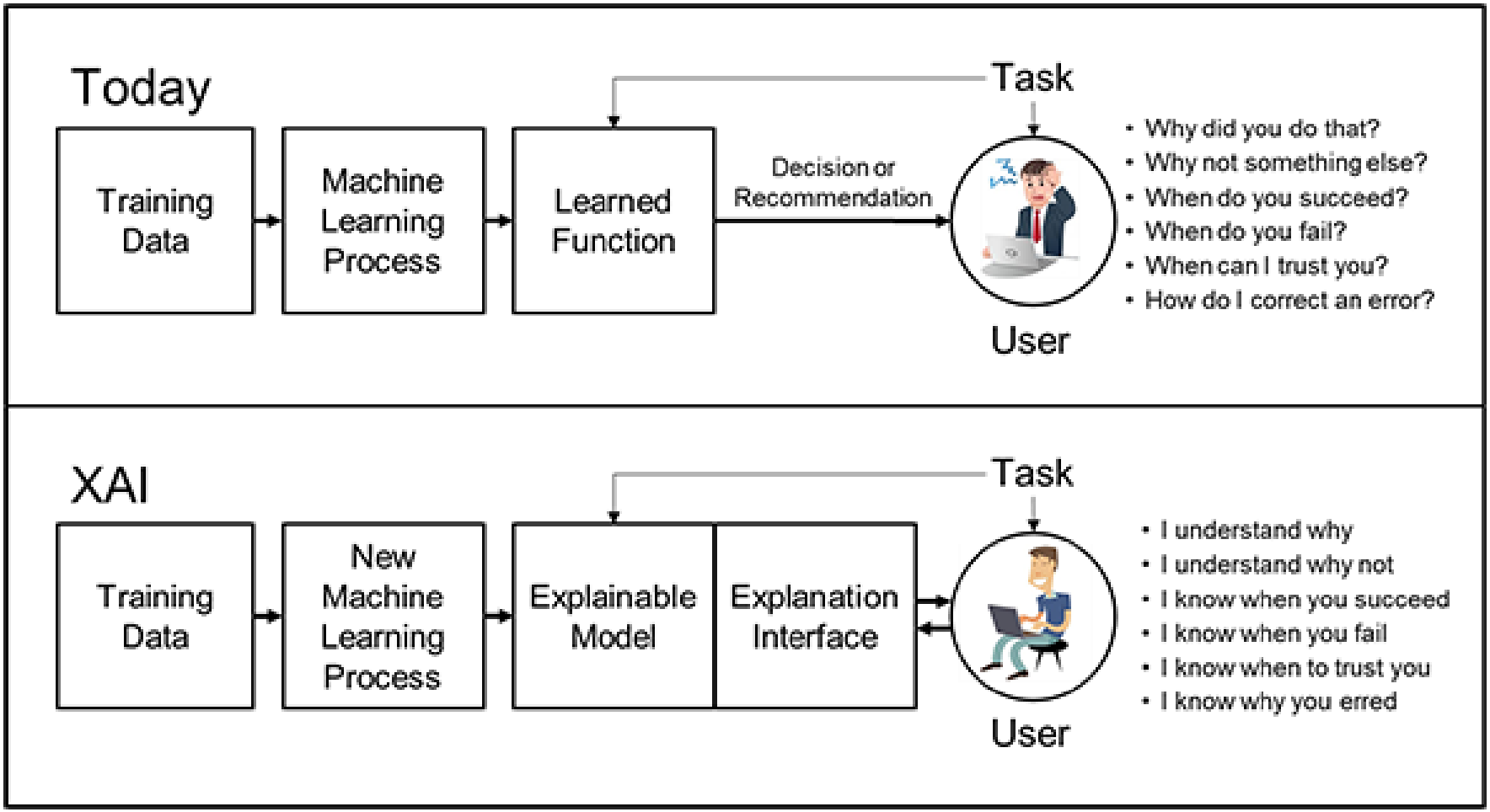
Adequacy-Fluency Evaluation of Natural Language

Li Haizhou (李海洲)
National University of Singapore (新加坡国立大学)

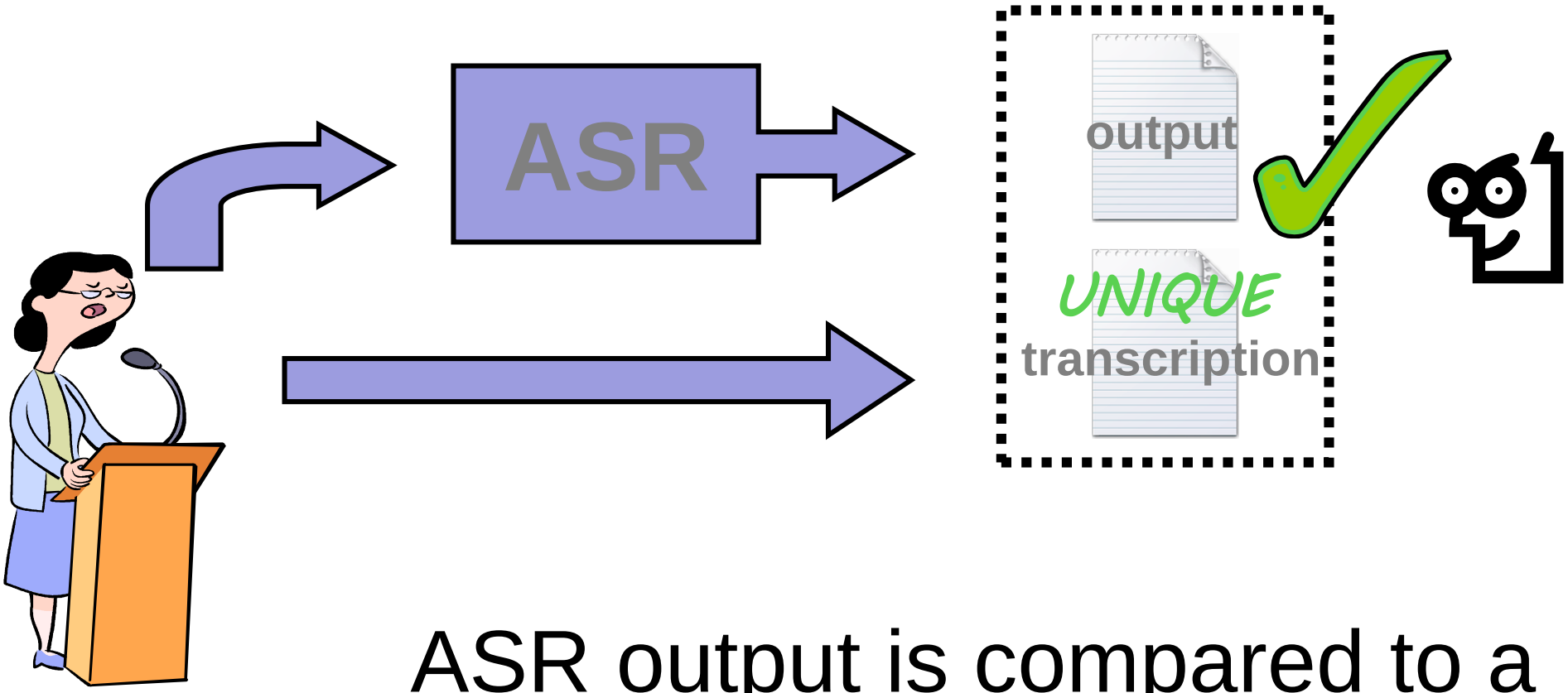
Acknowledgement: Rafael E. Banchs, Luis Fernando D'Haro

Explainable Artificial Intelligence



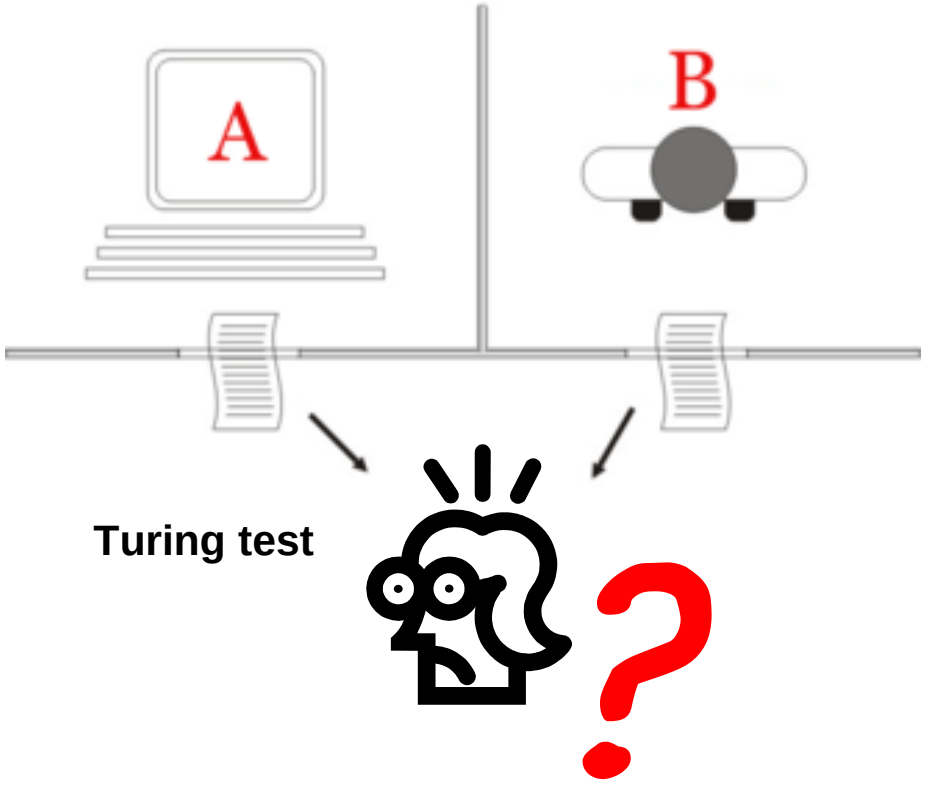
Dr. Matt Turek , Explainable Artificial Intelligence (XAI)
https://www.darpa.mil/program/explainable-artificial-intelligence

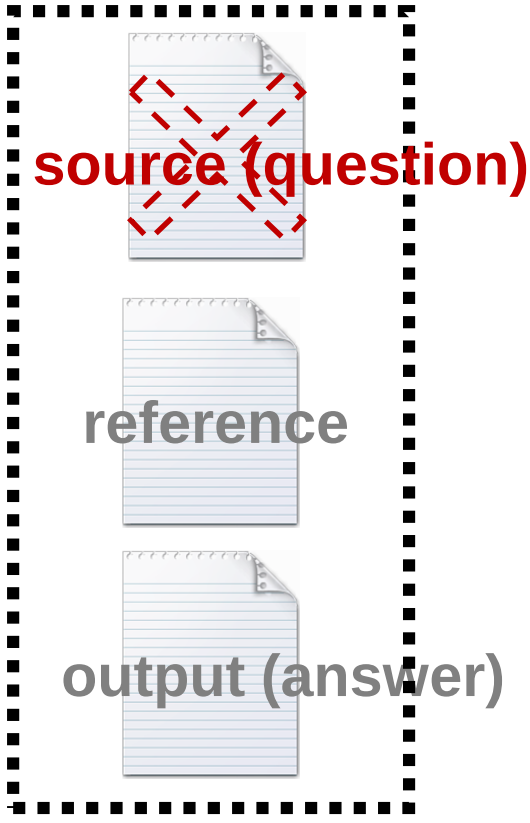
Automatic Evaluation of Automatic Speech Recognition



ASR output is compared to a **unique** reference transcription.

Automatic Evaluation of Dialogue Output





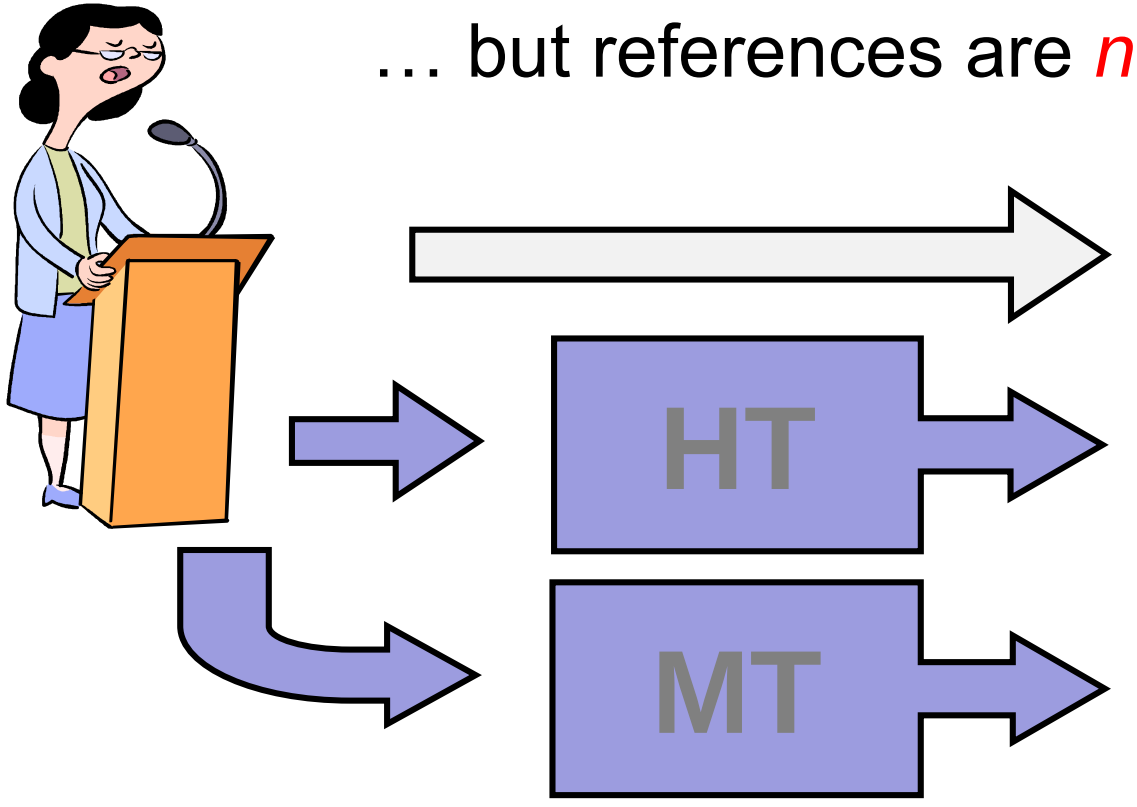
Dialogue output is compared to one or more reference answers,

… but references are *neither unique nor exact*!

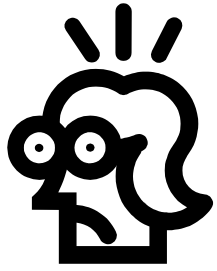
Automatic Evaluation of Machine Translation

MT output is compared to one or more reference translations.

… but references are *neither unique nor exact*!



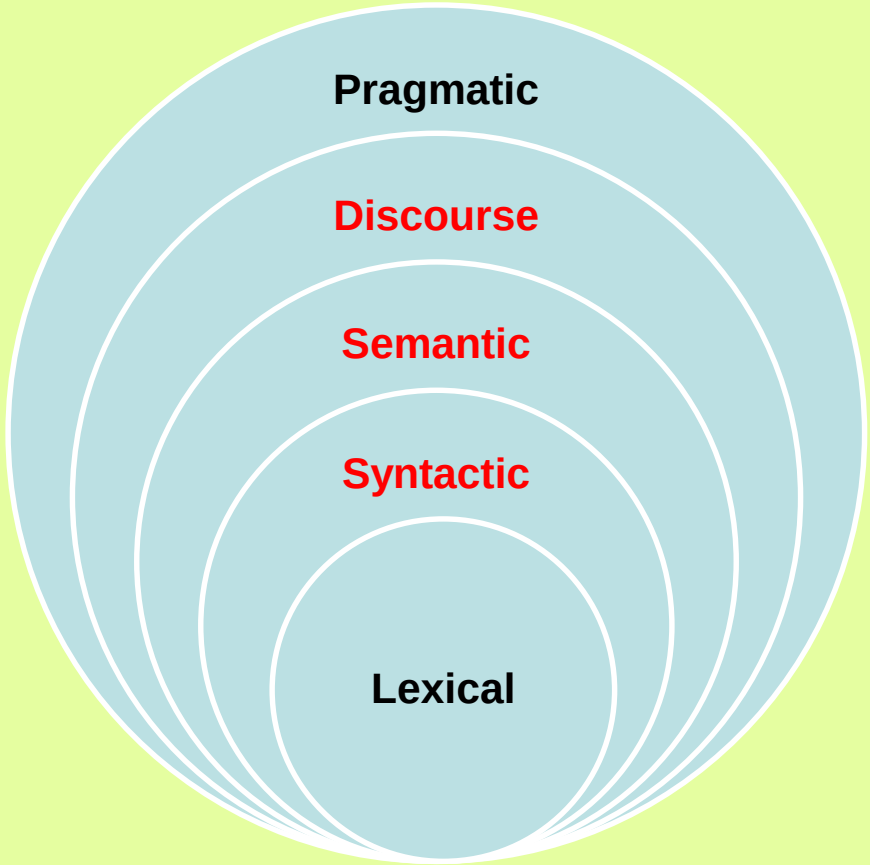
BLEU



K. Papineni *et al.*, “BLEU: a method for automatic evaluation of machine translation”, in *the 40th ACL*, Philadelphia, PA, USA, Jul 2002, pp. 311-318

What Do Humans Evaluate?

Dialogue Analysis

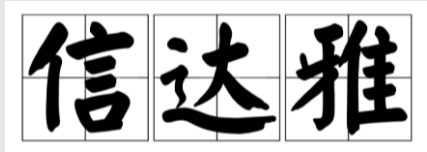


Machine Translation

Faithfulness

Expressiveness

Elegance



“译事三难：信、达、雅。求其信，已大难矣！顾信矣，不达，虽译，犹不译也，则达尚焉。”- 严复

Human Adequacy – Fluency MT Scores

Adequacy: Compare output （机器译文）with source （原文）and reference （参考译文）

Fluency: Mostly only look at the output （机器译文）

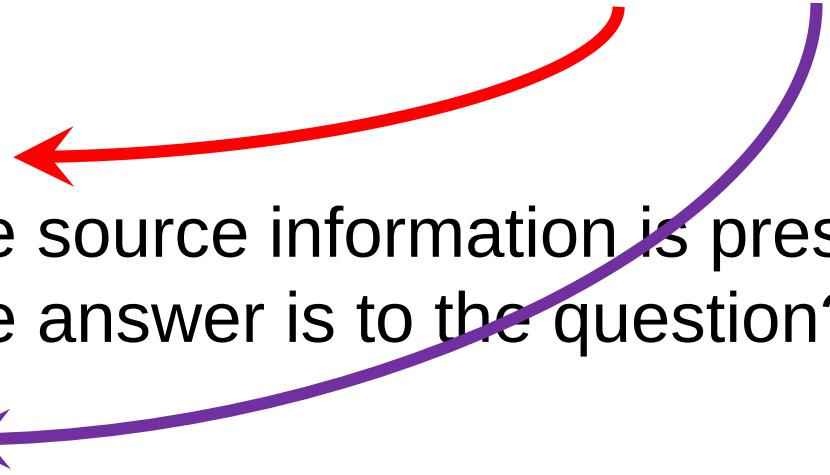
Metric	Score	Definition
ADEQUACY	1	None of the meaning is preserved
	2	Little of the meaning is preserved
	3	Much of the meaning is preserved
	4	Most of the meaning is preserved
	5	All the meaning is preserved
FLUENCY	1	Incomprehensible target language
	2	Disfluent target language
	3	Non-native kind of target language
	4	Good quality target language
	5	Flawless target language

* J.S. White, T. O'Cornell and F. O'Nava, "The ARPA MT evaluation methodologies: evolution, lessons and future approaches", in *Proc. of the Assoc. for Mach. Translation in the Amer.*, Oct 1994, pp. 193-205

Adequacy – Fluency

Dialogue Analysis: $P(A|Q) \approx P(Q|A)\ P(A)$

Machine Translation: $P(T|S) \approx P(S|T)\ P(T)$



ADEQUACY

How much of the source information is preserved?

How relevant the answer is to the question?

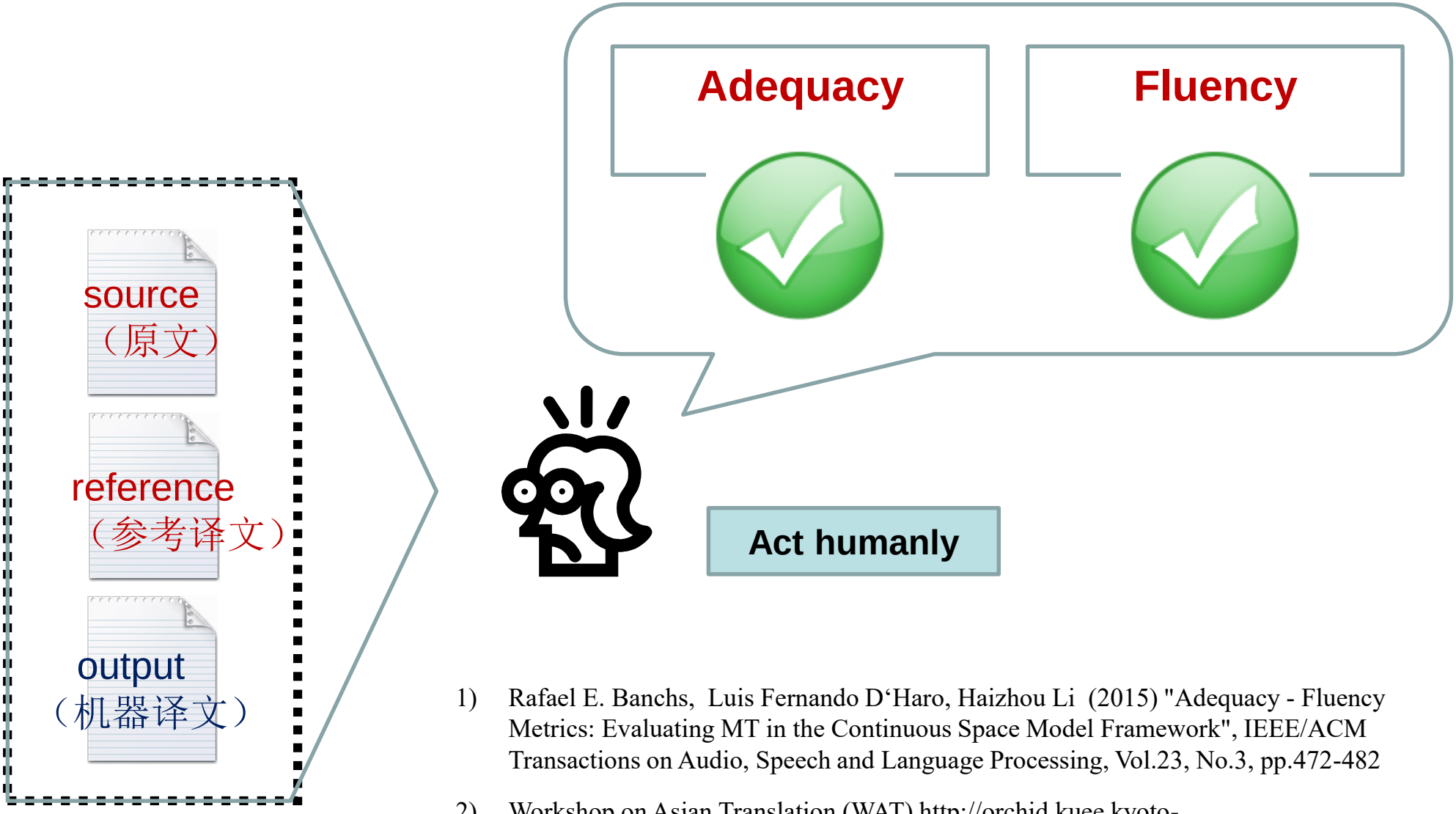
FLUENCY

How smooth a sentence is?


machine
language
translation
communication
code
word
automation
speech
symbol
field
industry
network
engineering
image
binary
laptop
tech
computer
process
downloading
robotic
translate
english
vocabulary
technology
global
cyberspace
order
emerging
fabricating
mathematics
glowing
display
virtual
generated
advice
internet
information
simulator
digital
robot
assembly
installation
software
graphic
online
linguistics
serial
text
format
stream
electronics
automatic
security
inscription
coding
digitally
equipment
technologies
mathematical
concept
twist
web
data


Translate server error
餐 厅


EXECUTION
IN PROGRESS
施工進行中



1) Rafael E. Banchs, Luis Fernando D'Haro, Haizhou Li (2015) "Adequacy - Fluency Metrics: Evaluating MT in the Continuous Space Model Framework", IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol.23, No.3, pp.472-482

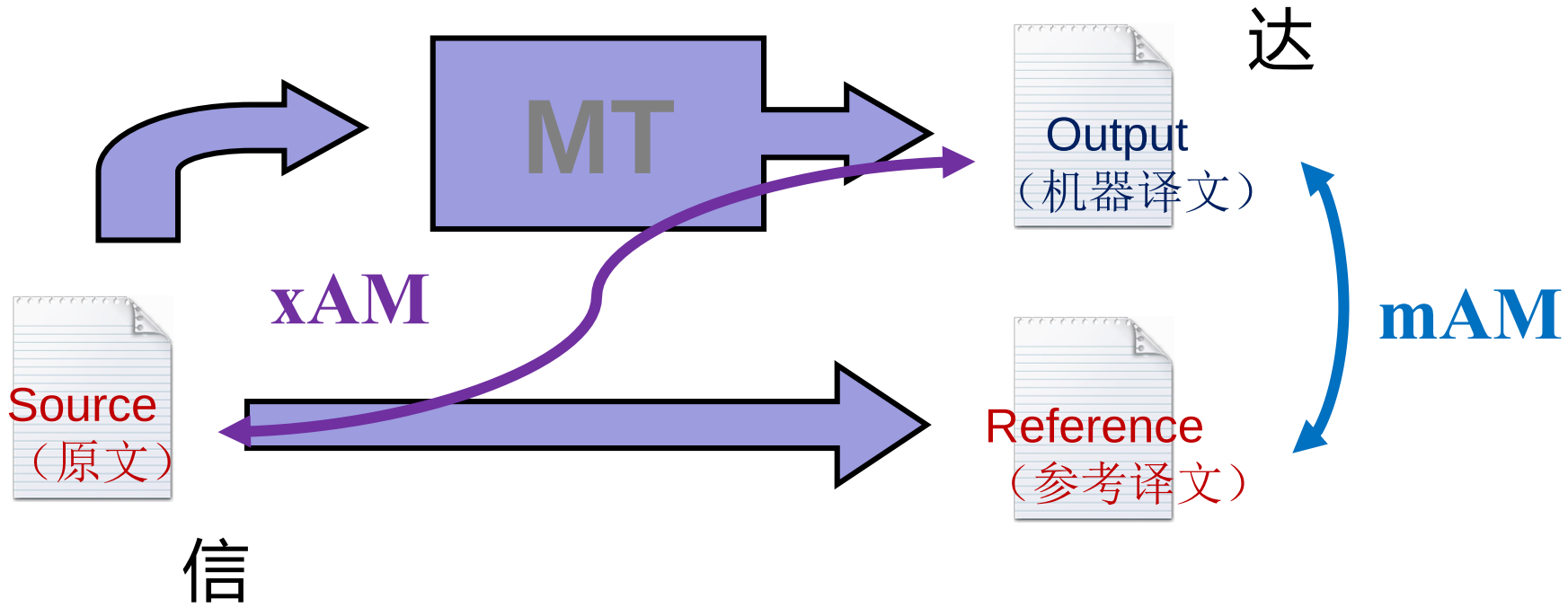
2) Workshop on Asian Translation (WAT) http://orchid.kuee.kyoto-u.ac.jp/WAT/WAT2019/index.html

AM: Adequacy-oriented Metric

Compare sentences in a semantic space

Monolingual AM (***mAM***): compare output vs. reference

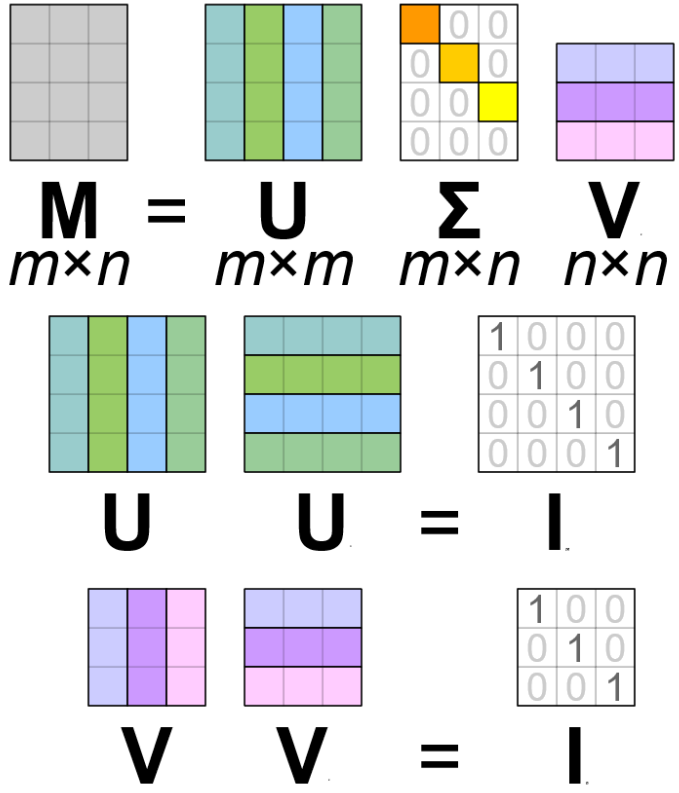
Cross-language AM (***xAM***): compare output vs. source



Latent Semantic Indexing by SVD

$$M_{M\times N} = U_{M\times M}\ \Sigma_{M\times N}\ V^{T}_{N\times N}$$

Term-Document Matrix



By Cmglee - Own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=67853297

Latent Semantic Indexing (mAM)

Documents projected into word space

$$U^{T}_{M\times M}\ M_{M\times N} = D_{M\times N}$$

Documents projected into reduced word space

$$U^{T}_{K\times M}\ M_{M\times N} = D_{K\times N}$$

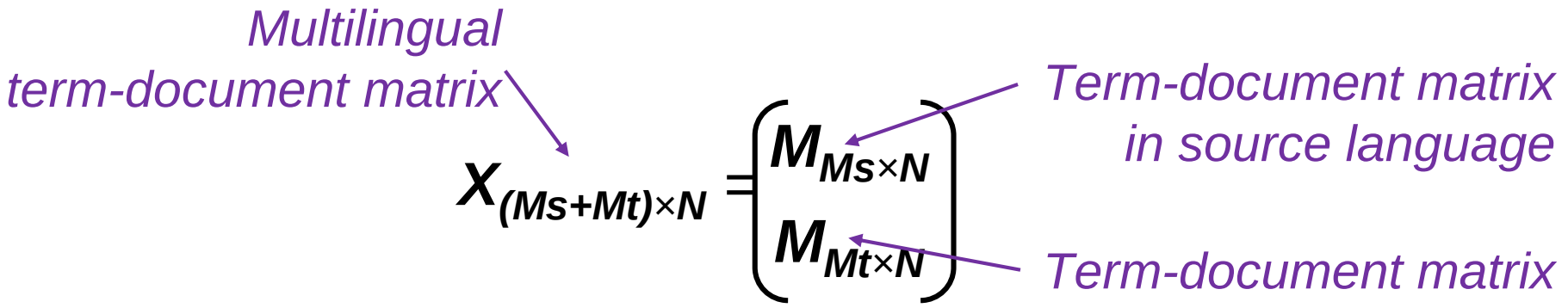
Translation output (to) and translation reference (tr) compared in reduced vector space

$$\langle U^{T}_{K\times M}\ to_{M\times 1}\ ,\ U^{T}_{K\times M}\ tr_{M\times 1} \rangle$$

$$U^{T}_{K\times M} = \begin{pmatrix} u_{11} \dots u_{1k} \dots u_{m1} \\ u_{21} \dots u_{2k} \dots u_{m2} \\ \vdots \quad \vdots \quad \vdots \\ u_{m1} \dots u_{mk} \dots u_{mm} \end{pmatrix}^{T}$$

* Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990), Indexing by latent semantic analysis, Journal of the American Society for Information Science, 41, pp.391-407

Cross-Language LSI (xAM)



SVD: $X = U \Sigma V^T$

Translation output (**to**) and translation input (**ti**) compared in cross-language vector space

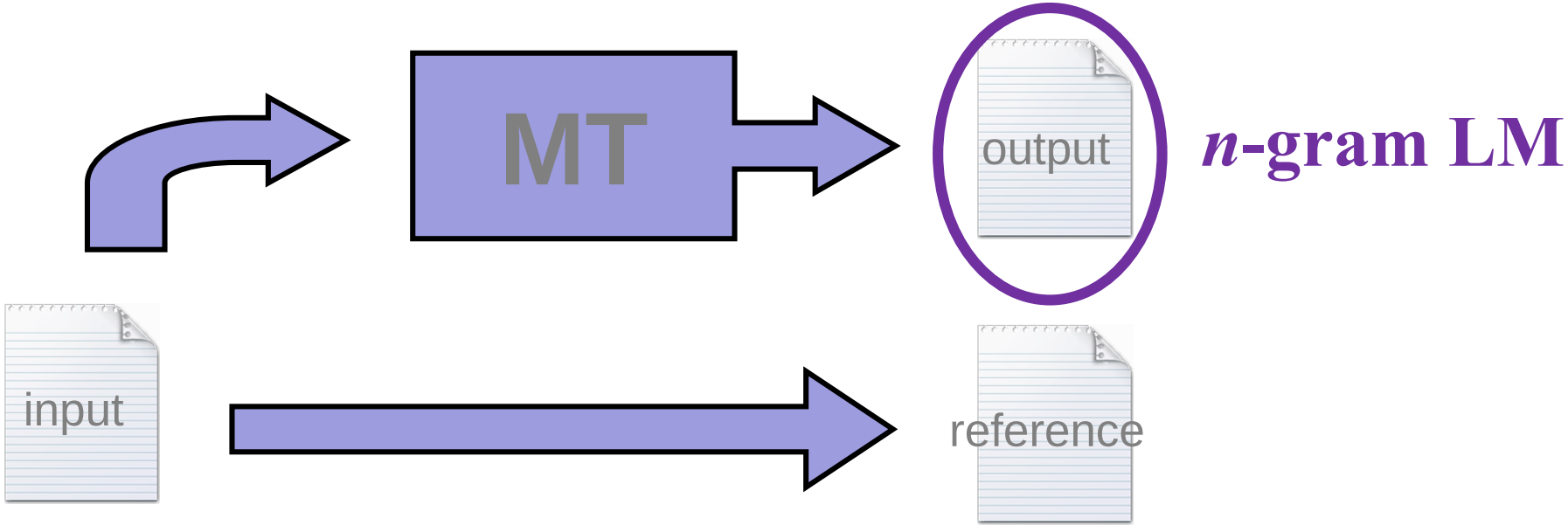
$$< U^T_{K\times(Ms+Mt)} \begin{bmatrix} 0_{Ms\times 1} \\ to_{Mt\times 1} \end{bmatrix}, U^T_{K\times(Ms+Mt)} \begin{bmatrix} ti_{Ms\times 1} \\ 0_{Mt\times 1} \end{bmatrix} >$$

* Dumais S.T., Letsche T.A., Littman M.L. and Landauer T.K. (1997), Automatic Cross-Language Retrieval Using Latent Semantic Indexing, in AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval, pp. 18-24

FM: Fluency-oriented Metric

Measures the quality （雅）of the target language with a language model

$$FM = \exp\left(\frac{1}{N}\sum_{n=1:N}\log\left(p(w_n \mid w_{n-1},\ldots)\right)\right)$$



AM-FM Combined Score

Both components can be combined into a single metric according to different criteria

Weighted Harmonic Mean: $$H\text{-}AM\text{-}FM = \frac{AM \cdot FM}{\alpha\, AM + (1-\alpha)\, FM}$$

Weighted Mean: $$M\text{-}AM\text{-}FM = (1-\alpha)\, AM + \alpha\, FM$$

Weighted L2-norm: $$N\text{-}AM\text{-}FM = \sqrt{(1-\alpha)\, AM^2 + \alpha\, FM^2}$$

*WMT-2007 Dataset**

Fourteen tasks:

five European languages (EN, ES, DE, FR, CZ) and

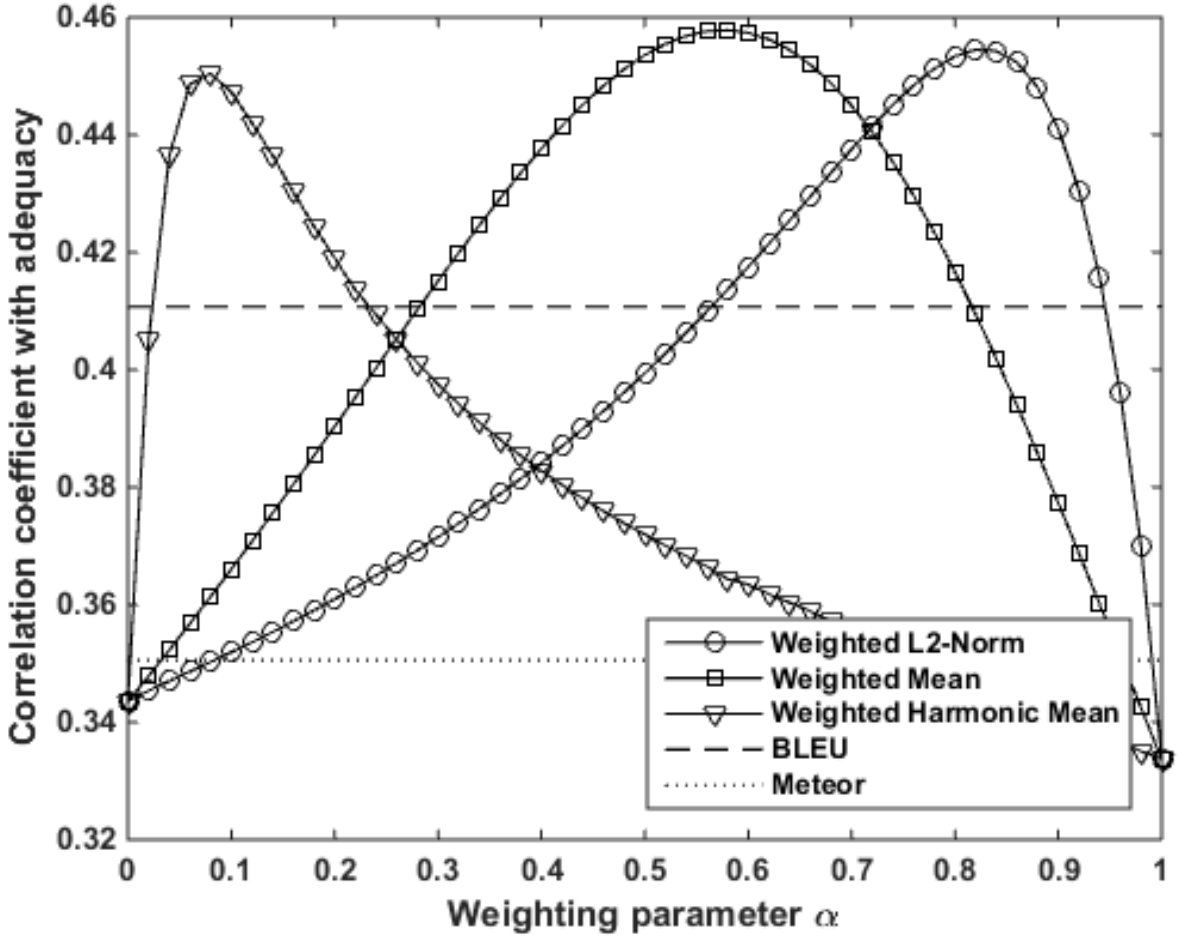
two different domains (News and EPPS).

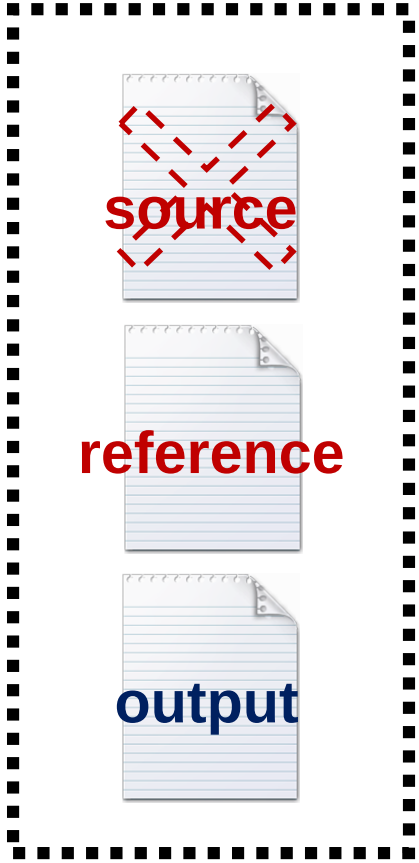
Systems outputs available from 14 teams that had participated in the evaluation. In total, 86 system outputs.

Overall 172,315 individual sentence translations, from which a total of 10,754 were rated for both adequacy and fluency by human judges.

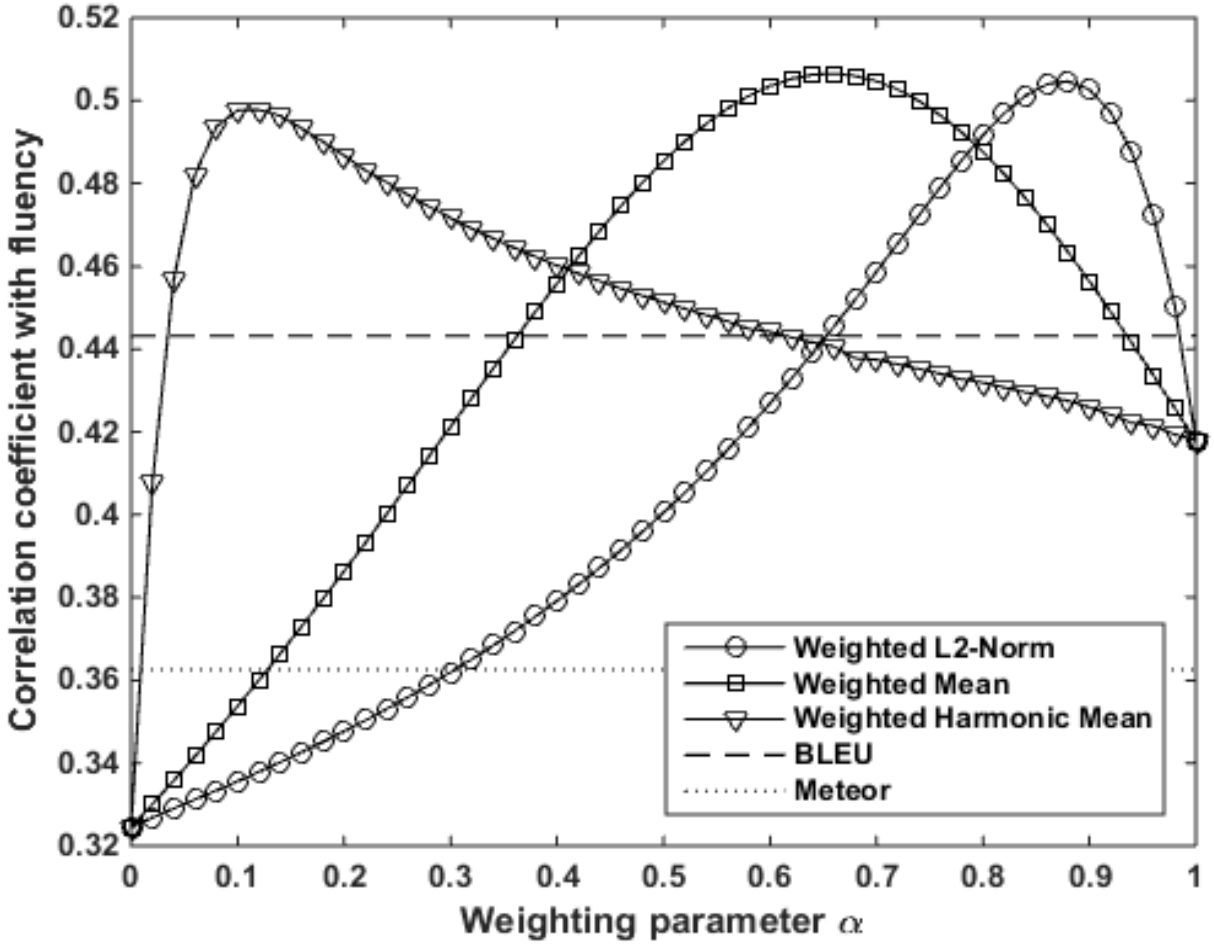
* Callison-Burch C., Fordyce C., Koehn P., Monz C. and Schroeder J. (2007), (Meta-) evaluation of machine translation, in Proceedings of Statistical Machine Translation Workshop, pp. 136-158

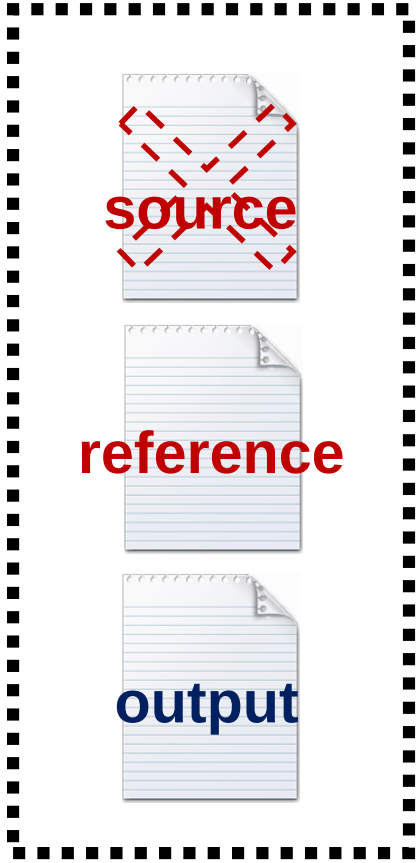
mAM-FM and Adequacy



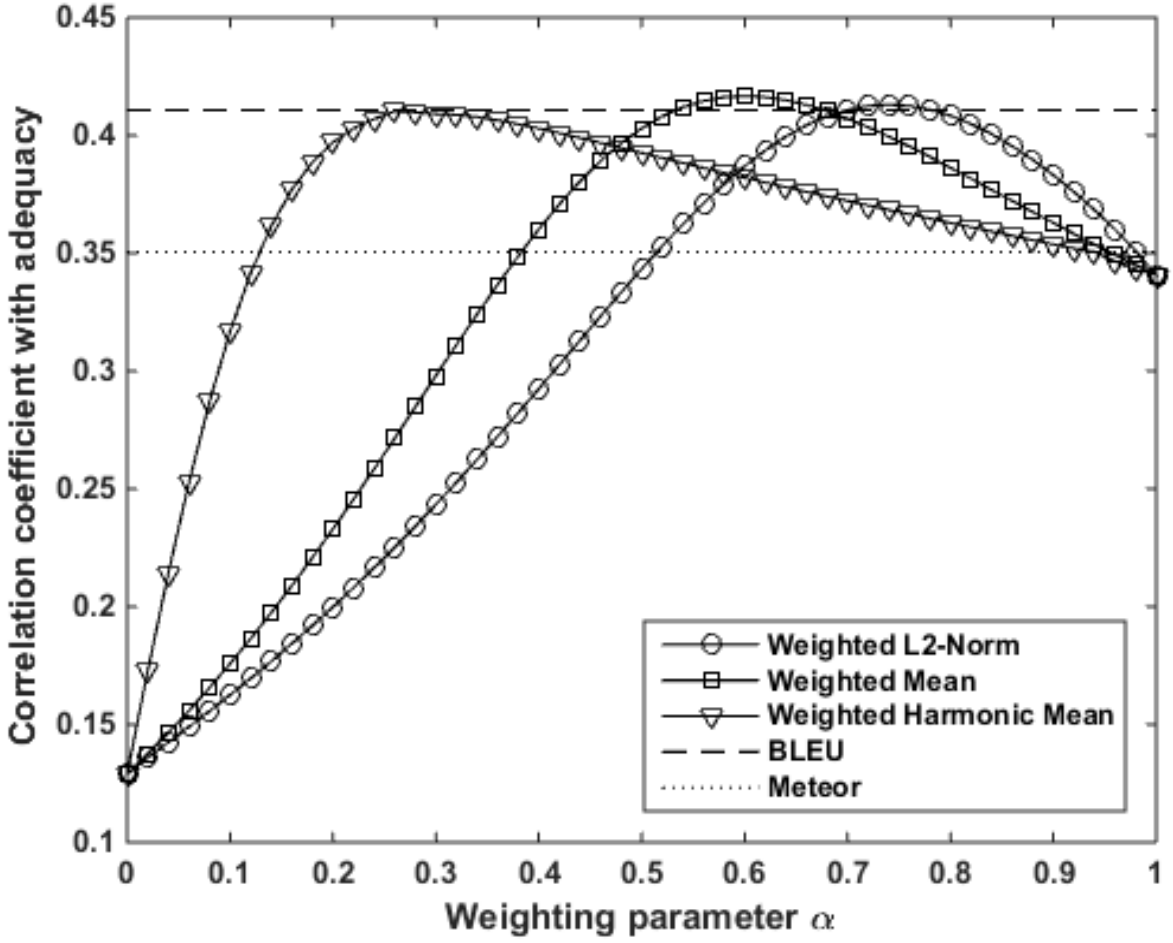


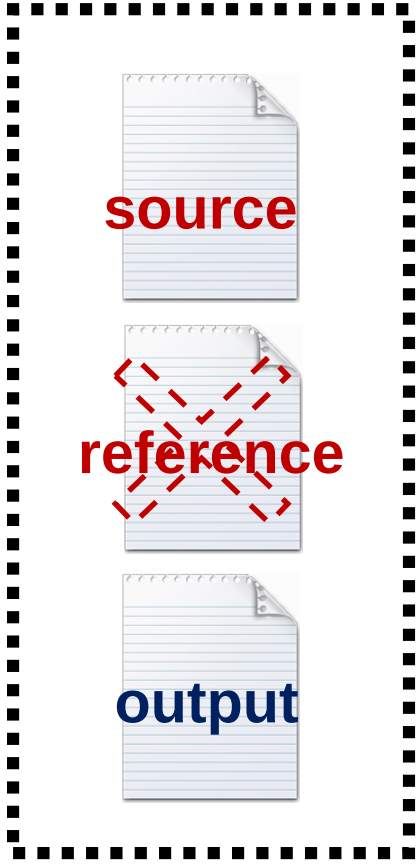
mAM-FM and Fluency



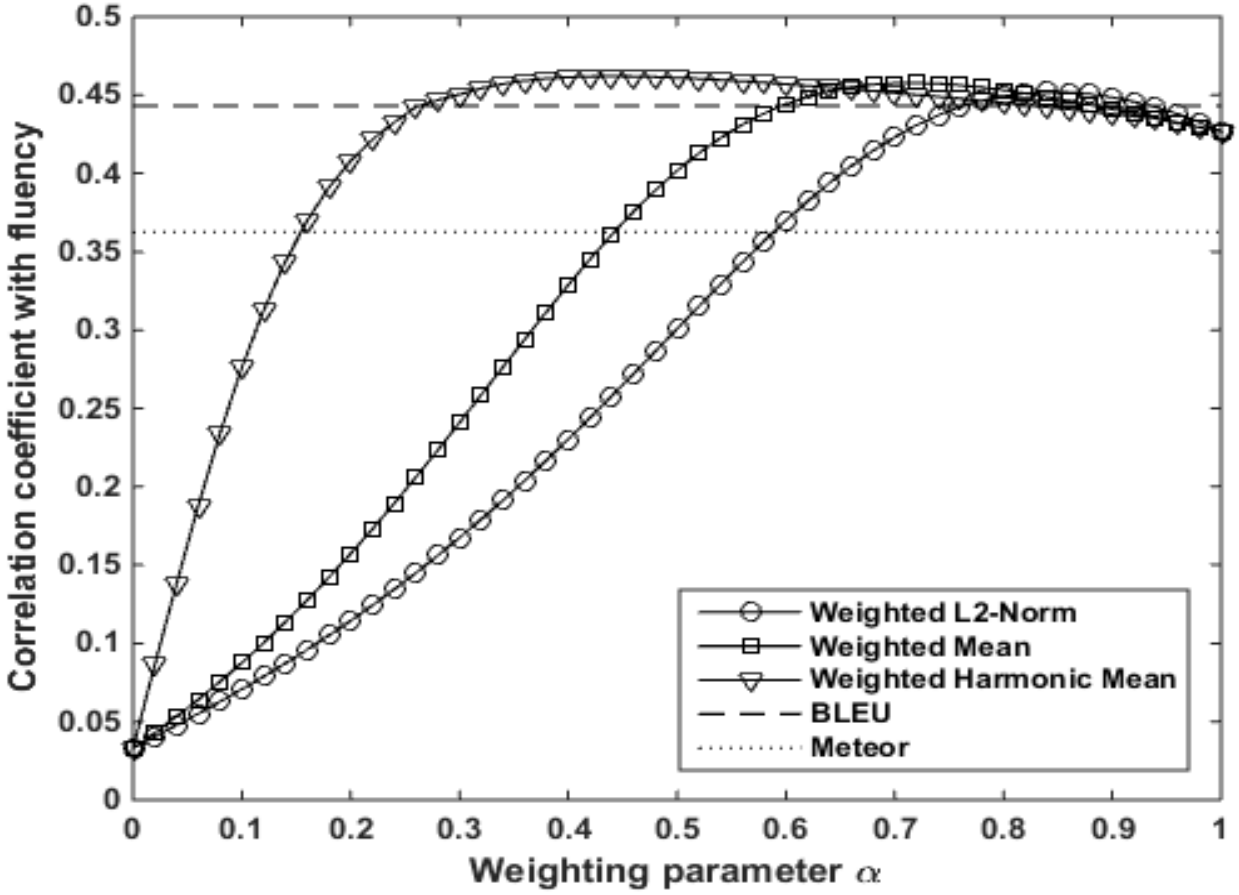


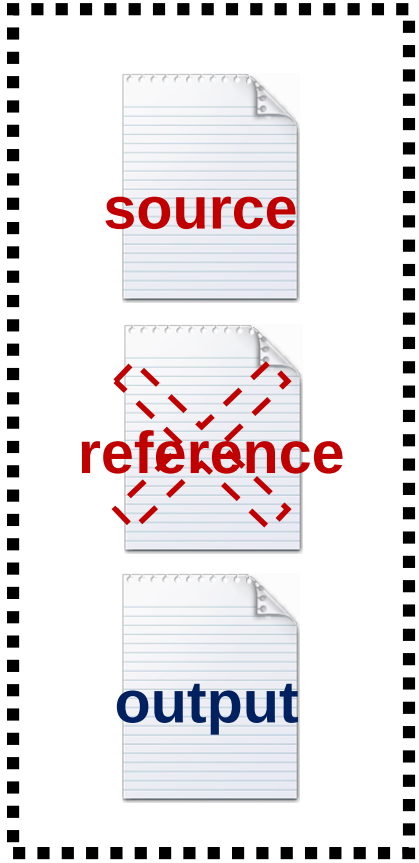
xAM-FM and Adequacy





xAM-FM and Fluency



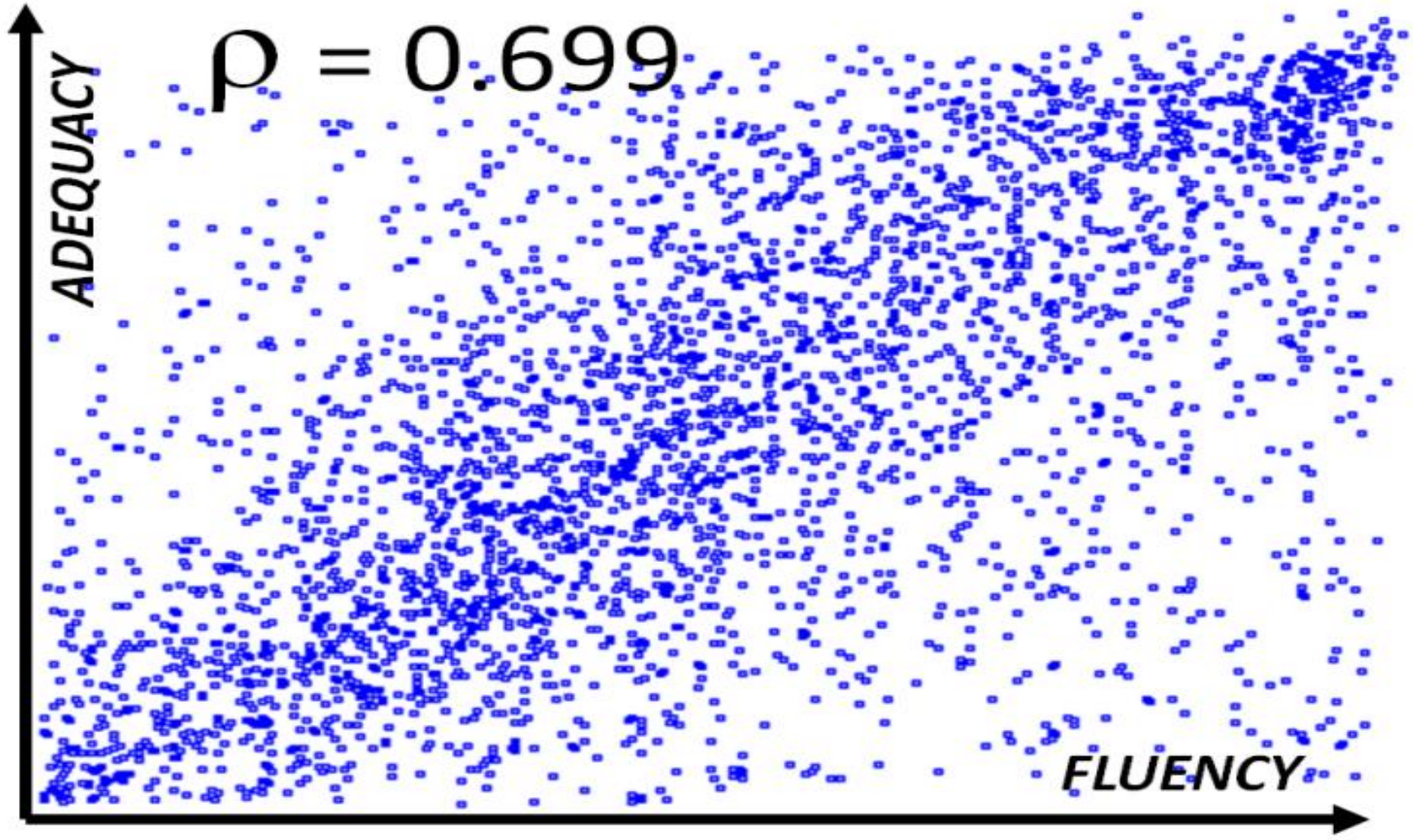


Comparative Evaluation Results

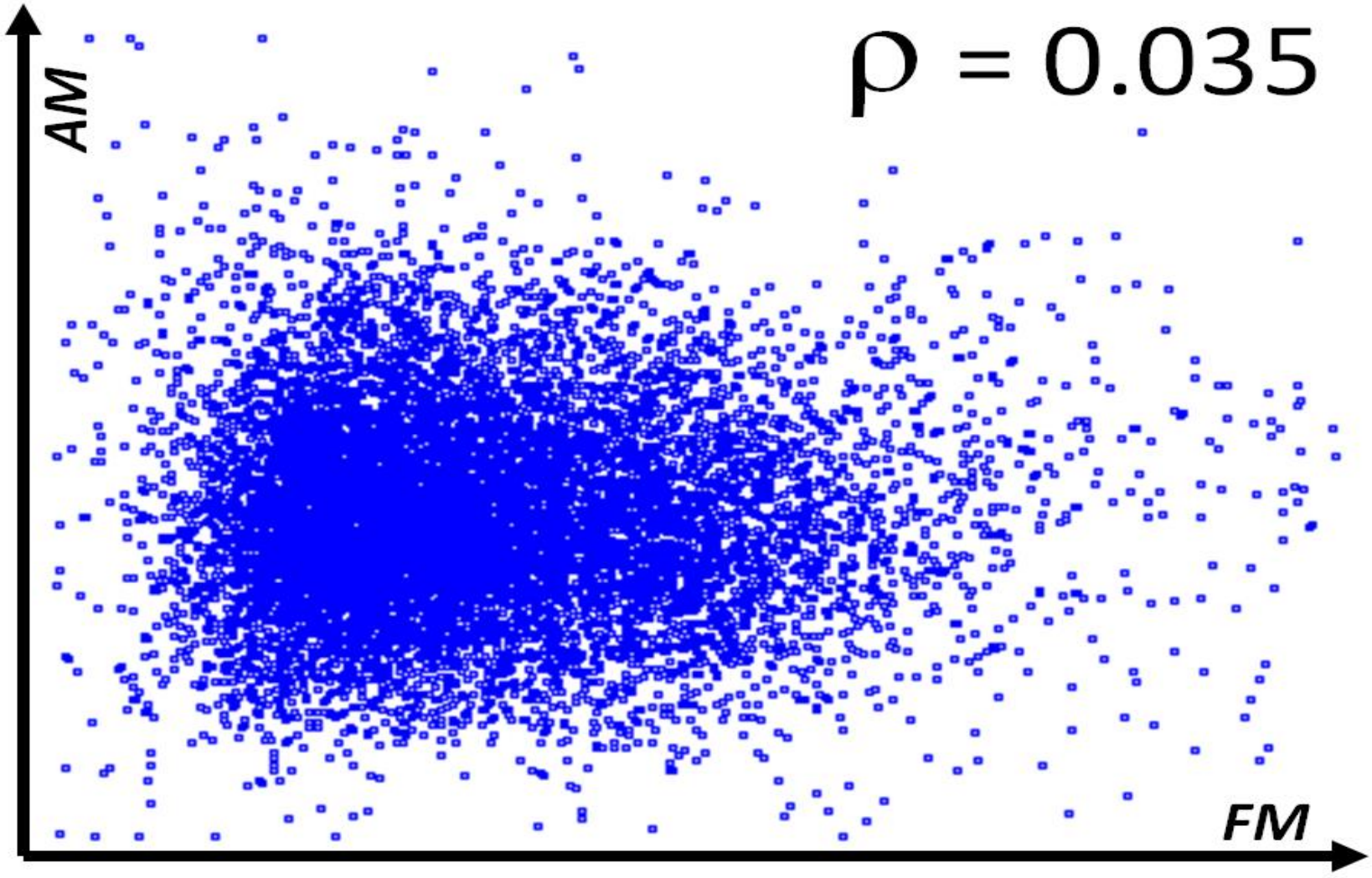
Metric	α	Adequacy	Fluency
BLEU	-	0.4107	0.4432
Meteor	-	0.3505	0.3626
NIST	-	0.3226	0.3444
TER-Plus	-	0.3068	0.3170
*m*AM	-	0.3435	0.3245
*x*AM	-	0.1291*	0.0330*
FM	-	0.3408	0.4267
*m*AM-FM $_{HM}$	0.10	0.4473	0.4977
*m*AM-FM $_{WM}$	0.60	**0.4574**	0.5036
*m*AM-FM $_{L2}$	0.86	0.4523	**0.5040**
*x*AM-FM $_{HM}$	0.30	0.4091	0.4503
*x*AM-FM $_{WM}$	0.60	0.4167	0.4442
*x*AM-FM $_{L2}$	0.80	0.4084	0.4493

All coefficients (except those marked with '*') are significant with $p<0.01$

Human Adequacy-Fluency Scores



Machine Adequacy-Fluency Scores



dialogue

DSTC6 Dataset

Computer Speech & Language
Volume 55, May 2019, Pages 1-25

Overview of the sixth dialog system technology challenge: DSTC6 ☆

Chiori Hori [a,*], Julien Perez [b], Ryuichiro Higashinaka [c], Takaaki Hori [a], Y-Lan Boureau [d], Michimasa Inaba [e], Yuiko Tsunomori [f], Tetsuro Takahashi [g], Koichiro Yoshino [h], Seokhwan Kim [i]

Show more

https://doi.org/10.1016/j.csl.2018.09.004

Highlights

- DSTC6: Dialog Challenge to improve performance of end-to-end dialog systems using Neural Network models and dialog breakdown detection.
- Track 1, End-to-End Goal Oriented Dialog Learning: selection of the best system response. Hybrid Code Network and Memory Network were the best models.
- Track 2, End-to-End Conversation Modeling: system response generation. 78.5% of the automatically generated sentences were rated as acceptable responses by humans.
- Track 3, Dialogue Breakdown Detection. The submitted systems performed as well as humans in detecting dialog breakdown, for both English and Japanese data-sets.

A Sample Dialogue

User: Hello!

System: How can I help you?

Dialogue Context

User: I'm planning to travel to Kyoto.

User Input

System: I recommend you to visit Kinkajuji.

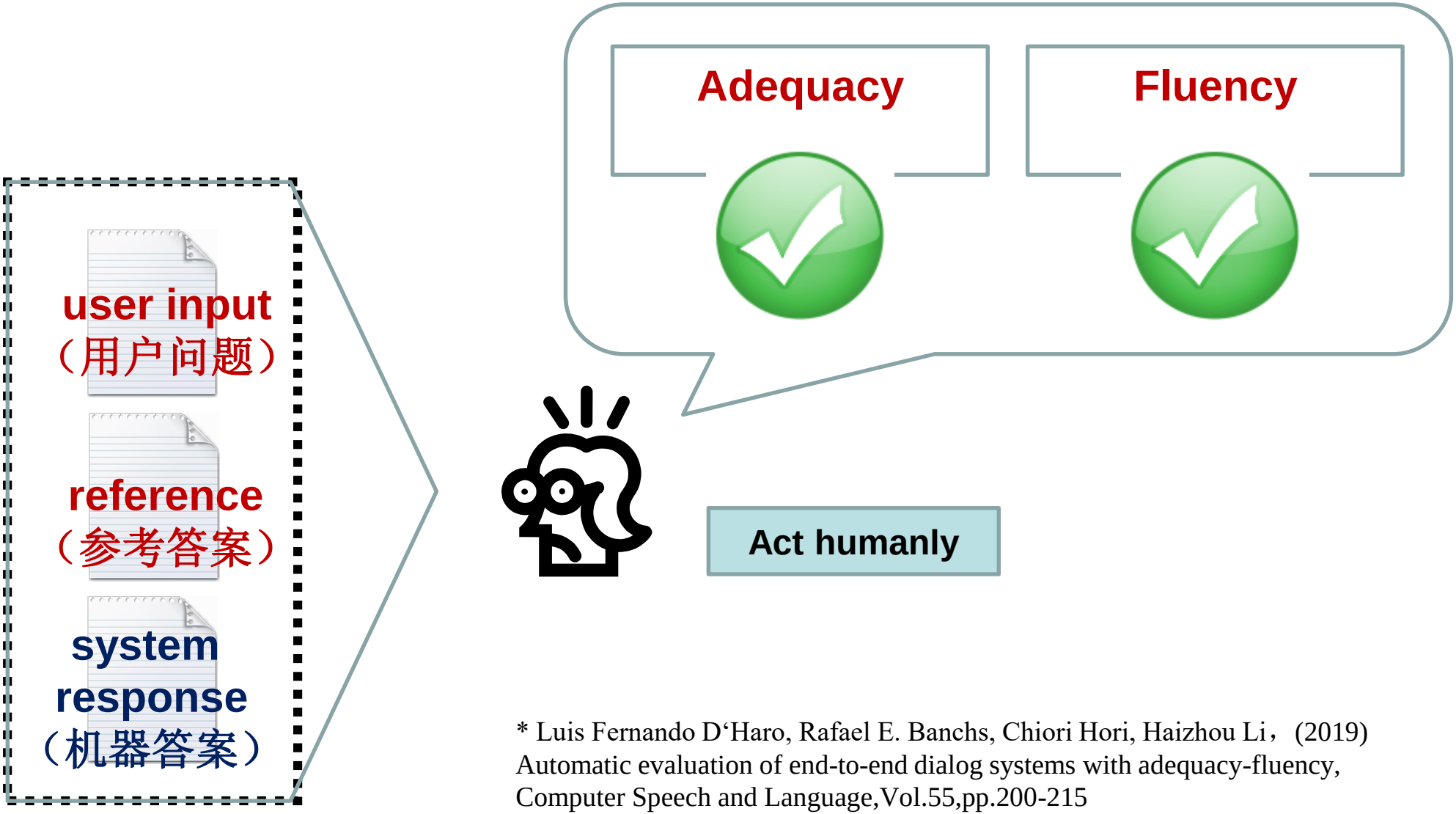
System Response

Reference 1: You may visit some temples.

Reference

Reference 2: Do you need local transport?

Reference



* Luis Fernando D'Haro, Rafael E. Banchs, Chiori Hori, Haizhou Li，(2019) Automatic evaluation of end-to-end dialog systems with adequacy-fluency, Computer Speech and Language,Vol.55,pp.200-215

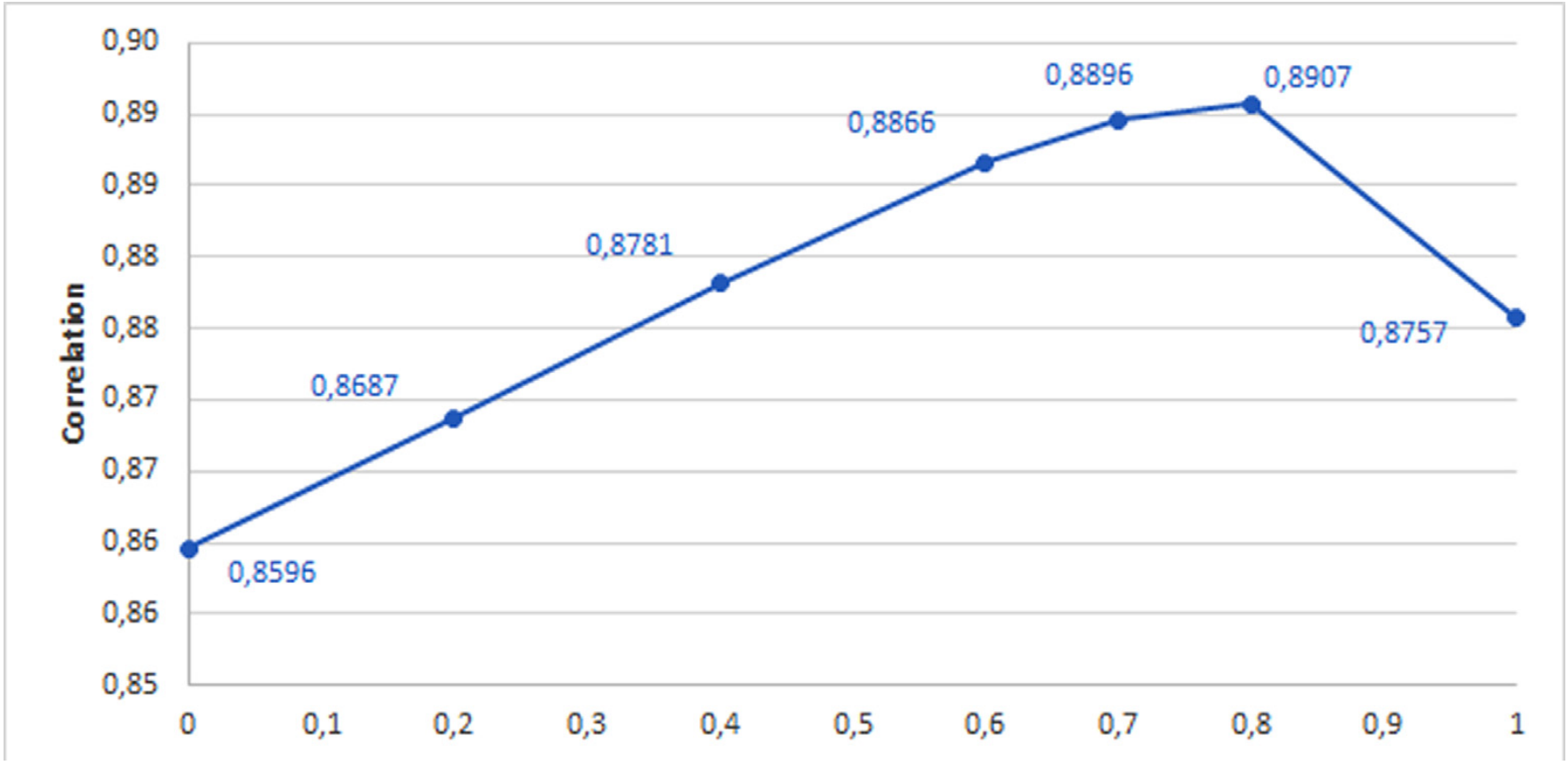
AM-FM vs Others

Table 1
Human evaluation scores, system correlations and *p*-values for different objective evaluation metrics and ours (using the best parameters) for the 20 evaluated systems at DSTC-6.

System	BLEU_4	METEOR	ROUGE_L	CIDEr	Skip thoughts	Embed avg.	Vector extrema	Greedy matching	AM–FM	Human	
										Mean	Std
S_1	0.1619	0.2041	0.3598	0.0825	0.6379	0.9155	0.6092	0.7543	0.7571	3.3638	1.047
S_2	0.1598	0.2020	0.3608	0.0780	0.6451	0.9113	0.6059	0.7527	0.7669	3.4415	1.024
S_3	0.1623	0.2039	0.3567	0.0828	0.6386	0.9060	0.6091	0.7524	0.7571	3.4298	1.026
S_4	0.1504	0.1826	0.3446	0.0803	0.6446	0.9093	0.5983	0.7488	0.7501	3.4453	1.024
S_5	0.2118	0.2140	0.3953	0.1060	0.7072	0.9281	0.6388	0.7724	0.7651	3.3894	1.045
S_6	0.1851	0.2040	0.3748	0.0965	0.6703	0.9136	0.6167	0.7571	0.7648	3.4778	1.026
S_7	0.1532	0.1833	0.3469	0.0800	0.6458	0.9099	0.5991	0.7499	0.7520	3.4382	1.025
S_8	0.2205	0.2210	0.4102	0.1279	0.6637	0.9277	0.6463	0.7773	0.7701	3.4332	1.014
S_9	0.1602	0.2016	0.3606	0.0782	0.6474	0.9103	0.6050	0.7512	0.7623	3.4504	1.022
S_10	0.1779	0.2085	0.3829	0.0978	0.6257	0.9215	0.6120	0.7647	0.7737	3.5239	1.027
S_11	0.1741	0.2024	0.3703	0.0994	0.6348	0.9021	0.6026	0.7515	0.7606	3.5082	1.011
S_12	0.1342	0.1762	0.3366	0.0947	0.6123	0.8831	0.5931	0.7315	0.7527	3.5107	1.004
S_13	0.1092	0.1731	0.3201	0.0702	0.6129	0.9014	0.5897	0.7344	0.7507	3.3919	1.033
S_14	0.1716	0.2071	0.3671	0.0898	0.6531	0.9127	0.6092	0.7548	0.7579	3.4431	1.024
S_15	0.1480	0.1813	0.3388	0.1025	0.6125	0.9104	0.5935	0.7394	0.7616	3.5209	0.997
S_16	0.0991	0.1687	0.3146	0.0708	0.5944	0.9010	0.5685	0.7204	0.7486	3.3054	1.064
S_17	0.1448	0.1839	0.3375	0.0940	0.6017	0.9103	0.5921	0.7397	0.7639	3.5396	0.998
S_18	0.1261	0.1754	0.3310	0.0945	0.6144	0.8996	0.5820	0.7287	0.7455	3.4546	1.011
S_19	0.1575	0.1918	0.3658	0.1112	0.6453	0.9094	0.6076	0.7490	0.7652	3.5098	0.997
S_20	0.2762	0.1656	0.3482	0.1235	0.6980	0.8054	0.5852	0.7202	0.6512	2.9906	0.935
Reference										3.7245	1.010
Pearson correlation	−0.5108	0.3628	0.1450	−0.1827	−0.4563	0.7768	0.2345	0.4028	**0.8907**	1.0000	
p-value	0.0214	0.1159	0.5420	0.4408	0.0432	0.0001	0.3196	0.0782	**1.41E−7**	0.0000	

AM-FM overall score

AM-FM= $\alpha \times AM + (1 - \alpha) \times FM$

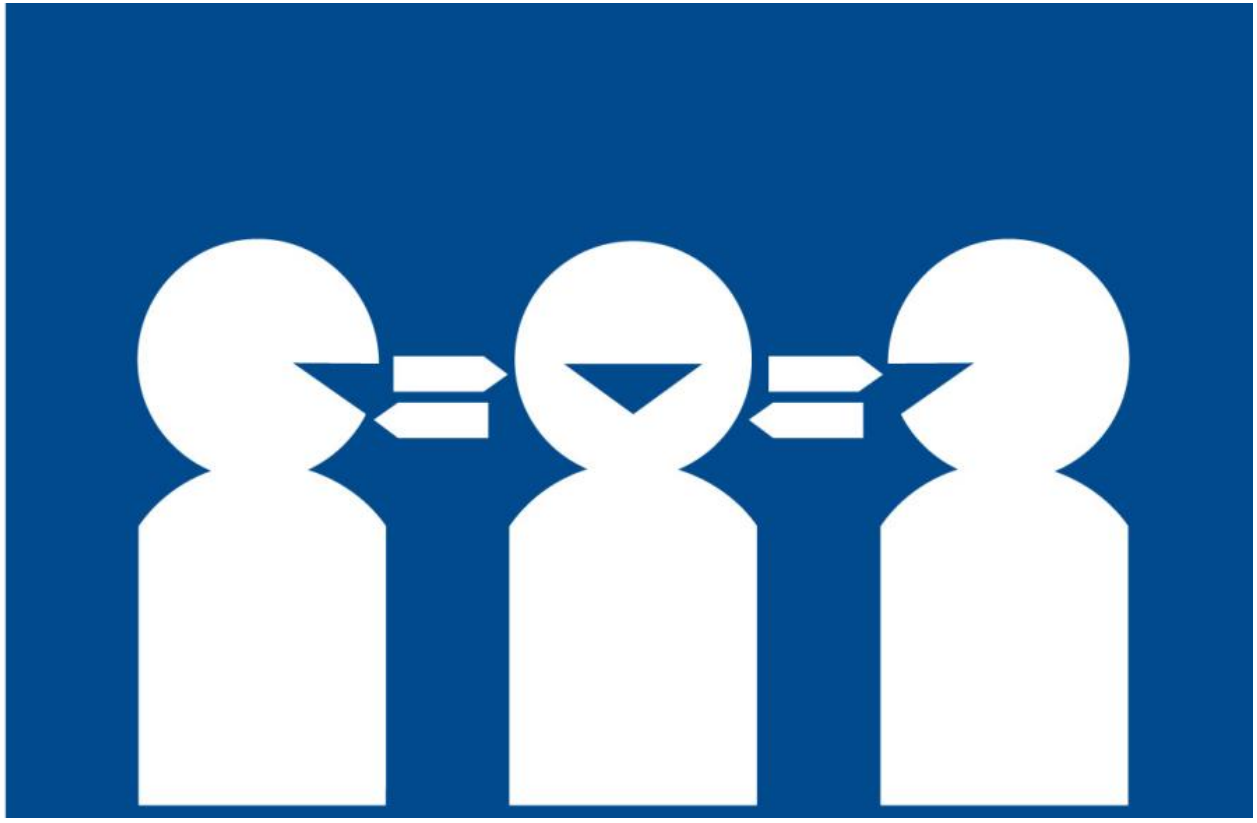


Qualitative Examples

Table 3
Qualitative examples where the metric correctly estimates the score in agreement with the average human scores, and when the metric fails to match the human scores. The provided text is the same that appears in the references, where we can see some misspellings and grammatical errors.

Prediction	Example
	**Dialog Context:** wins ! @infinitiusa ⟨URL⟩ **Ground Truth:** yay we 're happy you 're happy **Hypothesis:** we 're happy to have you as part of the infiniti family, ⟨USER⟩ ! **Human Score:** $\mu = 4.4 \pm 0.6$ **AM−FM Score:** 0.846 (AM: 0.817, FM: 0.962)
	**Dialog Context:** is there a problem with your online booking system this evening ? i 'm trying to book but nothing will load ? **Ground Truth:** i 'm sorry but we are currently facing technical issues with our site we will resolve them as soon as possible **Hypothesis:** thank you . **Human Score:** $\mu = 1.8 \pm 1.0$ **AM−FM Score:** 0.104 (AM: 0.081, FM: 0.194)
	**Dialog Context:** **U:** no one has come to collect car yet . hire car from motability got . not good service **S:** have you contacted jeep assist to collect ? who were you in agreement with ? **U:** jeep friern barnet spoke yesterday to mo **Ground Truth:** it is best to contact the dealership directly **Hypothesis:** if you'd like to speak with someone **Human Score:** $\mu = 1.4 \pm 0.5$ **AM−FM Score:** 0.912 (AM: 0.917, FM: 0.891)
	**Dialog Context:** if you haven't met @dandtwitr yet, you are missing out ! #fes2015 ⟨URL⟩ **Ground Truth:** we agree with you on that get more infor on our website **Hypothesis:** amen to that ! ! **Human Score:** $\mu = 4.3 \pm 0.6$ **AM−FM Score:** 0.215 (AM: 0.154, FM: 0.460)
	**Dialog Context:** do you delivery to serbia **Ground Truth:** i am sorry but we currently do not deliver to serbia **Hypothesis:** hi, yes we do deliver to ireland . **Human Score:** $\mu = 1.6 \pm 1.3$ **AM−FM Score:** 0.816 (AM: 0.792, FM: 0.912)

Conclusion



Thank you!

WMT-2007 Translation Task Details

Task	Domain	Source	Target	Systems	Sentences
T1	News	CZ	EN	3	727
T2	News	EN	CZ	2	806
T3	EPPS	EN	FR	7	577
T4	News	EN	FR	8	561
T5	EPPS	EN	DE	6	924
T6	News	EN	DE	6	892
T7	EPPS	EN	ES	6	703
T8	News	EN	ES	7	832
T9	EPPS	FR	EN	7	624
T10	News	FR	EN	7	740
T11	EPPS	DE	EN	7	949
T12	News	DE	EN	5	939
T13	EPPS	ES	EN	8	812
T14	News	ES	EN	7	668