自動評価方法の研究者の立場から(その1)

自動評価手法がもたらした歓喜と失望、そして、希望

北海学園大学工学部生命工学科
越前谷　博

2012/9/7

発表の流れ

- 自動評価手法のスタンダード
- 準スタンダードな自動評価手法
- WMT09における自動評価手法
- 自動評価手法IMPACT
- 結論

自動評価手法のスタンダード

人手評価の種類

- 絶対評価・・・FluencyとAdequacy
  - Fluency・・・文法的な正確さとイデオムワードの選択
  - Adequacy・・・出力文に入力文の意味が正しく伝達されているか

Adequacy	
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

- 相対評価・・・比較評価（「どちらの出力文が良いか？」を判定）

人手評価の問題点[1]

- 評価者ごとのばらつき
  - Fluency・・・評価者ごとの平均スコアは2.33から3.67
  - Adequacy・・・評価者ごとの平均スコアは2.56から4.13
- 時間とコストがかかる
- SMTの研究者の研究費は潤沢とは言えない。また、短期間に何度も繰り返しシステムを評価したい。

⇩

機械翻訳自動評価手法への期待の高まり

[1] Philipp Koehn(2010) “Statistical Machine Translation,” Cambridge University Press

理想的な自動評価手法

- low-cost metric
- tunable metric
- meaningful metric

どのような自動評価手法が使用されているのか？

- 圧倒的にBLEUが使用されている

適合率と再現率

SYSTEM A: *Israeli officials* ~~responsibility~~ of *airport* ~~safety~~

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: *airport security Israeli officials are responsible*

$$precision = \frac{correct}{output - length} \qquad recall = \frac{correct}{reference - length}$$

$$f - measure = \frac{precision \times recall}{(precision + recall)/2}$$

$$PER = 1 - \frac{correct}{reference - length}$$

Metric	System A	System B
precision	0.50	0.57
recall	0.43	0.57
f-measure	0.46	0.14
PER	0.57	0.43

WER: Word Error Rate [2]

- 手法：レーベンシュタイン距離（編集距離）に基づく手法

SYSTEM A: *Israeli officials* responsibility of *airport* safety

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security *Israeli officials are responsible*

$$WER = \frac{substitutions + insertions + deletions}{reference - length}$$

Metric	System A	System B
WER	0.57	0.71

[2] Geroge Leusch, Nicola Ueffing and Hermann Ney(2003) “A Novel String-to-String Distance Measure With Applications to Machine Translation Evaluation,” Proc. of MT Summit IX, pp.240-247

WERの特徴

- スコア：0.0以上（スコアが小さいほど高い評価）
- 高速処理が可能
- 文単位の評価にも有効
- 語順に厳しい

BLEU: A Bilingual Evaluation Understudy[3]

- 自動評価手法で最もポピュラー
- 手法：n-gram適合率に基づく手法

$$p_n = \frac{\sum_{c \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in} Count(n-gram')}$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases}$$

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu(2002) "*BLEU: a Method for Automatic Evaluation of Machine Translation*," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318

BLEU

SYSTEM A: *Israeli officials* responsibility of *airport* safety
2-gram match　1-gram match

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: *airport security*　*Israeli officials are responsible*
2-gram match　4-gram match

n=1～4までのn-gram適合率

Metric	System A	System B
p_1 (1-gram)	3/6	6/6
p_2 (2-gram)	1/5	4/5
p_3 (3-gram)	0/4	2/4
p_4 (4-gram)	0/3	1/3

BLEUの特徴

- スコア：0.0～1.0（スコアが大きいほど高い評価）
- 高速処理が可能
- 文単位の評価に適さない

NIST: National Institute of Standards and Technology[4]

- 手法：相互情報量より重み付されたn-gram適合率に基づく手法

$$Info(w_1 \dots w_n) = \log_2 \left(\frac{the\ \#\ of\ occurances\ of\ w_1 \dots w_{n-1}}{the\ \#\ of\ occurrences\ of\ w_1 \dots w_n} \right)$$

$$Score = \sum_{n=1}^{N} \left\{ \frac{\sum_{all\ w_1 \dots w_n that\ co-occur} Info(w_1 \dots w_n)}{\sum_{all\ w_1 \dots w_n in\ sys\ output}(1)} \right\} \cdot exp\left\{ \beta \log^2 \left[min\left(\frac{L_{sys}}{\bar{L}_{ref}}, 1 \right) \right] \right\}$$

[4] NIST(2002) "*Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*"

NISTの特徴

- スコア：0.0以上（スコアが大きいほど高い評価）
- 高速処理が可能
- 文単位の評価に適さない
- 情報量を重み付けに用いることで意味も考慮

METEOR: Metric for Evaluation of Translation with Explicit Ordering[5]

- 手法：適合率だけでなく再現率も重要視した一致率に基づく手法

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R}$$

$$Pen = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}$$

$$score = (1 - Pen) \cdot F_{mean}$$

α、β、γ：パラメータ

[5] Alon Lavie and Abhaya Agarwal(2007) "*Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments*," Proceedings of the Second Workshop on Statistical Machine Translation, pages 228–231

METEORの特徴

- スコア：0.0～1.0（スコアが大きいほど高い評価）
- 単語の語形変化、類義語（WordNet）を利用
- 処理時間：BLEUやNISTよりは長い
- 文単位の評価に有効
- 適合率と再現率の両方を考慮

準スタンダートな自動評価手法

ROUGE: Recall-Oriented Understudy for Gisting Evaluation[6]

- 手法:
  - Longest Common Subsequence(LCS)に基づく手法:ROUGE-L
    - 参照訳の単語数に対するLCSの長さの割合を再現率、MT訳の単語数に対するLCSの長さの割合を適合率として、再現率と適合率のF値をスコアとしている
  - Weighted Longest Common Subsequenceに基づく手法:ROUGE-W
    - LCSに対して重み付けをし、ROUGE-Lと同様、再現率と適合率のF値をスコアとしている
  - n-gramに対して、語順の連続性の制約を排除した一致率(Skip-Bigram)に基づく手法:ROUGE-S

SYSTEM: police killed the gunman

REFERENCE: police kill the gunman

$$P = \frac{3}{C(4,2)}$$

[6] Chin-Yew Lin and Franz Josef Och(2004) “Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics,” Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318

ROUGEの特徴

- スコア:0.0～1.0(スコアが大きいほど高い評価)
- 高速処理が可能
- 文単位の評価にも有効
- 語順に厳しい
- 適合率と再現率の両方を考慮

TER: Translation Error Rate[7]

- 手法：編集距離の3つ操作（substitution, insertion, deletion）にshiftの操作を加えた4つの操作に基づく手法

$$\mathrm{TER} = \frac{substitutions + insertion + deletion + shift}{average\ \#\ of\ reference\ words}$$

REFERENCE: SAUDI ARABIA denied THIS WEEK information published in the AMERICAN new york times
insertion=1

$$TER = \frac{4}{13} = 0.31$$

sift=1　　substitutions=2
SYSTEM: THIS WEEK THE SAUDIS denied information published in the new york times

[7] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul(2006) “A Study of Translation Edit Rate with Targeted Human Annotation,” Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA), pp. 223-231

TERの特徴

- スコア：0.0以上（スコアが小さいほど高い評価）
- 高速処理が可能
- 文単位の評価にも有効

自動評価手法で使用する情報

- その他の自動評価手法
  - GTM（General Text Matcher）：チャンクに基づく手法
  - CDER（Cover Disjoint Error Rate）：レーベンシュタイン距離に基づく手法
  - NMG_WN（Normalized Mean Grams-Word Number）：最大のgram数に基づく構文的自然性に着目した手法
- 適合率、再現率の使用
- 一致単語の利用
  - 訳語の正確さを評価
- 一致単語の位置や出現順の利用
  - 流暢さを評価

自動評価手法の問題点

- 絶対評価・・・異なるアーキテクチャ間のMTの評価が不十分
  - SMTに対する評価は高いが、RBMTに対する評価は低い
- 相対評価・・・MTの改良においてベースラインと提案手法間の評価が不十分

The 2009 Workshop on Statistical Machine Translation（WMT09）における自動評価手法

WMT09[8]

- WMT:2006年から毎年開催されているSMTの評価型ワークショップ
- WMT09における自動評価手法のランキング

Metric	de-en (21)	fr-en (21)	es-en (13)	cz-en (5)	hu-en (6)	Avg.
ulc	**.78**	.92	.86	**1**	.6	**.83**
maxsim	.76	.91	**.98**	.7	.66	.8
rte(absolute)	.64	.91	.96	.6	**.83**	.79
meteor-rank	.64	**.93**	.96	.7	.54	.75
rte(pairwise)	.76	.59	.78	.8	**.83**	.75
以下14 metricsは省略						

[8] Chris Callison-Burch, Philipp Koehn, Christof Monz and Josh Schroeder(2009) “Findings of the 2009 Workshop on Statistical Machine Translation,” Proceedings of the 4th Workshop on Statistical Machine Translation (WMT09), pp. 1-28

ULC: Uniform Linear Metric Combinations[9]

- 手法：様々な自動評価手法を一つのセットとしてスコアを決定。その際、種々の言語レベルの知識も利用。

$$ULC_{X(a,R)} = \frac{1}{|X|}\sum_{x \in X} x(a,R)$$

[9] Jesús Giménez and Lluís Márquez(2007) “Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations,” Proceedings of IJCNLP, pp. 319-326

MaxSim: Maximum Similarity Metric[10]

- 手法：語彙情報を用いたn-gram一致率に基づく手法

$$sim - score = \frac{1}{|S|}\sum_{s=1}^{|S|} score_s$$

$$score_s = \frac{1}{N}\sum_{n=1}^{N} F_{s,n}$$

[10] Yee Seng Chan and Hwee Tou Ng(2008) “MAXSIM: An Automatic Metric for Machine Translation Evaluation Based on Maximum Similarity,” Proceedings of the Metrics-MATR Workshop of AMTA-2008, pp. 319-326

RTE: Recognition of Textual Entailment [11]

- 手法：Textual Entailment(TE)に基づく手法

SYSTEM: The virus did not infect anybody.

↓ entailment ↑

REFERENCE: No one was infected by the virus.

SYSTEM: Virus was infected.

↓ non entailment ↑

REFERENCE: No one was infected by the virus.

[11] Sebastian Padó, Michel Galley, Dan Jurafsky, Christopher D. Manning(2009) “Textual Entailment Features for Machine Translation Evaluation,” Proceedings of the 4th Workshop on Statistical Machine Translation

自動評価手法 IMPACT

IMPACT: Intuitive comMon PArts ConTinuum[12]

- 手法
  - 最長一致のn-gramに対するセンシティブな手法[Philpp Koehn(2010), “Statistical Machine Translation”]
  - ROUGEシリーズを発展させた手法[特許庁(2011), “特許審査関連情報の機械翻訳による英語提供に対する精度評価に係る調査報告書”]
  - 共通部分の再帰的な決定に基づく手法
- 人間はどのようなヒューリスティックスを使って、共通部分を決定しているのか？

記号列1:a c b a d

記号列2:a b c b c d a

共通部分の長さと位置が重要

⇩

LCSはこのタスクにおいて非常に有効

[12] Hiroshi Echizen-ya and Kenji Araki(2007), “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

IMPACT

- 共通部分の自動決定

記号列1:a c b a d

記号列2:a b c b c d a

		0	1	2	3	4	5	6
		a	b	c	b	c	d	a
	0	0	0	0	0	0	0	0
0	a	0	1	1	1	1	1	1
1	c	0	1	1	2	2	2	2
2	b	0	1	2	2	3	3	3
3	a	0	1	2	2	3	3	4
4	d	0	1	2	2	3	3	4

LCSの長さ

[0,0][1,2][2,3][3,6]

記号列1:[a] [c b] [a] d

記号列2:[a] b [c b] c d [a]

[0,0][1,2][2,3][4,5]

記号列1:[a] [c b] a [d]

記号列2:[a] b [c b] c [d] a

IMPACT

■共通部分の自動決定

$$pos_w = \left(1.0 - \left|\frac{posX(c)}{m} - \frac{posY(c)}{n}\right|\right)$$

$$RS = \left(\sum_{c \in LCS} \left(length(c)^{\beta} \times pos_w\right)\right)^{\frac{1}{\beta}}$$

1 2 3 4 5 6 7

X: [a] [c b] [a] d

Y: [a] b [c b] c d [a]

$$pos_w = \left(1.0 - \left|\frac{4}{5} - \frac{7}{7}\right|\right) = 0.80$$

1 2 3 4 5 6 7

X: [a] [c b] a [d]

Y: [a] b [c b] c [d] a

$$pos_w = \left(1.0 - \left|\frac{5}{5} - \frac{6}{7}\right|\right) = 0.85$$

1 2 3 4 5 6 7

X: [a] [c b] a [d]

Y: [a] b [c b] c [d] a

ここで終わると残された共通部分が無視される

IMPACT

■共通部分の再帰的な決定

1. 決定された共通部分を無視する。除外すると文字の位置が変わってしまうので、あくまでも無視する
2. 前述の式に基づき新たな共通部分を一意に決定する

1 2 3 4 5 6 7

X: [a] [c b] a [d]

Y: [a] b [c b] c [d] a

1 2 3 4 5 6 7

X: [a] [c b] a [d]

Y: [a] b [c b] c [d] a

1 2 3 4 5 6 7

X: [a] [c b] [a] [d]

Y: [a] b [c b] c [d] [a]

全ての共通部分を一意に決定する手法を考案できた

これは何に利用できるのか？

自動評価手法に利用

IMPACT

■スコア計算（一致度）

■共通部分の長さに応じた重み付けに基づくマッチング（パラメータβ）

$$\sum_{c \in CC} length(c)^{\beta}$$

β:1.0以上

X:[a] [c b] a [d]

Y:[a] b [c b] c [d] a

β=2.0の場合

$1^{2.0} + 2^{2.0} + 1^{2.0} = 6$

■語順の違い応じた重み付けに基づくマッチング（パラメータα）

$$\sum_{i=0}^{RN} \left(\alpha^{i} \sum_{c \in CC} length(c)^{\beta} \right)$$

α:1.0以下

X:[a] [c b] a [d]

Y:[a] b [c b] c [d] a

α=0.5の場合

$0.5^{0} \times 6 = 6$

X:[a] [c b] [a] [d]

Y:[a] b [c b] c [d] [a]

$0.5^{1} \times 1^{2.0} = 0.5$

⇩

$6 + 0.5 = 6.5$

IMPACT

■スコア計算（正規化）

■適合率

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

■再現率

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

■スコア

$$IMPACT = \frac{(1+\gamma^{2})PR}{\gamma^{2}P + R}$$

$$\gamma = \frac{P}{R}$$

X:a c b a d

Y:a b c b c d a

$$P = \sqrt{\frac{6.5}{5^{2.0}}} = 0.5099$$

$$R = \sqrt{\frac{6.5}{7^{2.0}}} = 0.3643$$

γ=1.0の場合

$$IMPACT = \frac{2 \times 0.5099 \times 0.3643}{0.5099 + 0.3643} = 0.4250$$

IMPACT

- 先行研究とのアプローチの違い
  - ルーベンシュタイン距離に基づく手法WER, ROUGE-LはLCSの長さがわかればそれで終了
  - METEORやGTMはどのように共通部分（チャンク）の決定における多義性を解決しているのかよくわからない。
  - LCSやレーベンシュタイン距離に基づく手法の一番の問題は語順の異なる共通部分は無視される

IMPACT

- 特徴
  - マッチングはすべての共通単語を考慮。共通部分ごとにその長さに応じた重み付けをパラメータで制御
  - 語順の違いをパラメータで制御
  - 適合率と再現率のF値を用い、文単位の評価を重視

IMPACT

- 性能評価・・・NTCIR-7の特許翻訳データにおけるメタ評価[13]
  - 文単位におけるadequacyのピアソンの相関係数

Metric	JAPIO (RBMT)	Kyoto-U (EBMT)	Moses (SMT)	計14MTの相関係数の平均
IMPACT	0.5980	**0.7125**	**0.7621**	**0.6258**
ROUGE-L	**0.6111**	0.7042	0.7480	0.6109
BLEU	0.4230	0.5884	0.5459	0.4515
NIST	0.4142	0.4092	0.3521	0.3423
NMG-WN	0.6067	0.6068	0.6770	0.5653
METEOR	0.3907	0.4947	0.2987	0.3370
WER	0.5519	0.6570	0.7491	0.5937

[13] Hiroshi Echizen-ya, Terumasa Ehara, Sayori Shimohata, Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro and Noriko Kando(2009) “Meta-Evaluation of Automatic Evaluation Methods for Machine Translation using Patent Translation Data in NTCIR-7,” Proceedings of the 3rd Workshop on Patent Translation, Page.9-16

IMPACT

- 性能評価・・・NTCIR-7の特許翻訳データにおけるメタ評価
  - 文単位におけるfluencyのピアソンの相関係数

Metric	JAPIO (RBMT)	Kyoto-U (EBMT)	Moses (SMT)	計14MTの相関係数の平均
IMPACT	0.5821	0.6320	0.6034	**0.5115**
ROUGE-L	**0.5925**	0.6347	0.5889	0.4953
BLEU	0.4488	0.5521	0.4783	0.3983
NIST	0.4987	0.3811	0.3116	0.3030
NMG-WN	0.5434	0.5799	**0.6308**	0.4992
METEOR	0.4420	0.4153	0.3351	0.3022
WER	0.5220	**0.6421**	0.6228	0.4879

IMPACT

■性能評価・・・NTCIR-7の特許翻訳データにおけるメタ評価

■文単位におけるadequacyのスピアマンの相関係数

Metric	JAPIO (RBMT)	Kyoto-U (EBMT)	Moses (SMT)	計14MTの相関係数の平均
IMPACT	0.5992	**0.7067**	**0.7411**	**0.6014**
ROUGE-L	**0.6092**	0.6983	0.7280	0.5872
BLEU	0.4327	0.5827	0.4987	0.4014
NIST	0.4218	0.4424	0.2950	0.3269
NMG-WN	0.5579	0.6524	0.7015	0.5813
METEOR	0.4018	0.4776	0.2861	0.3413
WER	0.5478	0.6480	0.7131	0.5478

IMPACT

■性能評価・・・NTCIR-7の特許翻訳データにおけるメタ評価

■文単位におけるfluencyのスピアマンの相関係数

Metric	JAPIO (RBMT)	Kyoto-U (EBMT)	Moses (SMT)	計14MTの相関係数の平均
IMPACT	0.5572	0.6454	0.6319	0.4915
ROUGE-L	**0.5646**	0.6428	0.6159	0.4712
BLEU	0.4286	0.5419	0.4740	0.3612
NIST	0.4559	0.4193	0.3006	0.2909
NMG-WN	0.5381	0.5850	**0.6502**	**0.4992**
METEOR	0.4438	0.4267	0.3264	0.3034
WER	0.5087	**0.6505**	0.6501	0.4454

IMPACT

- 性能評価・・・NTCIR-7の特許翻訳データにおけるメタ評価
  - RBMTにおけるfluencyでの人手評価と自動評価IMPACTのスコアの例

		human	IMPACT
source sentence	これらのガスは、所定の割合で混合して用いてもよい。	4	0.3917
system	you may use these gases mixing it by the given percentage.		
reference No.1	these gases might be used in mixture in a prescribed proportion.		
reference No.2	these gases might be used by mixing at a predetermined percentage.		
reference No.3	these gases might be mixed at a prescribed ratio and be used		
reference No.4	these gases can be mixed and used at a predetermined ratio.		

IMPACT

- 性能評価・・・特許文書のブロック別における評価
  - 技術分野別（電気、物理、化学）の特性・・・「電気」は他の2分野に比べ、評価精度（人手評価との相関）が高い
  - 文献種別（公開特許公報分データ、拒絶理由通知分データ）の特性・・・「公開特許公報分データ」の方が「拒絶理由通知分データ」よりも評価精度が高い
  - 部分別（課題、手段、実施、請求項）の特性・・・「請求項」の評価精度が高く、「実施」の評価精度が低い
  - 原文の長さ別（文字数50～100、文字数101～150、文字数151～）の特性・・・明確な傾向は読み取れなかった

「特許庁(2011), “特許審査関連情報の機械翻訳による英語提供に対する精度評価に係る調査報告書”」より

IMPACT

■その他の関連文献

[14]越前谷 博, 江原 暉将, 下畑 さより, 藤井 敦, 内山 将夫, 山本 幹雄, 宇津呂 武仁, 神門 典子 (2009) “NTCIR-7データを用いた機械翻訳自動評価規準のメタ評価, 平成20年度AAMT/Japio特許翻訳研究会報告書,” 一般財団法人日本特許情報機構, pp.2-13

[15]江原 暉将, 越前谷 博, 下畑 さより, 藤井 敦, 内山 将夫, 山本 幹雄, 宇津呂 武仁, 神門 典子 (2009) “機械翻訳精度の各種自動評価の比較,” Japio 2009 Year Book, 一般財団法人日本特許情報機構, pp.272-275

[16] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya and Sayori Shimohata(2010) “Overview of the Patent Translation Task at the NTCIR-8 Workshop, Proceedings of 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access,” pp.371-376

[17]越前谷 博, 下畑 さより (2010) “機械翻訳自動評価における名詞句チャンキングの利用,” 平成21年度AAMT/Japio特許翻訳研究会報告書, 一般財団法人日本特許情報機構, pp.2-11

[18] Hiroshi Echizen-ya and Kenji Araki(2010) “Automatic Evaluation Method for Machine Translation using Noun-Phrase Chunking,” Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), pp.108-117

IMPACTにおける今後の展開

■処理時間の短縮

- ■現在、取組中ではあるが、大幅な改善を実現

■言語知識の活用

- ■WMT09においても言語知識に基づく自動評価手法（ULC, MaxSim, RTE）の評価精度が高い

結論

結論

- 現在の自動評価手法は実際に利用可能か？

⇩

  - 実用レベルに達しているとは言えない
  - もし、使用するのであれば、
    - 相対評価に限定して、かつ大幅な改善が見込まれる場合に使用
    - データの種別に応じて、自動評価手法を使い分ける

- 今後、自動評価手法に明るい展望はあるのか？

⇩

  - 言語知識の利用は有効（利用容易性は低下する）
  - 我々、開発者（研究者）の努力が不可欠
  - 翻訳、自然言語処理分野に携わる方々の協力も不可欠