研究会報告 2

「自動評価法を用いた機械翻訳の定量的評価」

研究報告２

自動評価法を用いた機械翻訳の定量的評価

越前谷博（北海学園大学）・磯崎秀樹（岡山県立大学）

目次

自動評価法とは

- なぜ必要なのか
  - 人間による評価は精度は高いが、時間やコストがかかり、再現性の点で問題がある
  - 機械翻訳システムの開発サイクルのスピードアップに有効

自動評価法とは

- 機械翻訳システムの訳文に対し、定量的な評価を完全自動で行うための技術
  - 入力：機械翻訳システムの訳文（システム訳）、人手による正しい訳文（参照訳）
  - 出力：スコア（例：0.0〜1.0）
- システム訳に対する評価単位：セグメントレベル（1文）、システムレベル（複数文）
- 自動評価法に対する評価（メタ評価）：自動評価法によるスコアと人手評価によるスコアと間の相関を求める（例：スピアンマンの相関係数）

自動評価法とは

- どんな自動評価法が求められているのか
  - 人間による評価との相関が高い
  - 処理速度が速い
  - 機械翻訳システムへのフィードバックに利用できる（どこが悪いのかを示してくれる）

自動評価法における動向
～Workshop on Statistical Machine Translationに参加して

自動評価法における動向：Workshop on Statistical Machine Translation（WMT）

- 2006年よりACL主催の国際会議のワークショップとして毎年開催されている。
- 機械翻訳に関するいくつかのタスクを選定し、タスクごとに評価ワークショップを実施
- EU言語を対象とした機械翻訳技術の進展を目的とするThe EuroMatrix（Statistical and Hybrid Machine Translation Between All European Languages）Projectの活動の一つとして始まった。

自動評価法における動向：WMT2014

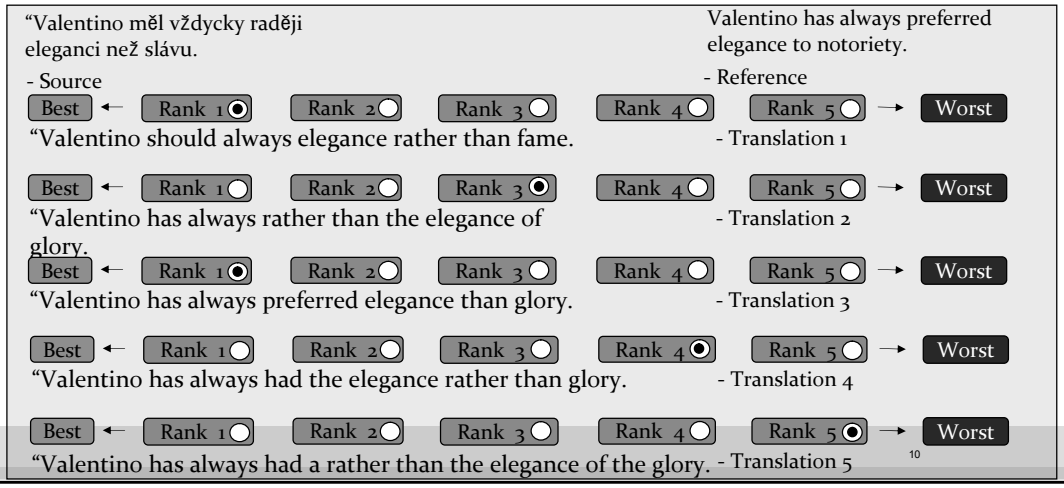
- WMT2014の概要
  - 2014年6月26日～27日、ACL2014のワークショップとしてボルチモアにて開催
  - 対象タスク
    - 翻訳タスク（Translation task）
    - 自動評価タスク（Metrics task）
    - 品質推定タスク（Quality Estimation task）
    - 医療翻訳タスク（Medical translation task）
    - その他：Data and Adaptation、Translation Models

自動評価法における動向：WMT2014

- 自動評価タスクにおけるテストコレクション
  - システム訳
    - 分野：オンラインニュース記事
    - 翻訳タスクに提出された110の機械翻訳システムのシステム訳を使用
    - 言語ペアとテストセット：French-English：3,003文, Hindi-English：2,507文, German-English：3,003文, Czech-English：3,003文, Russian-English：3,003文
    - 機械翻訳システム：cs-en:**5**システム、de-en:**13**システム、en-cs:**10**システム、en-de:**18**システム、en-fr:**13**システム、en-hi:**12**システム、en-ru:**9**システム、fr-en:**8**システム、hi-en:**9**システム、ru-en:**13**システム (en: English, cs: Czech, de: German, fr: French, hi: Hindi, ru: Russian)
    - セグメント数：cs-en:**15,015**文、de-en:**339,039**文、en-cs:**30,030**文、en-de:**49,266**文、en-fr:**39,039**文、en-hi:**30,084**文、en-ru:**27,027**文、fr-en:**24,024**文、hi-en:**22,563**文、ru-en:**39,039**文　　トータル：**315,126**文
- データの提出
  - システム訳と参照訳を用いて、開発した自動評価法よりスコアを求める
  - システムレベル：110スコア、セグメントレベル：315,126スコア

自動評価法における動向：WMT2014

- 自動評価タスクにおけるテストコレクション
  - 人手評価



自動評価法における動向：WMT2014

- 自動評価タスクにおける参加チーム
  - 12のグループより23の自動評価法が参加

Metrics	Sys	Seg	Authors
APAC	●	●	Hokkai-Gakuen University（Echizen'ya, 2014）
BEER		●	University of Amsterdam（Stanojevic and Sima'an,2014）
RED-＊	●	●	Dublin City University（Wu and Yu,2014）
DISCO TK-＊	●	●	Qatar Computing Research Institute（Guzman et al.,2014）
ELEXR	●		University of Tehran（Mahmoudi et al.,2014）
LAYERED	●		Indian Institute of Tech.（Gautam and Bhattacharyya,2014）
METEOR	●	●	Carnegie Mellon University（Denkowski and Lavie,2014）
AMBER	●	●	National Research Council of Canada（Chen and Cherry, 2014）
BLEU-NRC	●	●	National Research Council of Canada（Chen and Cherry,2014）
PARMESAN	●		Charles University in Prague（Barancikova,2014）
TBLEU	●		Charles University in Prague（Libovicky and Pecina,2014）
UPC-＊	●	●	Technical University of Catalunya（Gonzalez et al.,2014）
VERTA-＊	●	●	University of Barcelona（Comelles and Atserias,2014）

自動評価法における動向：WMT2014

- システムレベルのメタ評価
  - ピアソンの相関係数

$$r = \frac{\sum_{i=1}^{n}(H_i - \overline{H})(M_i - \overline{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \overline{H})^2}\sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2}}$$

MTシステムS_iに対する人手評価のスコア： H_i　人手評価のスコアの平均：$\overline{H}$
MTシステムS_iに対する自動評価法のスコア： M_i 自動評価法のスコアの平均：$\overline{M}$

  - 人手評価
    - TrueSkillを使用・・・ベイズ理論に基づくランキングアルゴリズム

自動評価法における動向：WMT2014

- システムレベルのメタ評価
  - 訳文：into English

From	fr	de	hi	cs	ru	Avg
DISCOTK-PARTY-TUNED	.98	**.94**	.96	.97	**.87**	**.94**
LAYERED	.97	.89	**.98**	.94	.85	.93
DISCOTK-PARTY	.97	.92	.86	.98	.86	.92
UPC-STOUT	.97	.91	.90	.95	.84	.91
VERTA-W	.96	.87	.92	.93	.85	.91
VERTA-EQ	.96	.85	.93	.94	.84	.90
TBLEU	.95	.83	.95	.96	.80	.90
BLEU-NRC	.95	.82	.96	.95	.79	.89
BLEU	.95	.83	.96	.91	.79	.89
UPC-IPA	.97	.89	.91	.82	.81	.88
CDER	.95	.82	.83	.97	.80	.87
APAC	.96	.82	.79	.98	.82	.87
REDSYS	**.98**	.90	.68	.99	.81	.87
REDSYSSENT	.98	.91	.64	**.99**	.81	.87
NIST	.96	.81	.78	.98	.80	.87
DISCOTK-LIGHT	.96	.93	.56	.95	.79	.84
METEOR	.98	.93	.46	.98	.81	.83
WER	.95	.76	.61	.97	.81	.82
AMBER	.95	.91	.51	.74	.80	.78
ELEXR	.97	.86	.54	.94	-.40	.58

自動評価法における動向：WMT2014

- システムレベルのメタ評価
  - 訳文：out of English

Into	fr	hi	cs	ru	Avg	de
NIST	.94	.98	.98	.93	**.96**	.20
CDER	.95	.95	.98	.94	.95	.28
AMBER	.93	**.99**	.97	.93	.95	.24
METEOR	.94	.98	.98	.92	.95	.26
BELU	.94	.97	.98	.91	.95	.22
PER	.94	.93	**.99**	**.94**	.95	.19
APAC	.95	.94	.97	.93	.95	.35
TBLEU	.93	.97	.97	.91	.95	.24
BLEU-NRC	.93	.97	.97	.90	.95	.20
ELEXR	.89	.96	.98	.94	.94	.26
TER	.95	.83	.98	.93	.92	.32
WER	**.96**	.52	.98	.93	.85	**.36**
PARMESAN	-	-	.96	-	.96	-
UPC-IPA	.94	-	.97	.92	.94	.28
REDSYSSENT	.94	-	-	-	.94	.21
REDSYS	.94	-	-	-	.94	.21
UPC-STOUT	.94	-	.94	.92	.93	.30

自動評価法における動向：WMT2014

- セグメントレベルのメタ評価
  - ケンドールの順位相関係数

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|}$$

自動評価法のスコアと人手評価のスコアが一致　：*Concordant*

自動評価法のスコアと人手評価のスコアが不一致：*Discordant*

$\tau \in [-1, 1]$

  - 人手評価

Best ← Rank 1 ◉ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst

"Valentino should always elegance rather than fame. - Translation 1

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ◉ Rank 4 ○ Rank 5 ○ → Worst

"Valentino has always rather than the elegance of glory. - Translation 2

自動評価法における動向：WMT2014

- セグメントレベルのメタ評価
  - ケンドールの順位相関係数
  - 例：

Human	Metric	結果
A<B	A<B	一致:1
C>A	C>A	一致:1
C>B	C<B	不一致:-1

$$\tau = \frac{2 \cdot 1 + 1 \cdot (-1)}{2 + 1} = \frac{1}{3}$$

  - WMT2014 variant
    - 自動評価法の結果のみが"="の場合は0とする
    - その場合、分母のみが増加

		Metric		
		<	=	>
Human	<	1	0	-1
	=	X	X	X
	>	-1	0	1

自動評価法における動向：WMT2014

- セグメントレベルのメタ評価
  - 訳文：into English
  - ペア数：fr-en：26,090
    de-en：25,260
    hi-en：20,900
    cs-en：21,130
    ru-en：34,460

From	fr	de	hi	cs	ru	Avg
DISCOTK-PARTY-TUNED	**.43**	**.38**	.43	**.33**	**.35**	**.39**
BEER	.42	.34	**.44**	.28	.33	.36
REDCOMBSENT	.41	.34	.42	.28	.34	.36
REDCOMBSYSSENT	.41	.34	.42	.28	.34	.36
METEOR	.41	.33	.42	.28	.33	.35
REDSYSSENT	.40	.34	.39	.28	.32	.35
REDSENT	.40	.34	.38	.28	.32	.35
UPC-IPA	.41	.34	.37	.27	.32	.34
UPC-STOUT	.40	.34	.35	.28	.32	.34
VERTA-W	.40	.32	.39	.26	.31	.34
VERTA-EQ	.41	.31	.38	.26	.31	.34
DISCOTK-PARTY	.39	.33	.36	.26	.31	.33
AMBER	.37	.31	.36	.25	.29	.32
BLEU-NRC	.38	.27	.32	.23	.27	.29
SENTBLEU	.38	.27	.30	.21	.26	.29
APAC	.36	.27	.29	.20	.28	.28
DISCOTK-LIGHT	.31	.22	.24	.19	.21	.23
DISCOTK-LIGHT-KOOL	.00	.00	.00	.00	.00	.00

自動評価法における動向：WMT2014

- セグメントレベルのメタ評価
  - 訳文：out of English
  - ペア数：en-fr：33,350
    en-de：54,660
    en-hi：28,120
    en-cs：55,900
    en-ru：28,960

Into	fr	de	hi	cs	ru	Avg
BEER	.29	**.27**	.25	**.34**	**.44**	**.32**
METEOR	.28	.24	.26	.32	.43	.31
AMBER	.26	.23	**.29**	.30	.40	.30
BLEU-NRC	.26	.20	.23	.30	.39	.28
APAC	.25	.21	.20	.29	.39	.27
SENTBLEU	.26	.19	.23	.29	.38	.27
UPC-STOUT	.28	.23	-	.28	.42	.30
UPC-IPA	.26	.23	-	.30	.43	.30
REDSENT	**.29**	.24	-	-	-	.27
REDCOMBSYSSENT	.29	.24	-	-	-	.27
REDCOMBSENT	.29	.24	-	-	-	.27
REDSYSSENT	.29	.24	-	-	-	.26

自動評価法における動向：WMT2014

- システムレベルの総評
  - 相関係数が0.8〜1.0の範囲であり、全体的に高い相関である
  - out of Englishにおいてベースライン（NIST, CDER, BLEU, PER）が高順位である
    - English-Hindiを除くとWERも高順位である
  - into Germanの相関係数が非常に低い
    - 機械翻訳システムの数（18）が他の言語間より多かった。
    - 自動評価法において、似たような性能のシステムを差別化することは難しい。
  - METEORではnon-Latin scriptから英語の順位が低い
- セグメントレベルの総評
  - 相関係数は約0.4であり、まだまだ不十分

⇩

自動評価タスクは変わらず興味深いタスクである
（12チームが参加）

自動評価法における動向：WMT2014

- WMT2014に参加しての感想
  - 提案手法（APAC）の位置づけの把握に有効
    - 参加前：システムレベルではそれほど有効ではないが、セグメントレベルでは有効
    - 結果：システムレベルはまあまあの順位だが、セグメントレベルの順位は低い
  - 似たような性能のシステムであっても正しく評価できなければならない

Into	fr	de	hi	cs	ru	Avg
APAC	.95	.35	.94	.97	.93	**.83**
CDER	.95	.28	.95	.98	.94	.82
METEOR	.94	.26	.98	.98	.92	.82
AMBER	.93	.24	**.99**	.97	.93	.81
NIST	.94	.20	.98	.98	.93	.81
ELEXR	.89	.26	.96	.98	.94	.81
BELU	.94	.22	.97	.98	.91	.80
TBLEU	.93	.24	.97	.97	.91	.80
TER	.95	.32	.83	.98	.93	.80
PER	.94	.19	.93	**.99**	**.94**	.80
BLEU-NRC	.93	.20	.97	.97	.90	.80
WER	**.96**	**.36**	.52	.98	.93	.75
PARMESAN	-	-	-	.96	-	.96
UPC-IPA	.94	.28	-	.97	.92	.78
UPC-STOUT	.94	.30	-	.94	.92	.78
REDSYSSENT	.94	.21	-	-	-	.58
REDSYS	.94	.21	-	-	-	.58

参考文献：

[1] M. Macháček and O. Bojar: Results of the WMT14 Metrics Shared Task, Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.293-301 (2014).

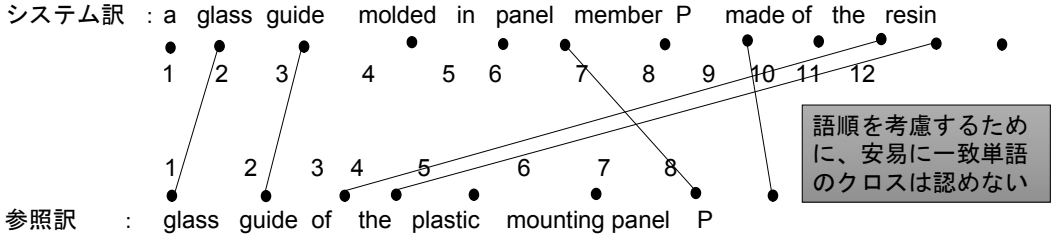
[2] O. Bojar, C. Buck, C. Federman, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia and A. Tamchyna: Findings of the 2014 Workshop on Statistical Machine Translation, Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.12-58 (2014).

自動評価法：APAC

自動評価法：APAC

- 特徴
  - 多義性のある一致単語列（チャンク）を大局的な観点から一意に決定：正しいチャンクを決定
  - 一致単語の語順の違いに柔軟に対応：パラメータの使用

- チャンクの決定方法



自動評価法：APAC

$$score = \sum_{c \in c\_num} \left(length(c)^{\beta} \times pos \right)$$

$$pos = \left(1.0 - \left| \frac{c_i}{m} - \frac{c_j}{n} \right| \right)$$

- チャンクの決定方法

候補1：

システム訳 ：a glass guide molded in panel member P made of the resin

参照訳 ：glass guide of the plastic mounting panel P

score = 3.499

候補2：

システム訳 ：a glass guide molded in panel member P made of the resin

参照訳 ：glass guide of the plastic mounting panel P

score = 3.446

パラメータβ：デフォルト値は1.2

自動評価法：APAC

- スコアの算出方法[1]

システム訳 a glass guide molded in panel member P made of the resin

参照訳 ：glass guide of the plastic mounting panel P

⇩ チャンクを再帰的に決定

システム訳 a glass guide molded in panel member P made of the resin

参照訳 ：glass guide of the plastic mounting panel P

$$R = \left(\frac{\sum_{i=0}^{RN-1} \left(\alpha^i \sum_{c \in c\_num} length(c)^{\beta} \right)}{m^{\beta}} \right)^{\frac{1}{\beta}}$$

$$P = \left(\frac{\sum_{i=0}^{RN-1} \left(\alpha^i \sum_{c \in c\_num} length(c)^{\beta} \right)}{n^{\beta}} \right)^{\frac{1}{\beta}}$$

$$AE\ score = \frac{(1+\gamma^2)RP}{R+\gamma^2 P}$$

パラメータ α：デフォルト値は0.1　パラメータβ：デフォルト値は1.2　AE score = 0.3268

[1] H. Echizen-ya and K. Araki:Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum, Proceedings of the Eleventh Machine Translation Summit (MT SUMMIT XI), pp.151-158 (2007).

自動評価法：APAC

- 改良[2]
  - 問題点：短い文のスコアが過度に小さくなる

⇩

短い文ほど不一致単語の重みが大きくなる

システム訳　:the doctor treated a patient
参照訳　　:　the doctor cured a patient

$$R = \left\{ \left(\frac{\sum_{i=0}^{RN-1} \left(\alpha^i \times Ch\_score \right)}{m^\beta} \right)^{\frac{1}{\beta}} + 0.5 \times Priz\_m \right\} / 2.0$$

$$Prize\_m = \frac{1}{\log(m) + 1}$$

$$P = \left\{ \left(\frac{\sum_{i=0}^{RN-1} \left(\alpha^i \times Ch\_score \right)}{n^\beta} \right)^{\frac{1}{\beta}} + 0.5 \times Priz\_n \right\} / 2.0$$

$$Prize\_n = \frac{1}{\log(n) + 1}$$

$$Ch\_score = \sum_{c \in c\_num} length(ch)^\beta$$

[2] H. Echizen-ya, K. Araki and E. Hovy: Application of Prize based on Sentence Length in Chunk-based Automatic Evaluation of Machine Translation, Results of the WMT14 Metrics Shared Task, Proceedings of the Ninth Workshop on Statistical Machine Translation, pp.381-386 (2014).

自動評価法：APAC

- 性能評価
  - WMT2012におけるシステムレベルの相関係数（Spearman's rank）

Metrics	cs-en(6)	de-en(16)	es-en(12)	fr-en(15)	Avg.	Rank
APAC	0.886	0.650	0.958	0.811	0.826	6
IMPACT	0.886	0.676	0.958	0.807	0.832	4
RIBES	0.943	0.732	0.944	0.814	0.858	2
METEOR	0.943	0.841	0.979	0.818	0.895	1
BLEU	0.886	0.674	0.958	0.796	0.828	5
NIST	0.943	0.700	0.944	0.779	0.841	3

  - WMT2012におけるセグメントレベルの相関係数（Kendall tau rank）

Metrics	cs-en (11,155)	de-en (12,042)	es-en (9,880)	fr-en (11,682)	Avg.	Rank
APAC	0.185	0.204	0.209	0.226	0.206	3
IMPACT	0.189	0.207	0.208	0.226	0.207	2
RIBES	0.055	0.125	0.114	0.115	0.102	4
METEOR	0.223	0.279	0.248	0.243	0.248	1

自動評価法：APAC

- 性能評価
  - WMT2013におけるシステムレベルの相関係数（Spearman's rank）

Metrics	cs-en(11)	de-en(17)	es-en(12)	fr-en(13)	ru-en(19)	Avg.	Rank
APAC	0.900	0.904	0.916	0.934	0.709	0.873	4
IMPACT	0.909	0.909	0.937	0.934	0.721	0.882	2
RIBES	0.900	0.912	0.930	0.978	0.670	0.878	3
METEOR	0.982	0.946	0.923	0.967	0.889	0.941	1
BLEU	0.945	0.897	0.853	0.951	0.614	0.852	5
NIST	0.900	0.828	0.804	0.786	0.465	0.757	6

  - WMT2013におけるセグメントレベルの相関係数（Kendall tau rank）

Metrics	cs-en (85,469)	de-en (128,668)	es-en (67,832)	fr-en (80,741)	ru-en (151,422)	Avg.	Rank
APAC	0.144	0.163	0.169	0.139	0.121	0.147	3
IMPACT	0.148	0.167	0.176	0.142	0.123	0.151	2
RIBES	0.044	0.063	0.056	0.018	0.003	0.037	4
METEOR	0.222	0.236	0.241	0.194	0.226	0.224	1

自動評価法：APAC

- 性能評価(JE)
  - NTCIR-7におけるシステムレベルの相関係数（Spearman's rank）

Metrics	Adequacy(15)	Fluency(15)	Avg.	Rank
APAC	0.872	0.805	0.839	2
IMPACT	0.872	0.805	0.839	2
RIBES	0.963	0.918	0.941	1
METEOR	0.424	0.380	0.402	6
BLEU	0.582	0.586	0.584	4
NIST	0.578	0.568	0.573	5

  - NTCIR-7におけるセグメントレベルの相関係数（Kendall tau rank）

Metrics	Adequacy(1,500)	Fluency(1,500)	Avg.	Rank
APAC	0.494	0.489	0.491	1
IMPACT	0.482	0.476	0.479	2
RIBES	0.370	0.341	0.356	4
METEOR	0.366	0.383	0.375	3

自動評価法：APAC

- 性能評価(JE)
  - NTCIR-9におけるシステムレベルの相関係数（Spearman's rank）

Metrics	Adequacy(19)	Acceptance(14)	Avg.	Rank
APAC	0.182	0.298	0.240	2
IMPACT	0.182	0.298	0.240	2
RIBES	0.660	0.640	0.650	1
METEOR	-0.081	0.015	-0.033	5
BLEU	-0.123	0.059	-0.032	4
NIST	-0.344	-0.275	-0.309	6

  - NTCIR-9におけるセグメントレベルの相関係数（Kendall tau rank）

Metrics	Adequacy(5,700)	Acceptance(5,700)	Avg.	Rank
APAC	0.250	0.261	0.256	2
IMPACT	0.242	0.250	0.246	3
RIBES	0.281	0.339	0.310	1
METEOR	0.167	0.217	0.192	4

自動評価法：APAC

- APACの特徴
  - Chef's tips for evaluation

データ		優劣
WMT	システムレベル	METEOR > RIBES > APAC
	セグメントレベル	METEOR > APAC > RIBES
NTCIR	システムレベル	RIBES > APAC > METEOR
	セグメントレベル	APAC > METEOR > RIBES (NTCIR-7)
	セグメントレベル	RIBES > APAC > METEOR (NTCIR-9)

  - 相対的には安定した性能を示している。

自動評価法：RIBES

自動評価法：RIBES

- システム訳と参照訳の間の**語順の近さ**を測定
- 日英・英日の翻訳において**人手評価と強い相関**がある

NTCIR-7 日英翻訳でのメタ評価

妥当性とのシステムレベルの相関、単一参照訳、スピアマンの相関係数

BLEU	METEOR	ROUGE-L	IMPACT	RIBES
0.515	0.490	0.903	0.826	**0.947**

自動評価法：RIBES

- EMNLP版[1]のRIBESは以下の式で定義される

$$\mathrm{RIBES} \overset{\mathrm{def}}{=} \mathrm{NKT} \times P^{\alpha}$$

- NKT $\overset{\mathrm{def}}{=} \frac{\tau + 1}{2}$ は正規化したKendall's τ
  - システム訳と参照訳で共通する単語の**語順の近さ**を表す。
- *P*は単語の適合率
  - α（$0 \leq \alpha \leq 1$）は*P*の影響を制御するパラメータ
  - デフォルト値は0.2
- （低評価）0.0 ≤RIBES ≤1.0（高評価）

[1] H. Isozaki, T. Hirao, K. Duh, K. Sudoh and H. Tsukada: Automatic Evaluation of Translation Quality for Distant Language Pairs, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP2010), pp.944-952 (2010).

自動評価法：RIBES

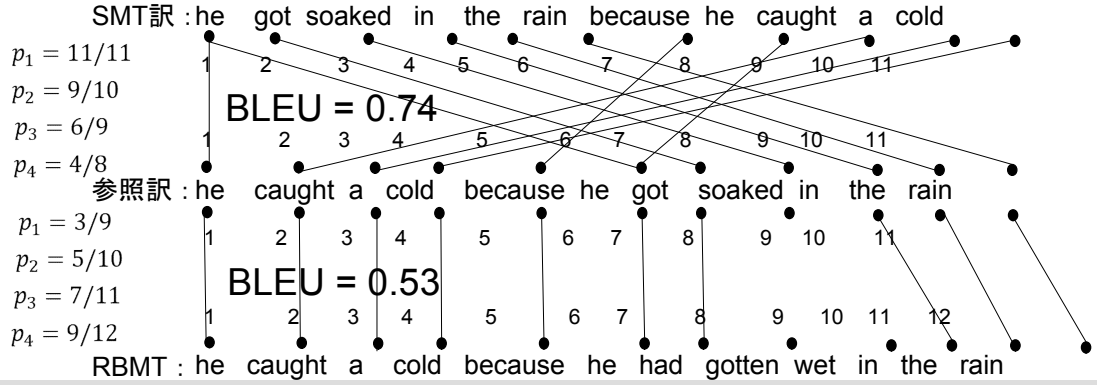
- BLEUの問題点
  - SMTの語順が大きく誤っていても高いスコアとなる。
  - 因果関係が逆の例

参照訳：
He caught a cold because he got soaked in the rain.

SMT訳：
He got soaked in the rain because he caught a cold.

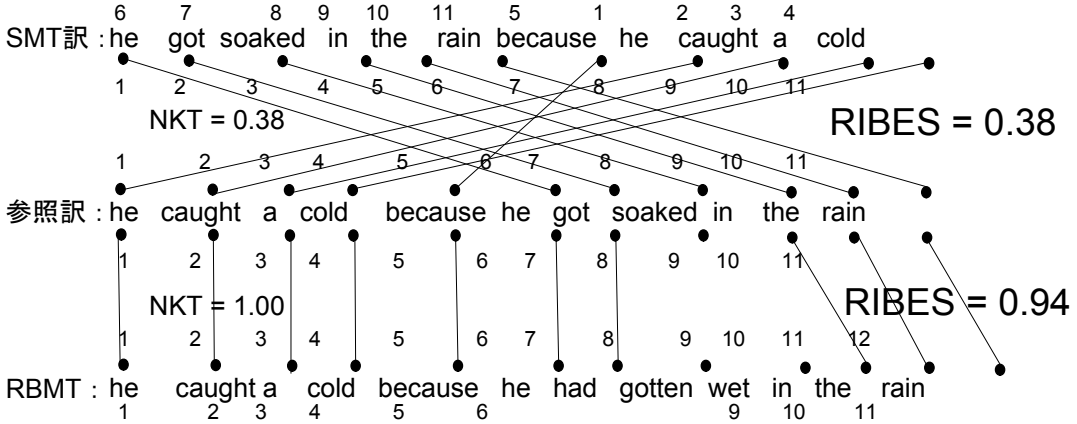
自動評価法：RIBES

- BLEUの問題点
  - SMTの語順が大きく誤っていても高いスコアとなる。
  - 因果関係が逆の例



自動評価法：RIBES

- RIBESの評価
  - SMT訳よりもRBMTを高く評価



自動評価法：RIBES

- RIBESの改良
  - EMNLP版のRIBESに対して、BLEUのBrevity Penaltyを導入

参照訳： John went to a restaurant yesterday

システム訳 : to a

語順（NKT）もユニグラム適合率（P）も完全一致なので、従来だと1.0となってしまう。

  - 以下の式で定義[2]

$$\mathrm{RIBES} \stackrel{\mathrm{def}}{=} \mathrm{NKT} \times P^{\alpha} \times \mathbf{BP}^{\beta}$$

  - デフォルト値はα=0.25、β=0.10

http://www.kecl.ntt.co.jp/icl/lirg/ribes

[2] 平尾、磯崎、須藤、Duh、塚田、永田： 語順の相関に基づく機械翻訳の自動評価法、自然言語処理、Vol. 21、No. 3, pp.421-444 (2014).

自動評価法：RIBES

- 性能評価
  - NTCIR-9, 10 Patent MTがRIBESを標準的な自動評価法として採用

NTCIR-9, 10 Patent MTでのメタ評価

妥当性とのシステムレベルの相関、単一参照訳、スピアマンの相関係数

	BLEU	NIST	RIBES
NTCIR-9 JE	**-0.042**	-0.114	**0.632**
NTCIR-9 EJ	**-0.029**	-0.074	**0.716**
NTCIR-10 JE	**0.31**	0.36	**0.88**
NTCIR-10 EJ	**0.36**	0.22	**0.79**

  - 現在、日英・英日翻訳のほとんどの論文がRIBESを使用
  - 言語処理学会第20回年次大会（NLP2014）にて18本の機械翻訳の論文がRIBESを使用

自動評価法：RIBES

- RIBESのさらなる改良

  日本語は語順が比較的自由（スクランブリング）。

  太郎はイタリアでピザを食べた。

  イタリアで太郎はピザを食べた。

  日本語訳の評価をする場合に、この点を考慮すべき。

  与えられた参照文の係り受け木から、他の語順を自動生成して参照訳に追加

- RIBESの文レベルの相関係数が若干改善された。

  NTCIR-7 Mosesベースラインで Spearman's ρ が 0.607から 0.670 に向上など。

H. Isozaki, N. Kouchi, T. Hirao:
Dependency-based Automatic Enumeration of Semantically Equivalent Word Orders for Evaluating Japanese Translations, WMT-2014.

まとめ

- 現時点での最適な自動評価法は何か
  - 求めるものによって変わる
  - 一般的な翻訳データ(WMT)、特許翻訳データ(NTCIR)、対象言語、システムレベル、セグメントレベル
- 今後の課題
  - セグメントレベルの評価精度（相関係数）の向上