Adequacy-Fluency Metrics (AM-FM) for Machine Translation (MT) Evaluation

Haizhou Li

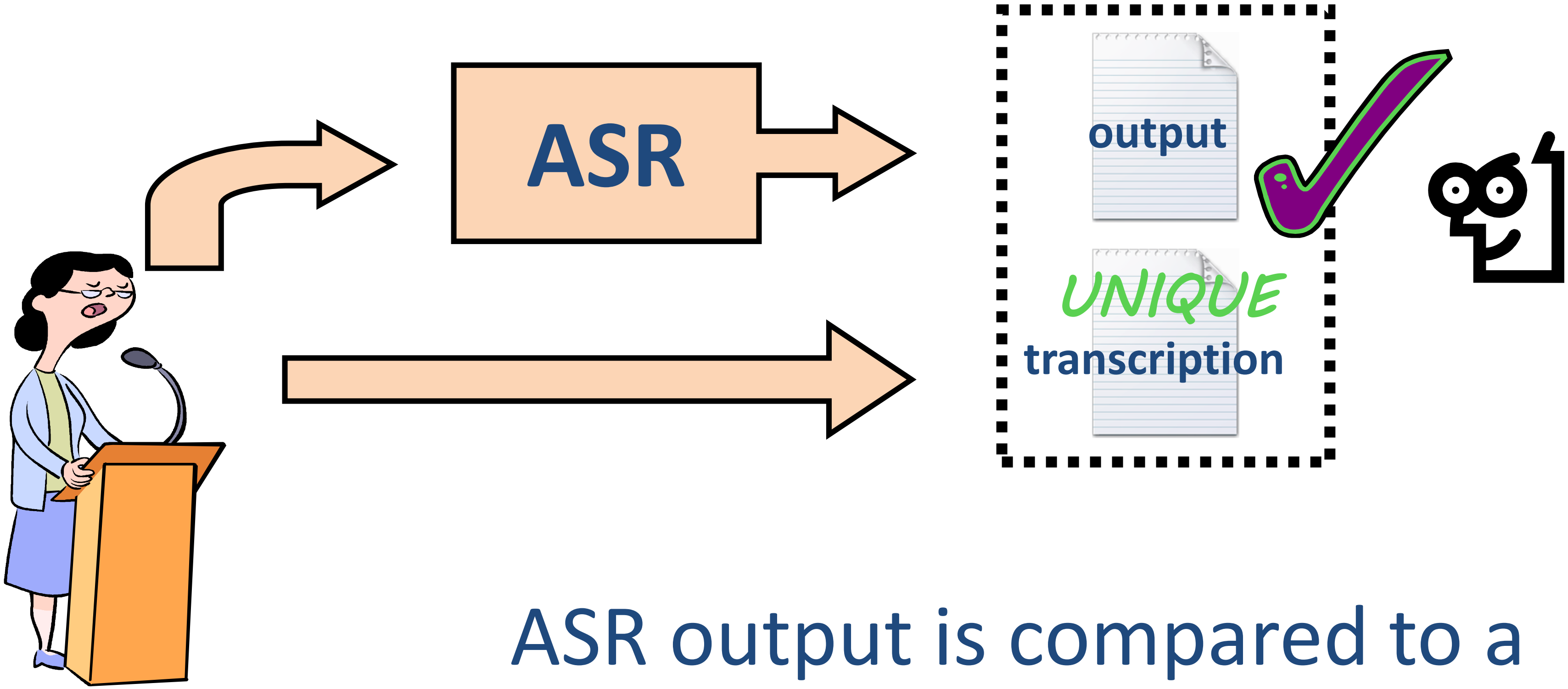
Invited Talk: WAT 2015, 16 Oct 2015, Kyoto, Japan

Banchs R. E., D'Haro L.F., Li H. (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

Agenda

- The evaluation of ASR and MT
- How do machines evaluate translations today?
- How do humans evaluate translations?
- The Adequacy-Fluency Metrics (AM-FM)
- The mathematical formulation
- The experiments

Automatic Evaluation of Automatic Speech Recognition



ASR output is compared to a reference transcription.

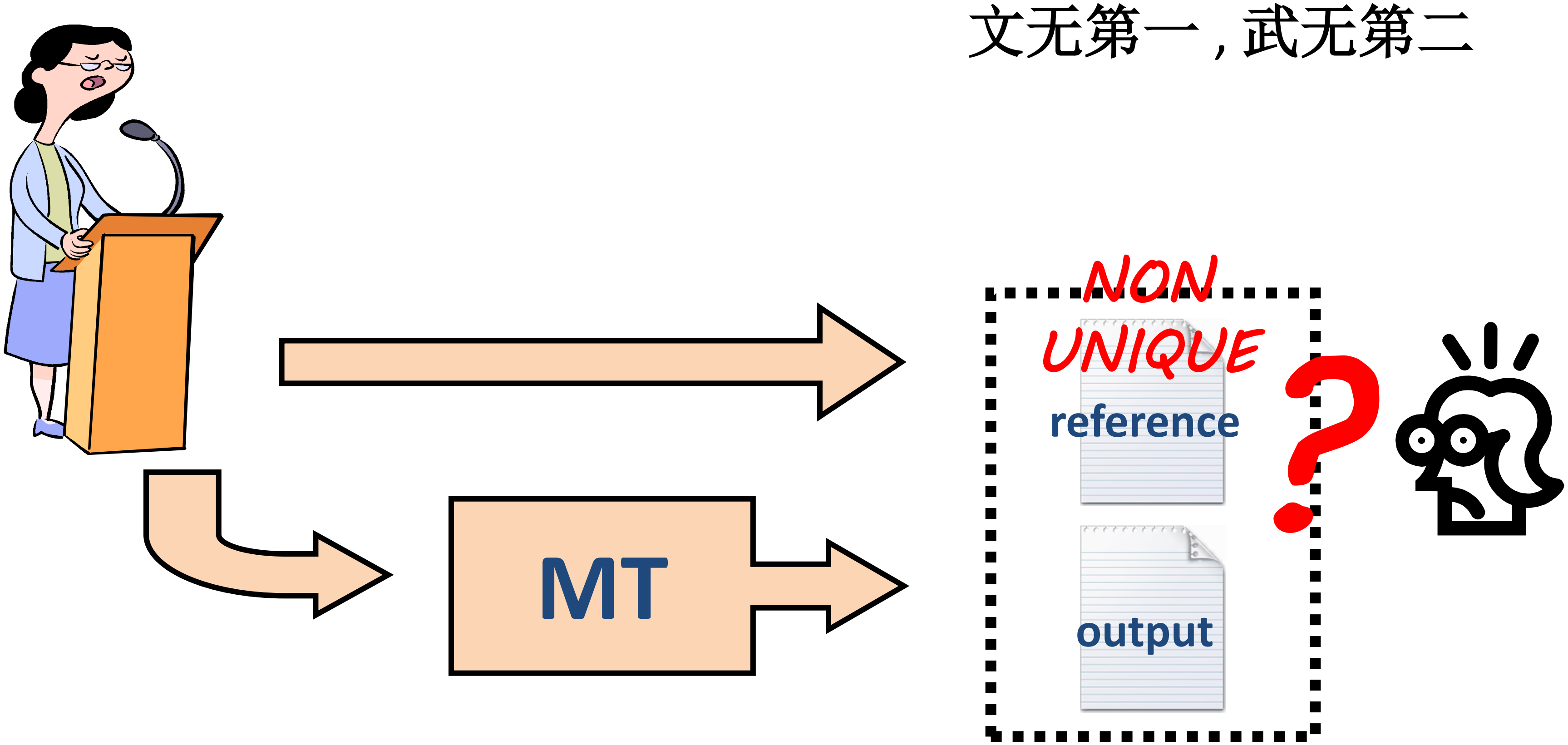
The reference transcription is unique!

Automatic Evaluation of Machine Translation

MT output is compared to reference translations.

… but references are not unique!

文无第一，武无第二



Agenda

- The evaluation of ASR and MT
- How do machines evaluate translations today?
- How do humans evaluate translations?
- The Adequacy-Fluency Metrics (AM-FM)
- The mathematical formulation
- The experiments

Traditional Evaluation Approach

Compare the output with a set of references

WER[1], PER[1] → **Compare words**

BLEU[2], NIST[3] → **Compare *n*-grams**

1.- C. Tillmann *et al.*, "Accelerated DP Based Search for Statistical Translation", in *Proc. of the 5th European Conf. on Speech Commun. and Tech.*, Rhodos, Greece, Sept 1997, pp. 2667–2670.

2.- K. Papineni *et al.*, "BLEU: a method for automatic evaluation of machine translation", in *Proc. of the 40th Annu. Meeting of the Assoc. for Computational Linguistics*, Philadelphia, PA, USA, Jul 2002, pp. 311-318

3.- G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics", in *Proc. of the Human Lang. Tech. Conf.*, San Diego, CA, USA, Mar 2002

Traditional Approach: Good Scores

Translation Output → **This is a toilet.**

Reference Translation → **This is a toilet.**

word matches = 4/4

n-gram matches = 5/5

Good Score → **Good Translation**

Traditional Approach: Bad Scores

Translation Output → **It's the Water Closet.**

Reference Translation → **This is a toilet.**

word matches = 0/4

n-gram matches = 0/5

Bad Score → **?**

Traditional Approach: Better Scores?

Translation Output → **This isn't a toilet.**

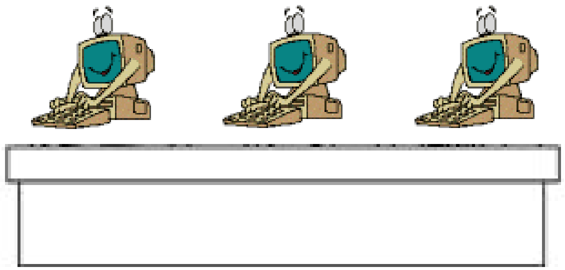
Reference Translation → **This is a toilet.**

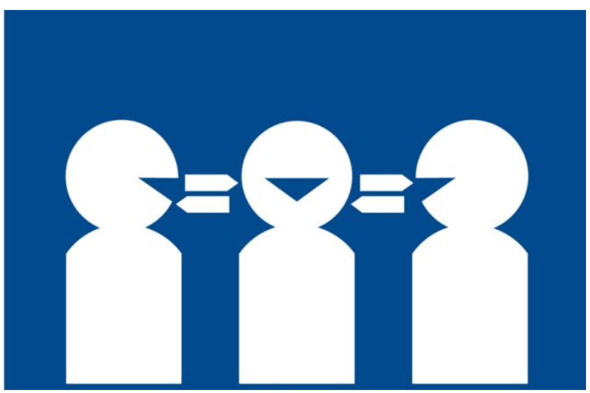
word matches = 3/4

n-gram matches = 3/5

Better Score → **?**

How Machines Evaluate Translations?





- **Only look at outputs and references**
- **Without knowledge support**

A Semantic Framework is Needed

Automatic MT evaluation must move beyond words and *n*-grams! Some recent proposals:

METEOR[1] → **Compare stems and synonyms**

TER[2] → **Compute edit distances**

MEANT[3] → **Compare semantic roles**

1.- A. Lavie and M.J. Denkowski, “The Meteor metric for automatic evaluation of machine translation”, *Machine Translation*, vol. 23, pp. 105-115, May 2009

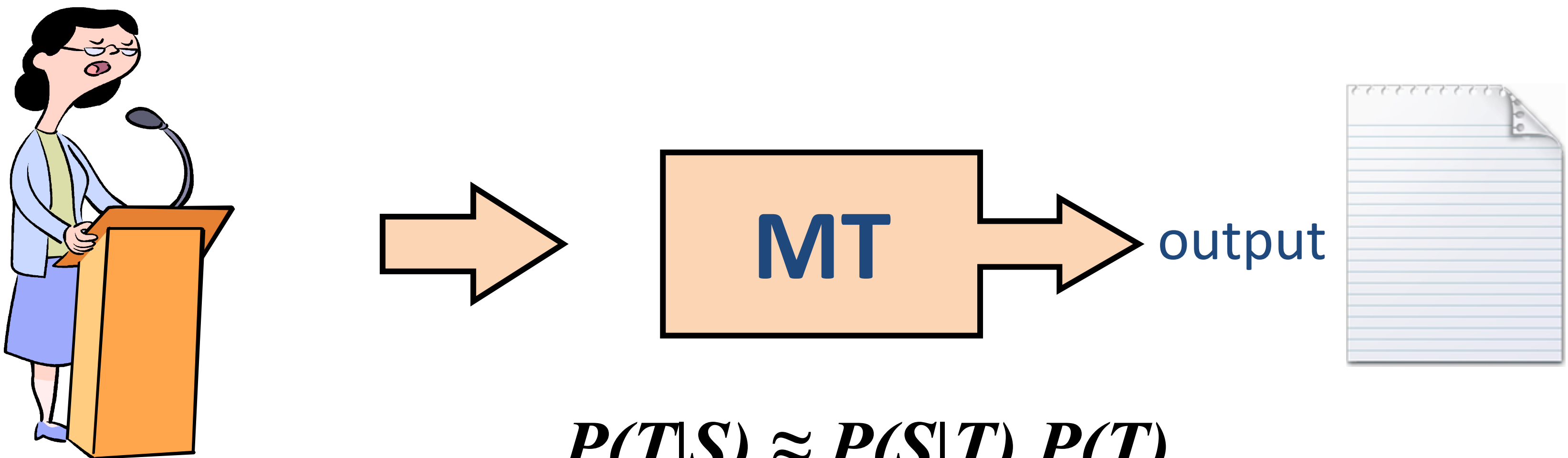
2.- M. Snover *et al.*, “Study of Translation Edit Rate with Targeted Human Annotation”, in *Proc. of the 7th Biennial Conf. of the Assoc. for Mach. Translation in the Amer.*, Cambridge, MA, USA, Aug 2006

3.- C.K. Lo and D. Wu, “MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on seman-tic roles”, in *Proc. of the 49th Annu. Meeting of the Assoc. for Computational Linguistics,* Portland, OR, USA, Jun 2011, pp. 220-229

Agenda

- The evaluation of ASR and MT
- How do machines evaluation translations today?
- How do humans evaluation translations?
- The Adequacy-Fluency Metrics (AM-FM)
- The mathematical formulation
- The experiments

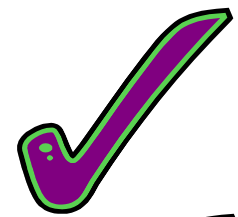
How Humans Evaluate Translations?* (I)



$$P(T|S) \approx P(S|T)\,P(T)$$

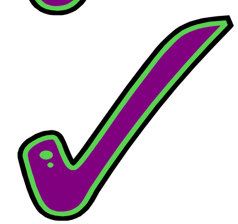
ADEQUACY

How much of the source information is preserved?



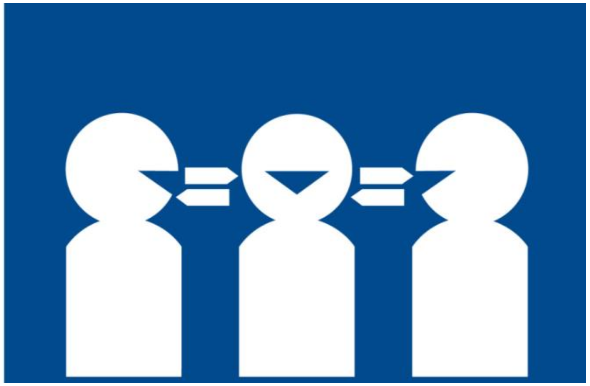
FLUENCY

How good is the generated target language quality?



* 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

How Humans Evaluate Translations ? (II)



- **Look at both outputs and inputs**
- **Language and cultural knowledge**

*Adequacy Evaluation Scale**

How much of the source information is preserved in the translation?

(Look at both inputs and outputs!)

Score	Definition
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

* 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

*Fluency Evaluation Scale**

How good is translation regarding the target language quality?

(Only look at the outputs!)

Score	Definition
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

Agenda

- The evaluation of ASR and MT
- How do machines evaluate translations today?
- How do humans evaluate translations?
- The Adequacy-Fluency Metrics (AM-FM)
- The mathematical formulation
- The experiments

*The Proposed Evaluation Framework**

- Approximate adequacy and fluency by means of independent models:
  - Use a "semantic approach" for adequacy
  - Use a "syntactic approach" for fluency
- Combine both evaluation metrics into a single evaluation score

* Banchs R.E., D'Haro L.F., Li H. (2015) "Adequacy - Fluency Metrics: Evaluating MT in the Continuous Space Model Framework", IEEE/ACM Transactions on Audio, Speech and Language Processing, Special issue on continuous space and related methods in NLP, Vol.23, No.3, pp.472-482

State of the Art in MT Evaluation*

Assessment Level	**Need for References**	**Cross-Language Approach**	**Humans in the Loop**
Words	WER, PER	-	-
Word *n*-grams	BLEU, NIST	-	-
Stems & Synonyms	METEOR	-	-
Edit Distances	TER	-	HTER
Semantic Roles	MEANT	XMEANT	HMEANT
Continuous Space	*m*AM-FM	*x*AM-FM	-

* Banchs R.E., D'Haro L.F., Li H. (2015) "Adequacy - Fluency Metrics: Evaluating MT in the Continuous Space Model Framework", IEEE/ACM Transactions on Audio, Speech and Language Processing, Special issue on continuous space and related methods in NLP, Vol.23, No.3, pp.472-482

Properties of Continuous Spaces

The Distributional Hypothesis

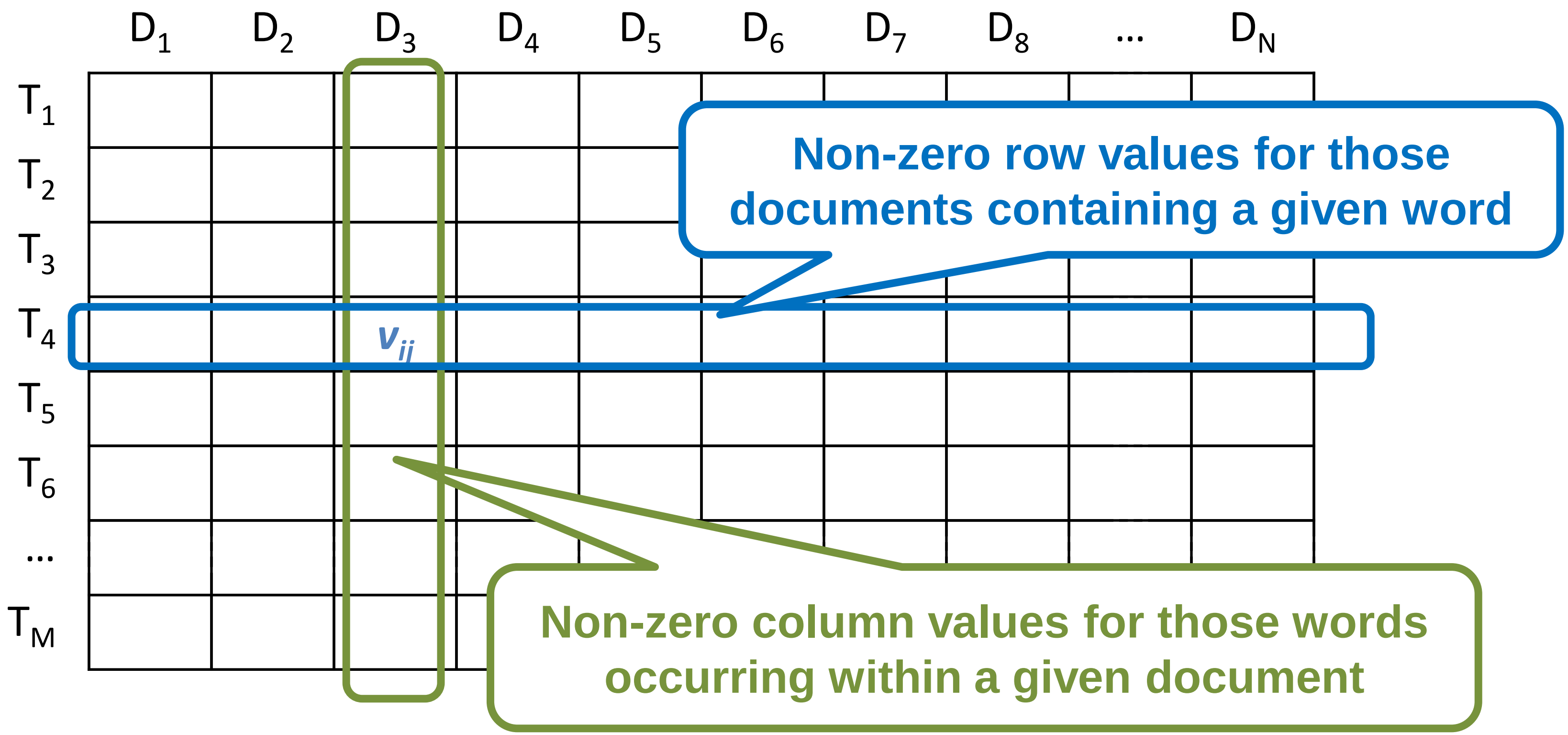
“a word is characterized for the company it keeps” (Firth 1957)

meaning is mainly determined by the context rather than from individual language units

- **Continuous spaces represent semantic similarities by means of the geometric concept of proximity**
- **Offer much “better” smoothing capabilities**
- **Not constrained to the Markovian assumption**

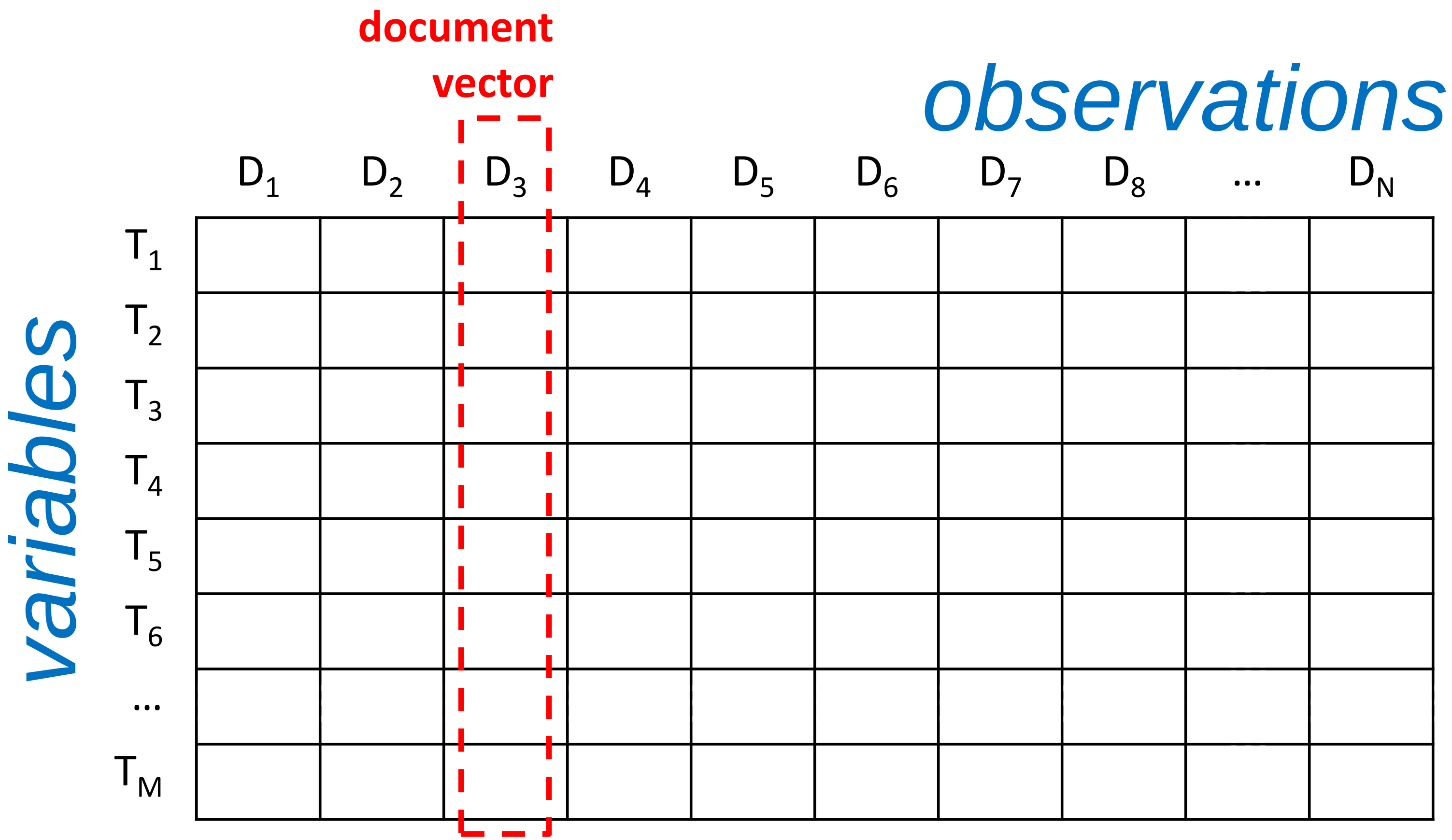
The Term-Document Matrix

- A model representing joint distributions between words and documents



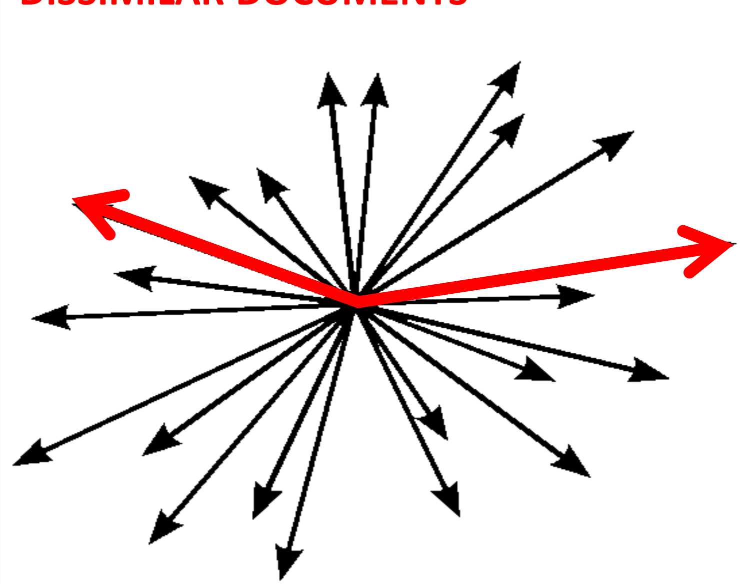
Document Vector Spaces

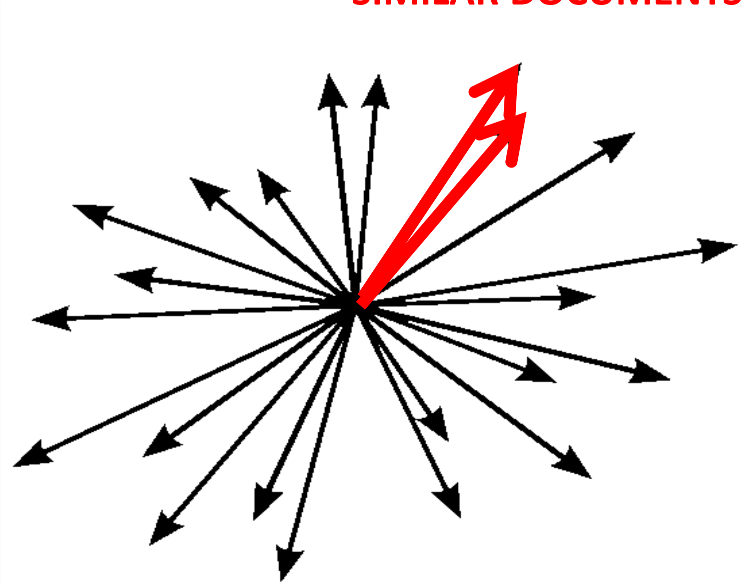
Pay attention to the columns of the term-document matrix



Semantic Association in Vector Spaces

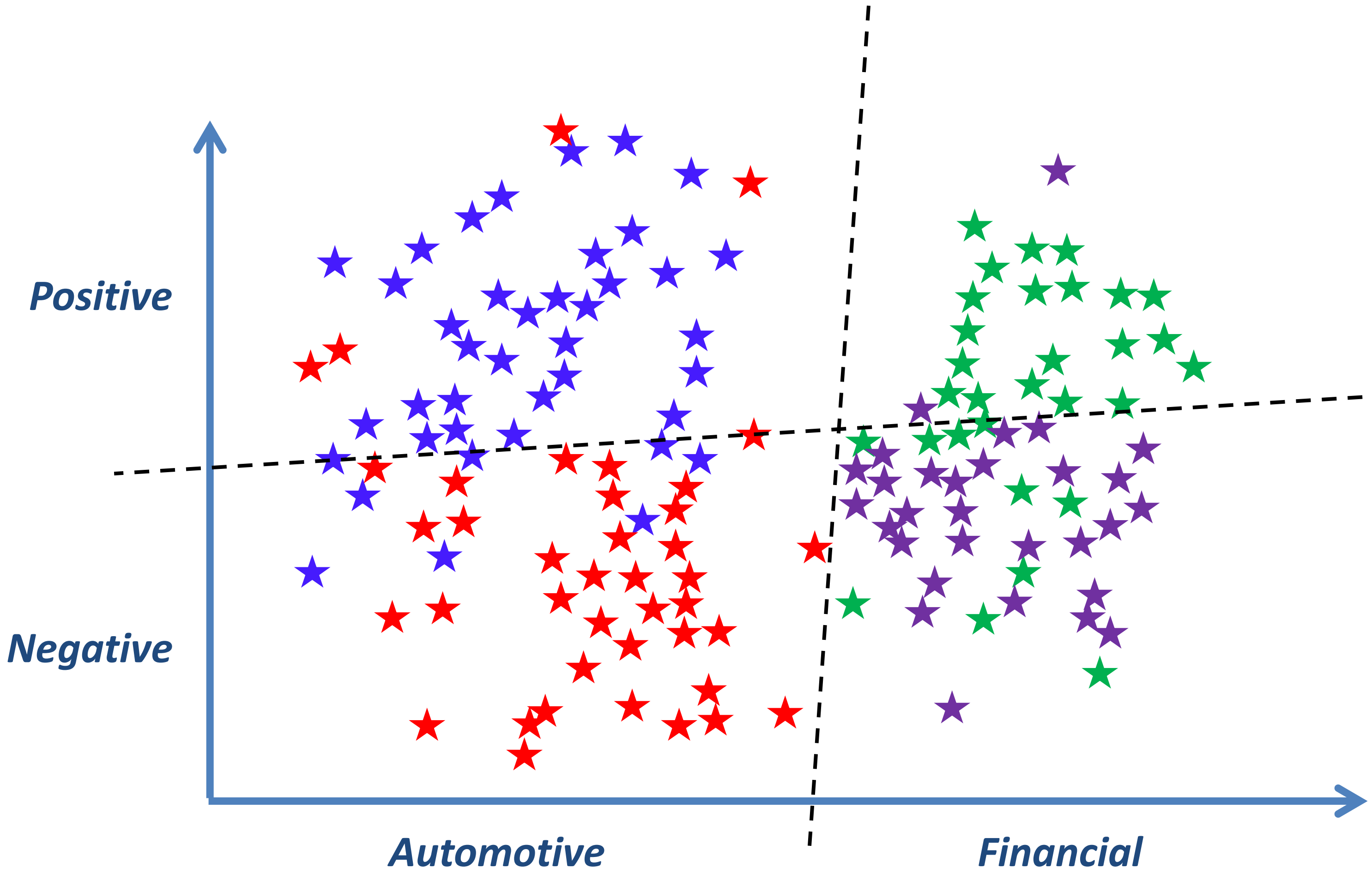
Association scores and similarity metrics can be used to assess the degree of semantic relatedness among documents





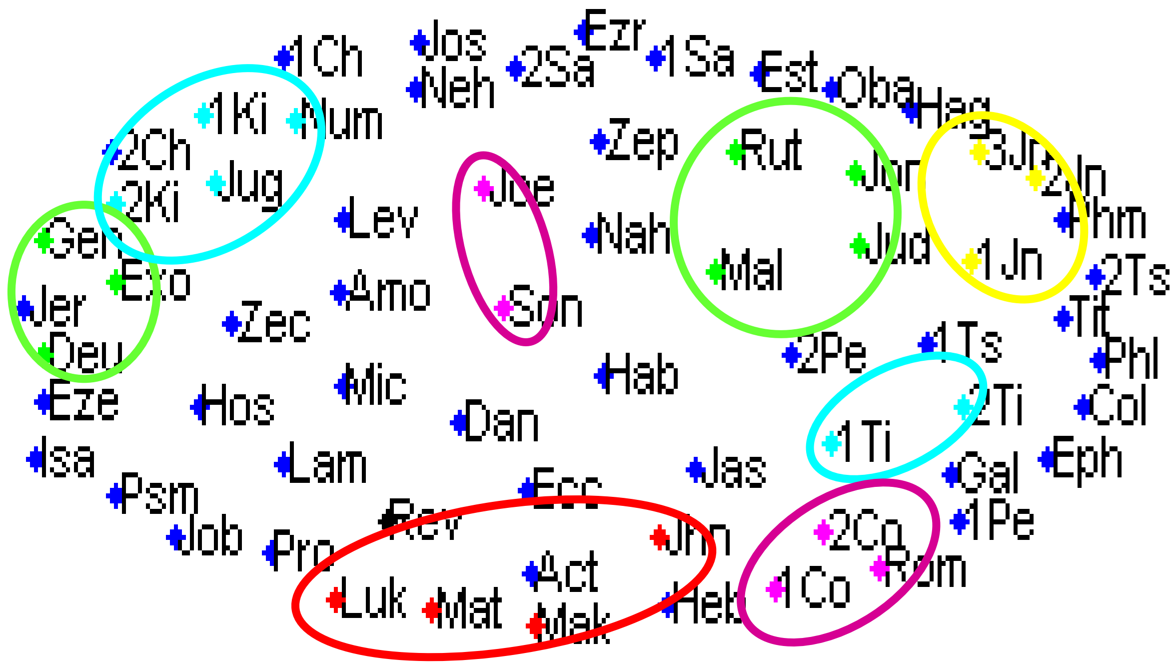
Semantic Map for Data Collection (1)

Opinionated content from rating website



Semantic Map for Data Collection (2)

66 Books from The Holy Bible: English version



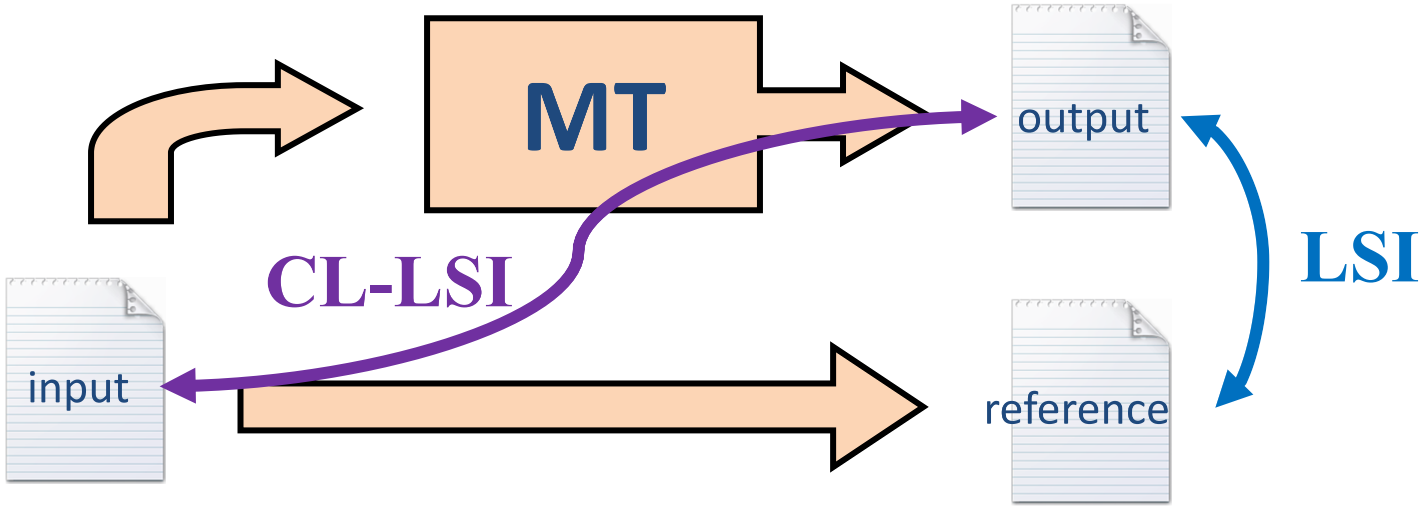
(vocabulary size: 8121 words)

Agenda

- The evaluation of ASR and MT
- How do machines evaluate translations today?
- How do humans evaluate translations?
- The Adequacy-Fluency Metrics (AM-FM)
- The mathematical formulation
- The experiments

AM: Adequacy-oriented Metric

- Compare sentences in a semantic space
  - Monolingual AM (***mAM***): compare output vs. reference
  - Cross-language AM (***xAM***): compare output vs. input



Latent Semantic Indexing (LSI)*

SVD: $\boldsymbol{M}_{M\times N} = \boldsymbol{U}_{M\times M}\,\boldsymbol{\Sigma}_{M\times N}\,\boldsymbol{V}^{T}_{N\times N}$

$$\boldsymbol{U}^{T}_{M\times M}\,\boldsymbol{M}_{M\times N} = \boldsymbol{D}_{M\times N}$$

Documents projected into word space

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

$$\boldsymbol{U}^{T}_{K\times M}\,\boldsymbol{M}_{M\times N} = \boldsymbol{D}_{K\times N}$$

Documents projected into reduced word (semantic) space

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

$$<\boldsymbol{U}^{T}_{K\times M}\,\boldsymbol{to}_{M\times 1}\,,\ \boldsymbol{U}^{T}_{K\times M}\,\boldsymbol{tr}_{M\times 1}>$$

* 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*

Multilingual term-document matrix

$$X_{(Ms+Mt)\times N} = \begin{pmatrix} M_{Ms\times N} \\ M_{Mt\times N} \end{pmatrix}$$

Term-document matrix in source language

Term-document matrix in target language

SVD: $\boldsymbol{X = U\,\Sigma\,V^T}$

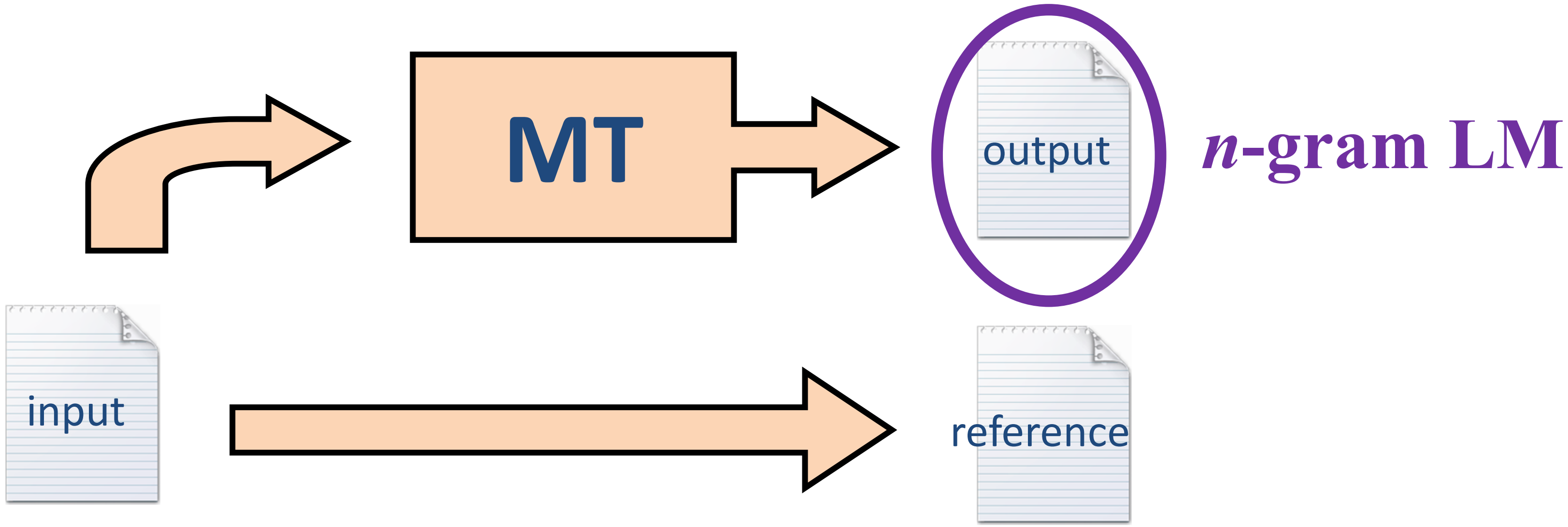
Translation output (**to**) and translation input (**ti**) compared in cross-language vector space

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

* 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 of the target language with a language model
- Uses a compensation factor to avoid effects derived from differences in sentence lengths



Compensated Language Model

n**-gram probabilities**

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

compensation factor

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

Agenda

- The evaluation of ASR and MT
- How do machines evaluate translations today?
- How do humans evaluate translations?
- The Adequacy-Fluency Metrics (AM-FM)
- The mathematical formulation
- The experiments

*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.

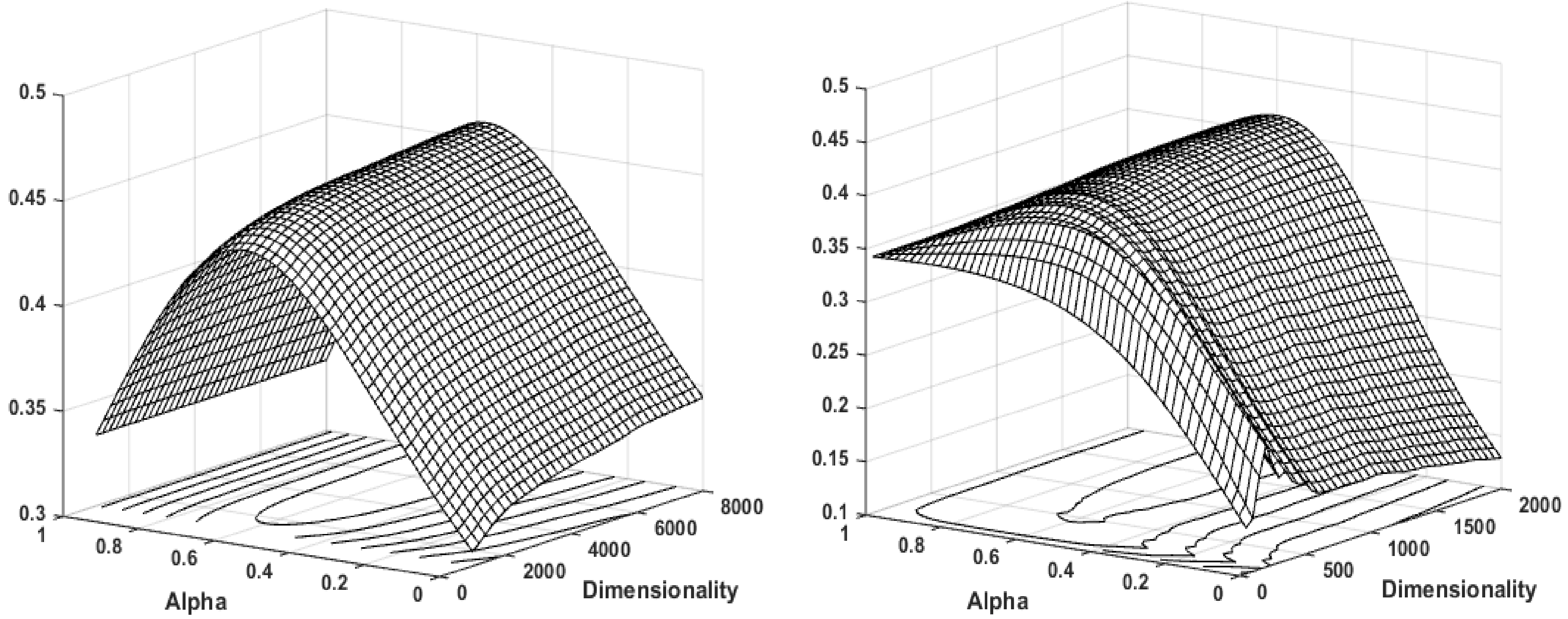
* Callison-Burch C., Fordyce C., Koehn P., Monz C. and Schroeder J. (2007), (Meta-) evaluation of machine translation, in Proceedings of Statistical Machine Translation Workshop, pp. 136-158

WMT-2007 Translation Task Details

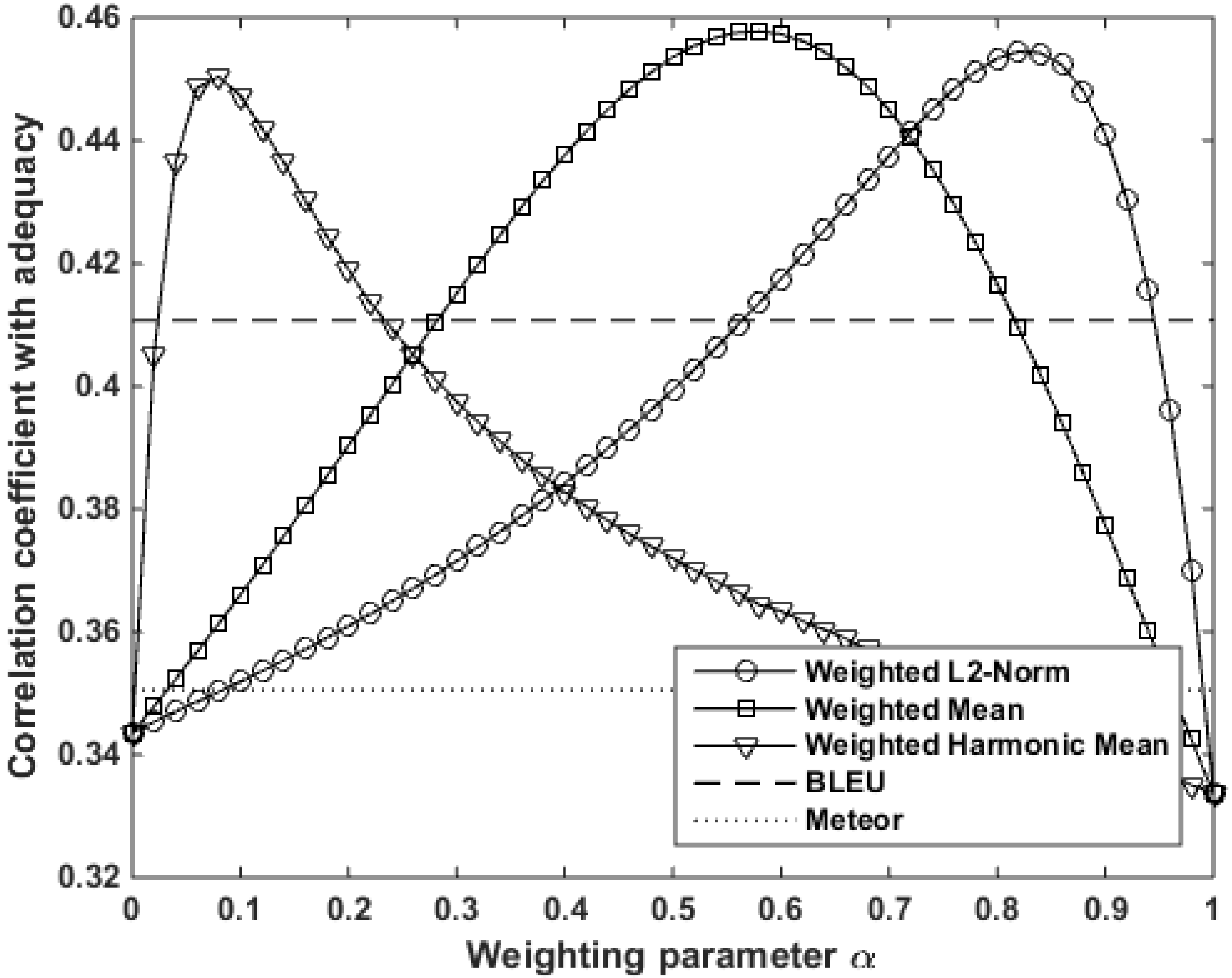
Task	Domain	Source	Target	Systems	Sentences
T1	News	CZ	EN	3	727
T2	News	EN	CZ	2	806
T3	EPPS	EN	FR	7	577
T4	News	EN	FR	8	561
T5	EPPS	EN	DE	6	924
T6	News	EN	DE	6	892
T7	EPPS	EN	ES	6	703
T8	News	EN	ES	7	832
T9	EPPS	FR	EN	7	624
T10	News	FR	EN	7	740
T11	EPPS	DE	EN	7	949
T12	News	DE	EN	5	939
T13	EPPS	ES	EN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

Metric Correlation with Human Scores

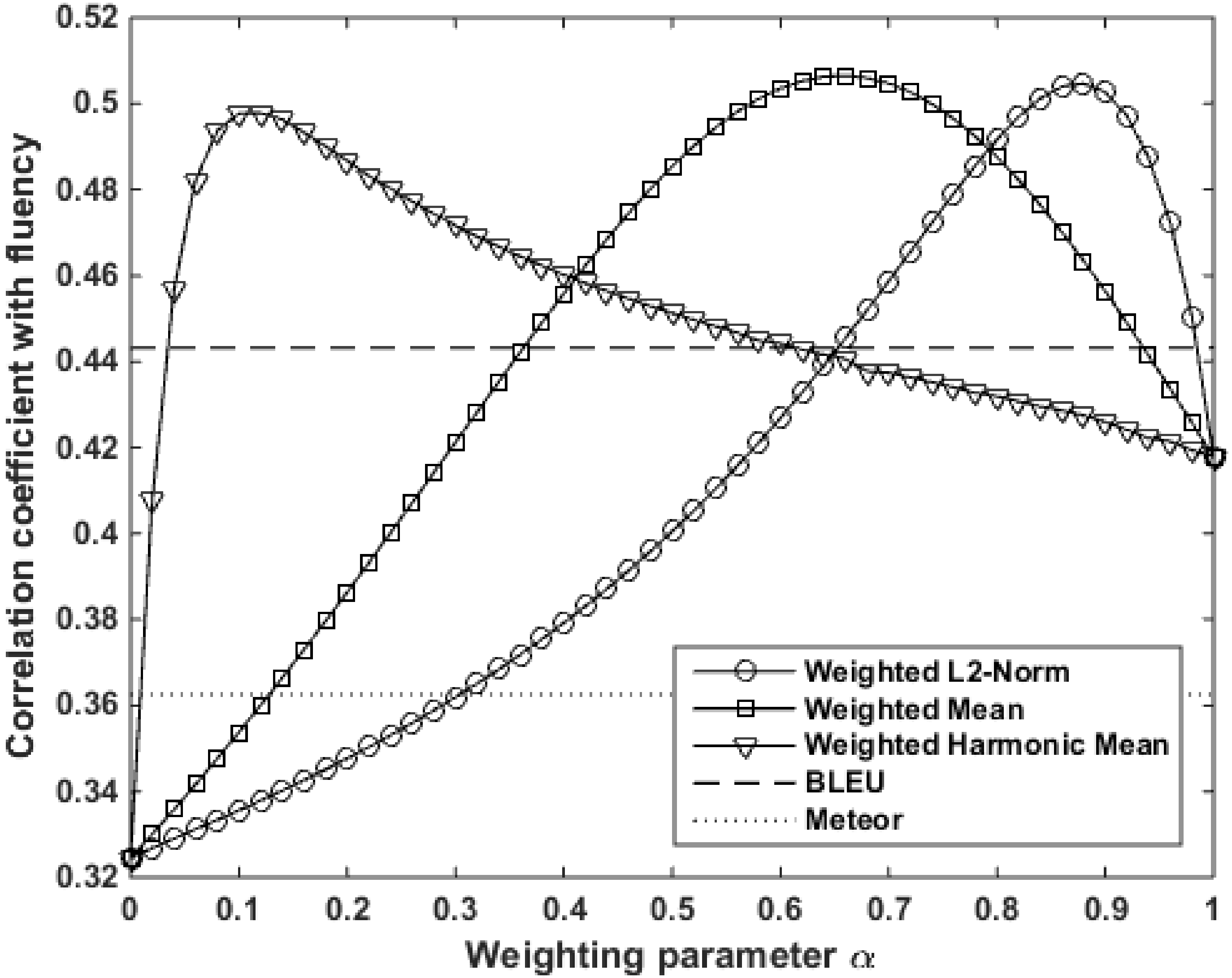
Pearson's correlation coefficients between the ***mAM-FM*** *Weighted Mean* (left) and ***xAM-FM*** *Weighted Mean* (right) components and human-generated scores for adequacy



mAM-FM and Adequacy

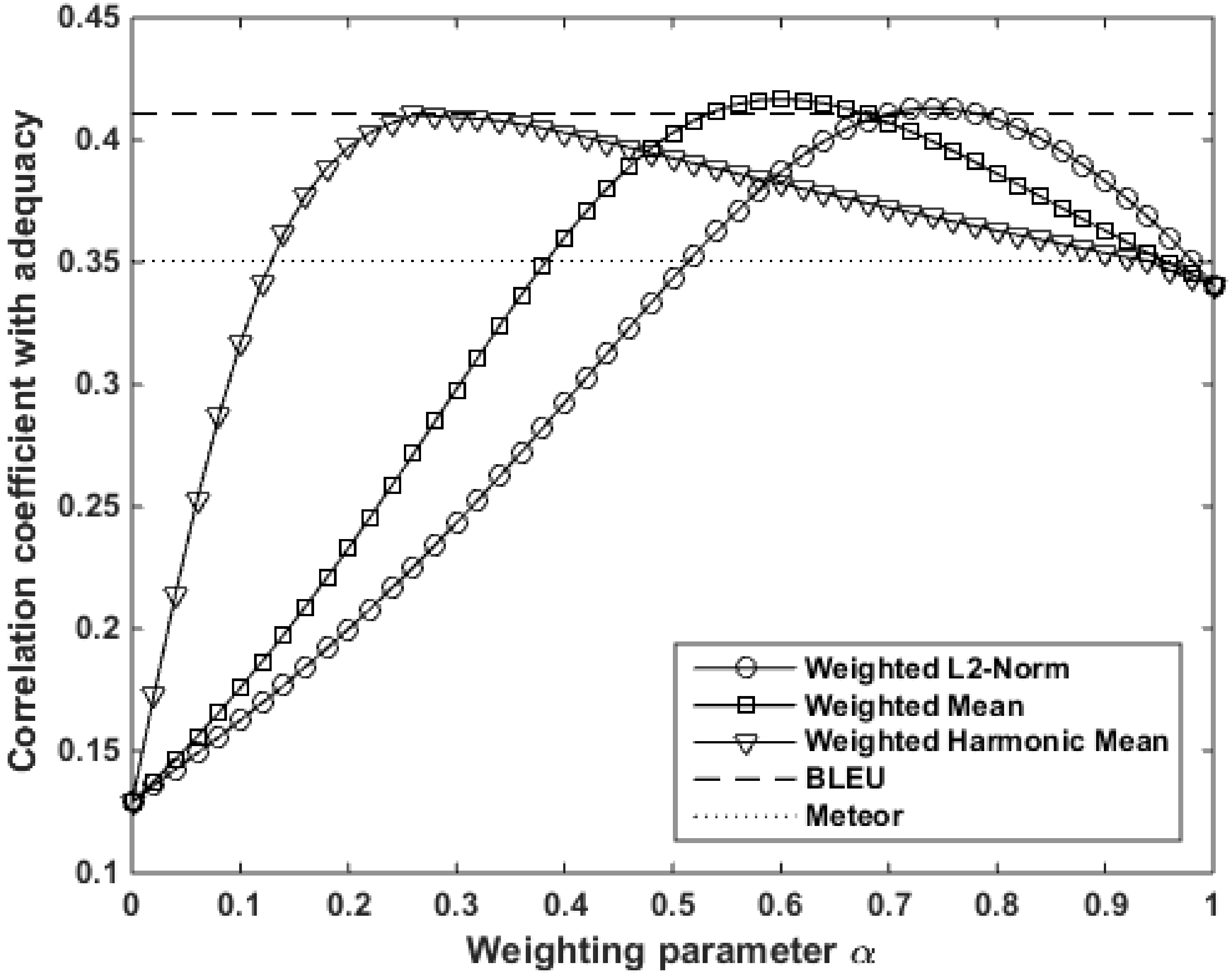


mAM-FM and Fluency



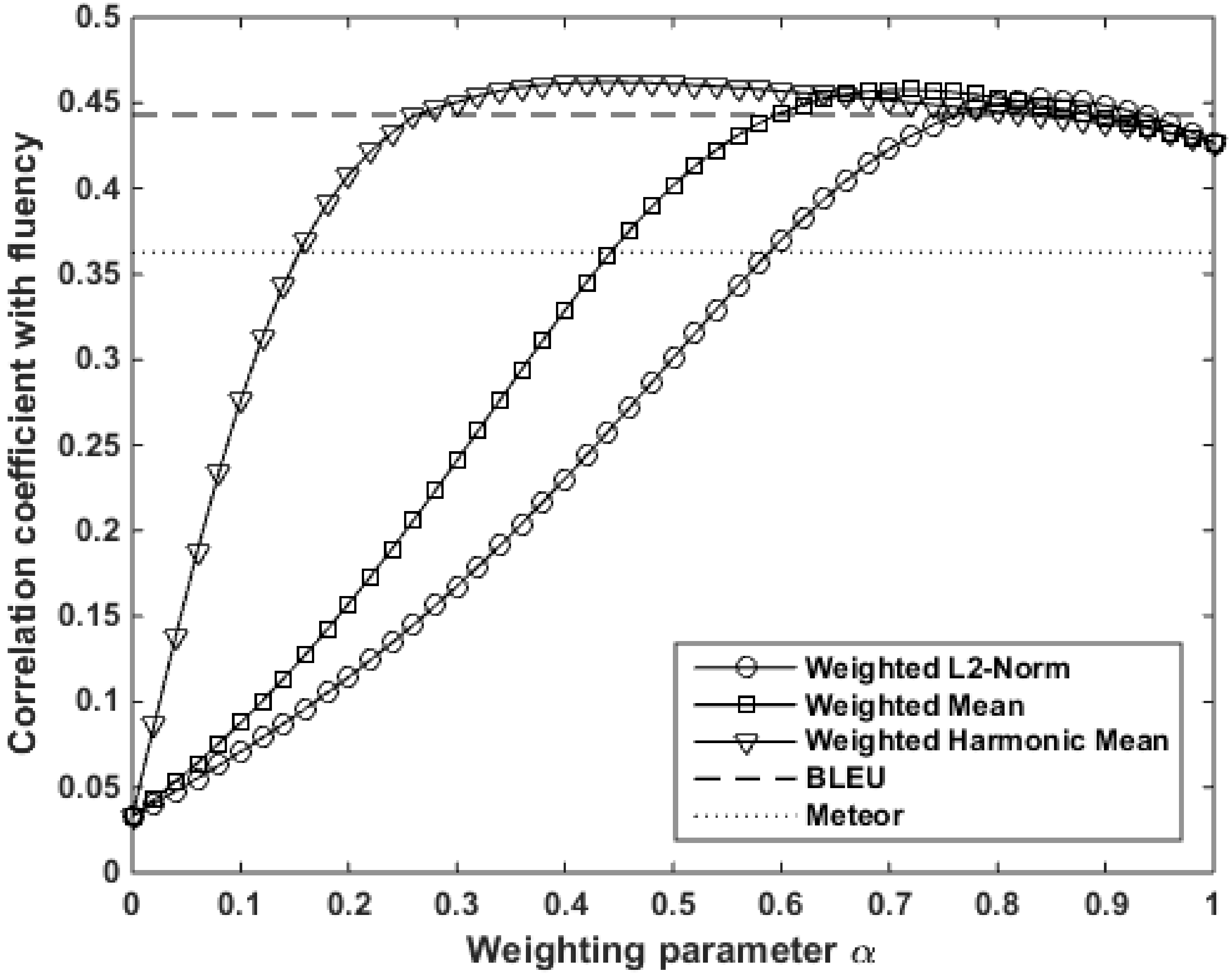
Institute for Infocomm Research (I[2]R)

xAM-FM and Adequacy



Institute for Infocomm Research (I2R)

xAM-FM and Fluency

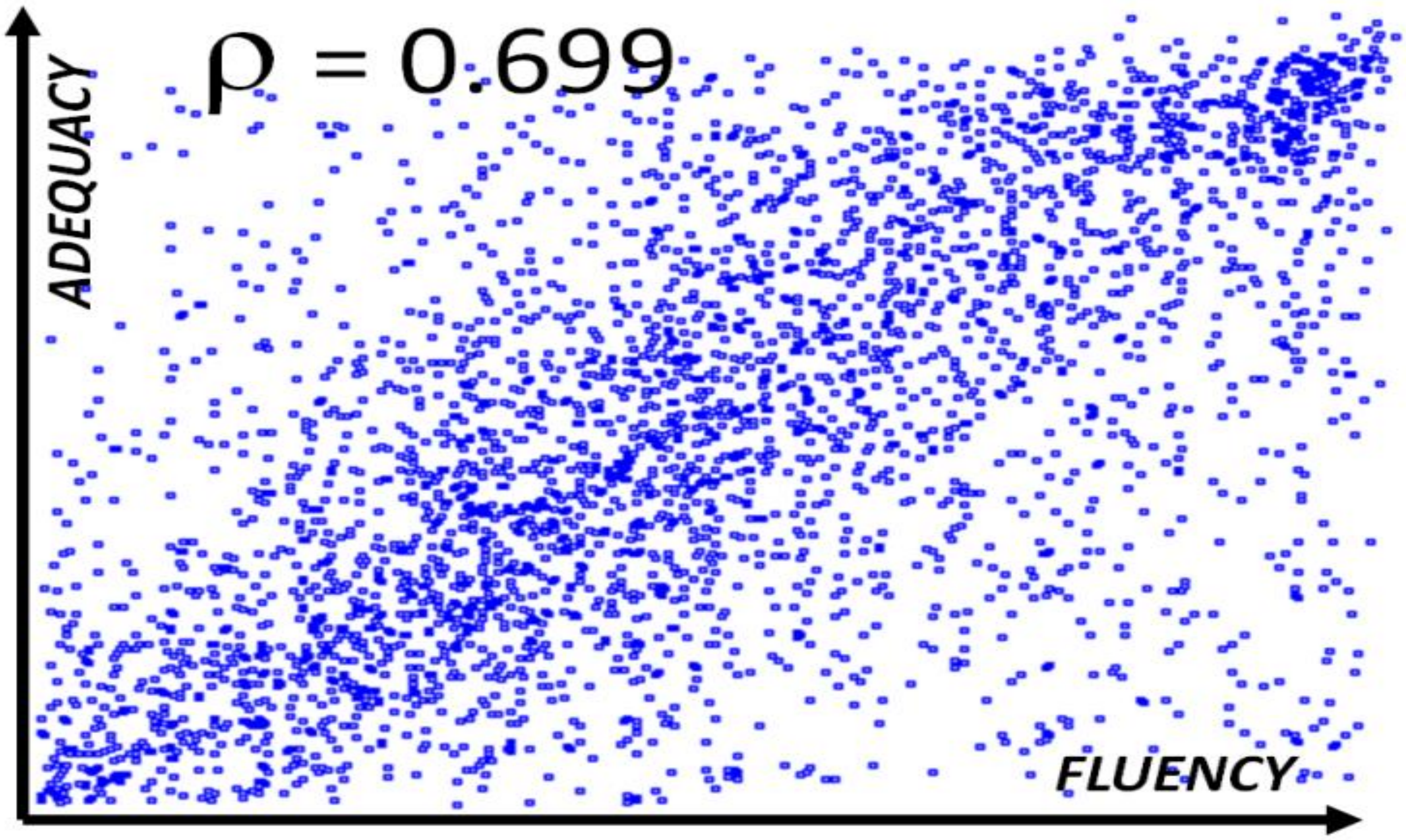


Comparative Evaluation Results

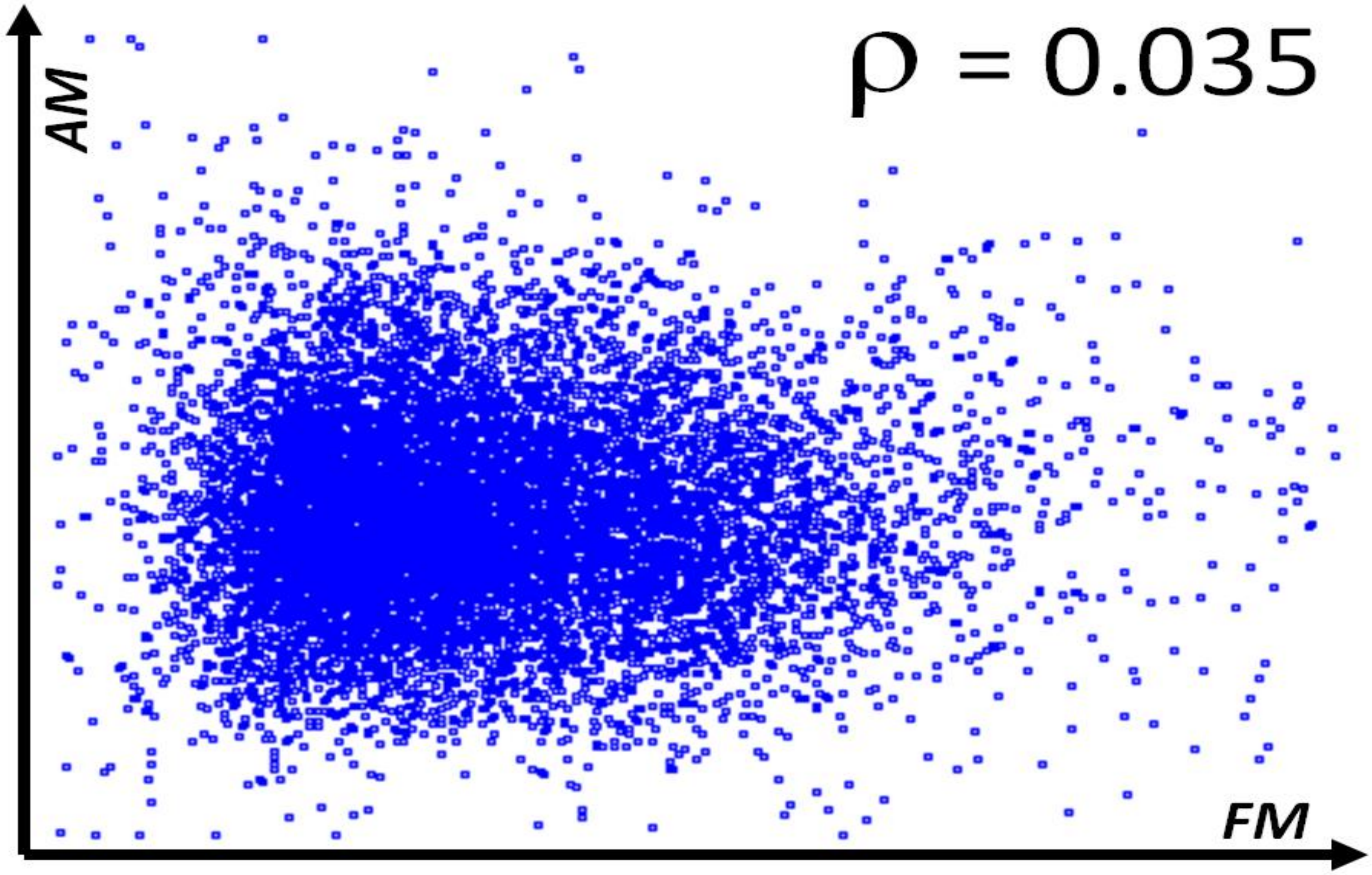
Metric	α	Adequacy	Fluency
BLEU	-	0.4107	0.4432
Meteor	-	0.3505	0.3626
NIST	-	0.3226	0.3444
TER-Plus	-	0.3068	0.3170
mAM	-	0.3435	0.3245
xAM	-	0.1291*	0.0330*
FM	-	0.3408	0.4267
$m\mathrm{AM\text{-}FM}_{\mathrm{HM}}$	0.10	0.4473	0.4977
$m\mathrm{AM\text{-}FM}_{\mathrm{WM}}$	0.60	**0.4574**	0.5036
$m\mathrm{AM\text{-}FM}_{\mathrm{L2}}$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

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

Human Adequacy and Fluency



AM and FM Metrics



Conclusions

- We have proposed a new evaluation framework for MT evaluation operating on a continuous space
- mAM-FM achieve better correlations with human evaluations for both adequacy and fluency than other conventional metrics
- xAM-FM allows for quality assessment without the need for a set of reference translations, its performance is still comparable to other state-of-the-art automatic evaluation metrics

Institute for
Infocomm Research
A*STAR

Thank You

Online:

www.i2r.a-star.edu.sg
www.facebook.com/i2r.research