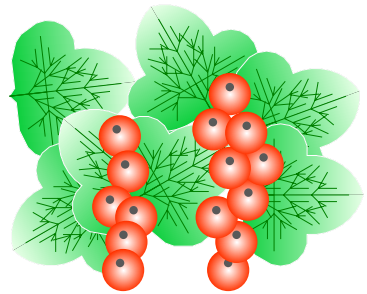
最近の自動評価法の研究動向とRIBES（ライビーズ）（最終版）

磯崎　秀樹

岡山県立大学　情報工学部



2012 年 9 月 7 日 (金)

平成 24 年度 AAMT/Japio 特許翻訳研究会
特許文書の機械翻訳結果評価方法検討会資料集

要旨

機械翻訳を改良するうえで、翻訳の品質を評価することは不可欠である。

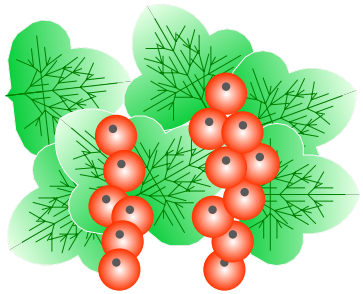
しかし、**人間の評価は時間やコストがかかり**、再現性にも問題がある。

そこで**翻訳自動評価法が不可欠**である。

近年、デファクト・スタンダードの自動評価法である BLEU の問題が指摘され、**新しい自動評価法の模索**が始まっている。

本発表では、**2010 年以降に発表**された翻訳自動評価法を紹介する。

Outline



BLEU の復習

参照訳 (人間が与えた理想的な訳) との近さを測定することで採点。

▼クリックすると、論文が表示されます。

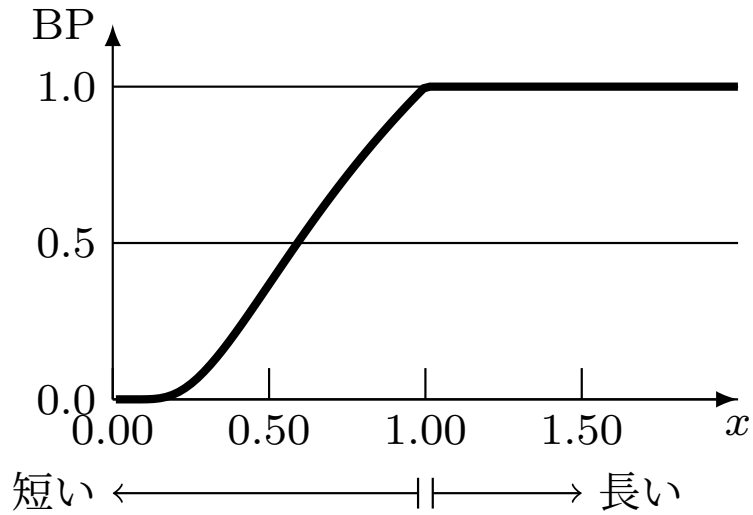
BLEU P02-1040 の基本は
1, 2, 3, 4 グラムの適合率の幾何平均 $\sqrt[4]{p_1 p_2 p_3 p_4}$。

適合率なので、自信のある部分だけ出力する方が得。

そうしたずるい出力に対して課されるのが **Brevity Penalty (BP)**。

$$\mathrm{BP} = \min\left(1, \exp\left(1 - \frac{\text{参照訳の語数}}{\text{機械訳の語数}}\right)\right)$$

$x = \dfrac{\text{機械訳の語数}}{\text{参照訳の語数}}$ としたとき、右図の形。



BLEU の復習

$$\textbf{BLEU} = \mathrm{BP} \times \sqrt[4]{p_1 p_2 p_3 p_4}$$

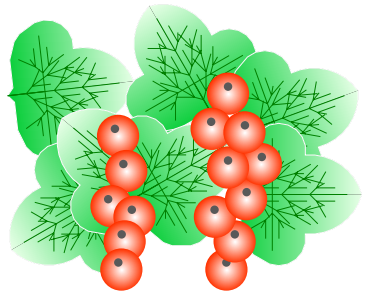
4 グラムがひとつはマッチしていないと 0 点

文単位で採点すると、ほとんどの文が 0 点になってしまう。

BLEU を使うときには、参照訳をいくつも用意して、しかも文単位ではなく、もっと大きい単位で評価するのが普通。

細かい採点ができないので、成績が悪い原因を突き止めることが難しい。

Outline



翻訳自動評価に関わることになったきっかけ

英日翻訳で **Head Finalization** WMT-2010 W10-1736, NLP10-B4-2, NTCIR9-SudohK, ACM TALIP-2012 という preordering (翻訳前並べ替え) を考案

日本語の語順にどれくらい近づいたか知りたかった。

順番を比べるなら**順位相関係数**、で、Kendall の τ を利用

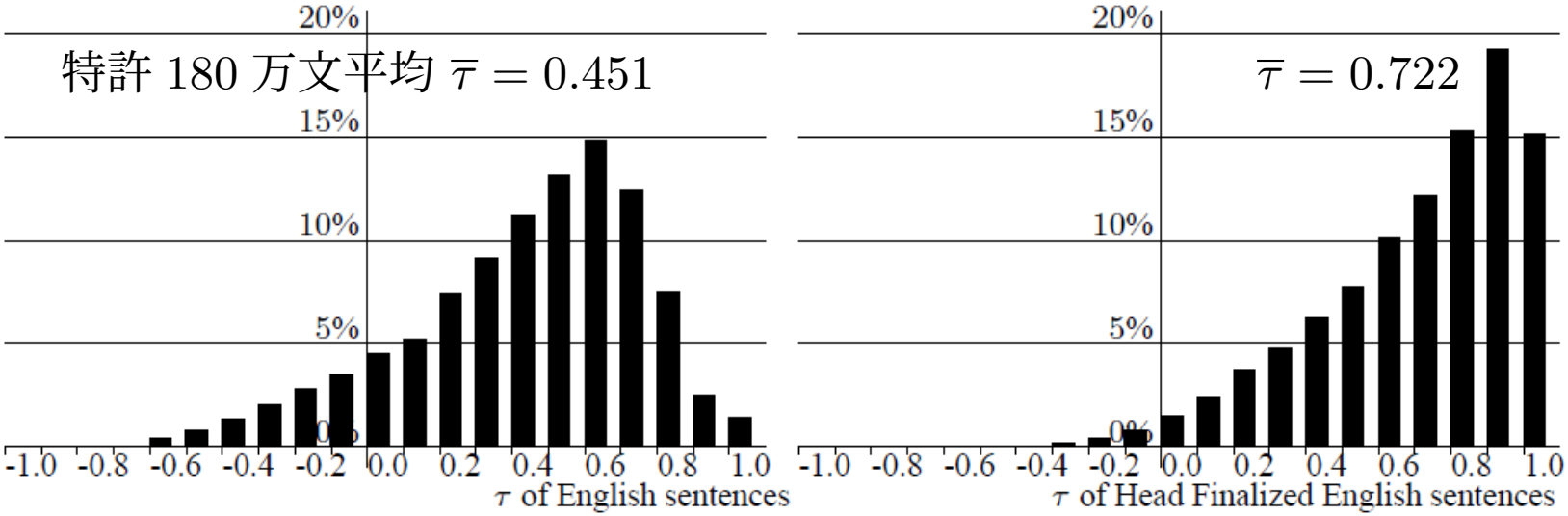
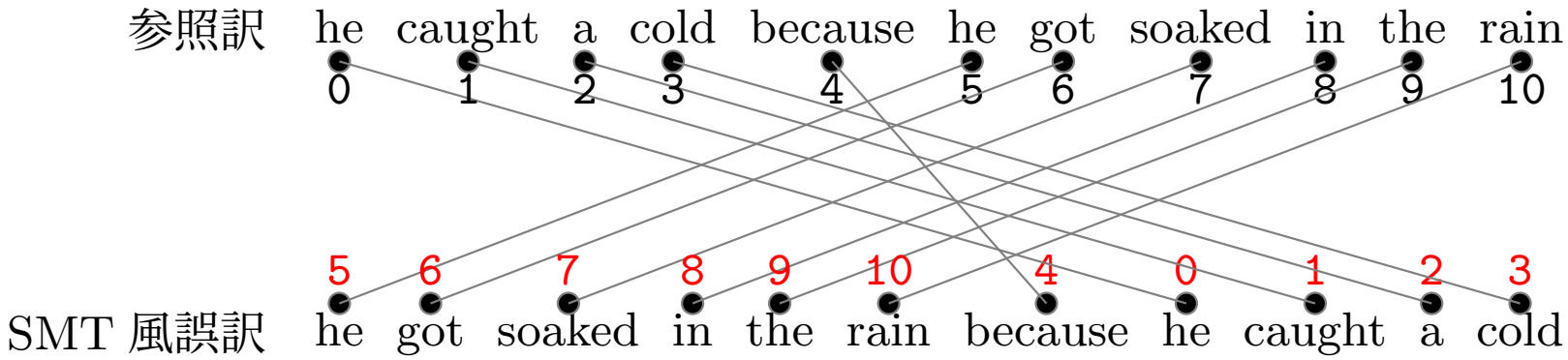


Figure 4: Distribution of τ

翻訳前の評価に使えるのなら、翻訳後の評価にも使えるんじゃないか？

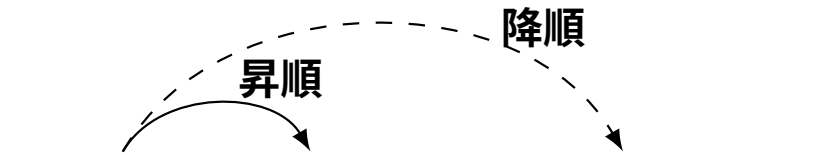
語順の順位相関を求めてみよう（SMT 風誤訳）

原文 彼は雨に濡れたので、風邪を引いた。



[5, 6, 7, 8, 9, 10, 4, 0, 1, 2, 3] の τ は？

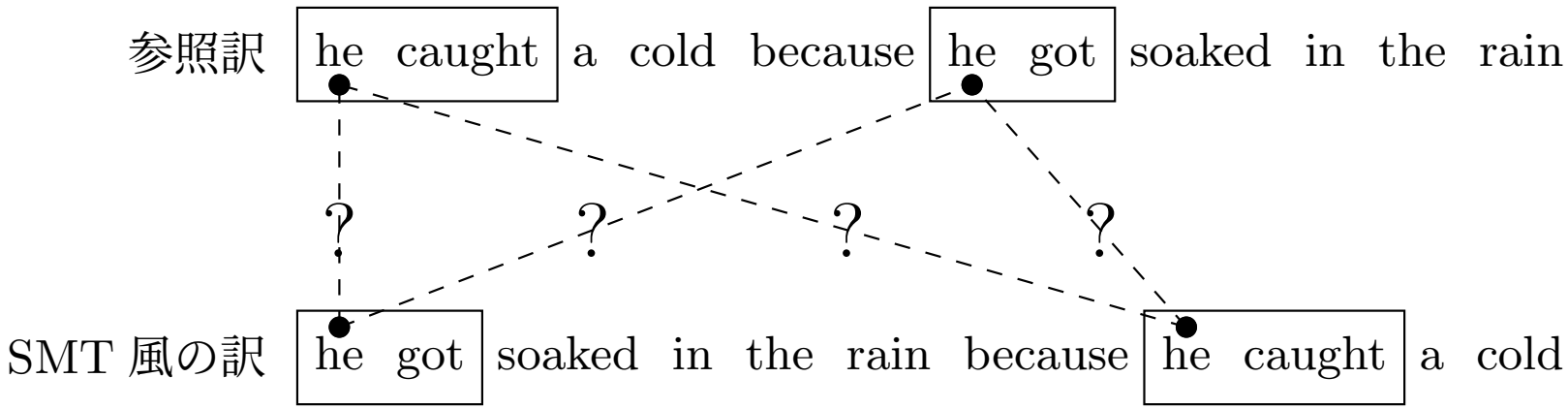
タイがない場合の計算式は、$\tau = \dfrac{\text{昇順ペア数}}{\text{全ペア数}} \times 2 - 1$



$$\tau = \frac{21}{{}_{11}C_2} \times 2 - 1 = -0.236$$

語順の順位相関を求めてみよう（SMT 風誤訳）

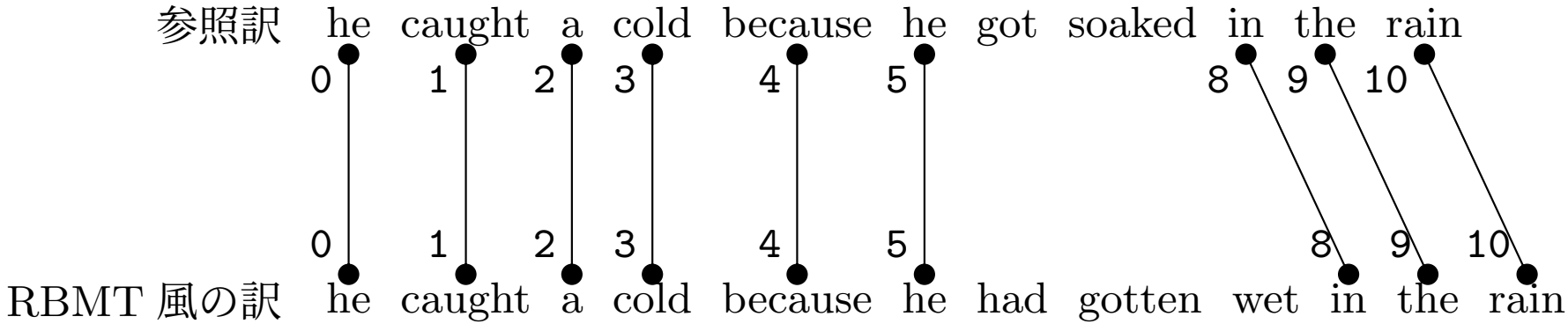
同じ単語が１文中に複数ある場合に、ユニグラムでは対応がつかない。



そこで論文の実験で用いたプログラムではバイグラムを見て判定。
公開中のプログラムは、トライグラム以上も見て判定しているそうです。

語順の順位相関を求めてみよう（RBMT 風直訳）

原文 彼は雨に濡れたので、風邪を引いた。



$[0, 1, 2, 3, 4, 5, 8, 9, 10]$ の τ は？
全ペアが昇順なので $\tau = 1.000$

対応しなかったところは無視していいの？

対応しなかったところは無視していいの？

got soaked と had gotten wet が対応しているか、意味がどれくらい違うか判定するには、**言い換え**の辞書などが必要。

NTCIR-7 の翻訳タスクの提出結果を見ると、RBMT も SMT も訳語選択の間違いは気にならない。

- SMT は語順が滅茶苦茶なので、何が何の誤訳なのかわからない。
- RBMT の訳語は参照訳と違うことが多いが、間違いは少なそう。

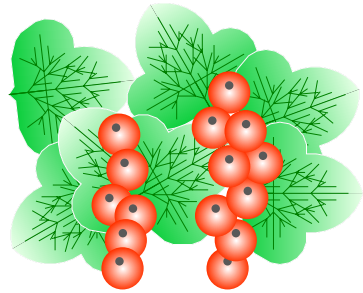
性善説に立ち、機械翻訳の訳語の違いは、気にしない。

ただし、τ が同じ場合にも、**対応した単語数の多寡**は考慮したい。

RIBES (NTT)

EMNLP-2010 で提案 磯崎他 D10-1092

2011 年の言語処理学会年次大会で命名 NLP11-D5-2



RIBES: Rank-based Intuitive Bilingual Evaluation Score

（ribes はカシスの仲間を表す属名）

$$\text{RIBES} = \frac{\tau + 1}{2} \times P^{\alpha} \quad (0 \leq \alpha \leq 1)$$

P はユニグラム適合率。α は NTCIR-7 JE のデータでチューニング。

τ の値域を $(\tau + 1)/2 \left(= \frac{\text{昇順ペア数}}{\text{全ペア数}} \right)$ により $[0, 1]$ に変換。

		妥当性	BLEU	RIBES
原文	**彼は雨に濡れたので、風邪を引いた。**			
参照訳	He caught a cold because he got soaked in the rain.			
RBMT	He caught a cold because he had gotten wet in the rain.	○	**0.53**	**0.93**
SMT	He got soaked in the rain because he caught a cold.	×	**0.74**	**0.38**

BLEU は人間の妥当性評価と逆!

NTCIR-7 日英特許翻訳データでの性能

RIBES のシステム・レベルのスコアは、文レベルのスコアの単純平均

人手評価とのシステム・レベルの順位相関（Spearman の ρ）

自動評価法	妥当性	流暢性
RIBES($\alpha = 0.2$)	**0.947**	0.879
ROUGE-L	0.903	**0.889**
IMPACT	0.826	0.751
METEOR	0.490	0.508
BLEU	0.515	0.500

NLP11-D5-2

翻訳自動評価の手法としてあまり使われていない Chin-Yew Lin の ROUGE-L P04-1077 もよい。

このような、評価法の評価を「**メタ評価**」という。

NTCIR-9 特許翻訳オーガナイザによるメタ評価

後藤他 NTCIR9-GotoI, NLP12-E2-9

NTCIR-9 では、RIBES が BLEU と並んで自動評価法として採用された

NTT からソースコードを公開

http://www.kecl.ntt.co.jp/icl/lirg/ribes/index-j.html

NTCIR-9 での人手評価（妥当性）との順位相関（Spearman の ρ）

自動評価法	日英	英日	中英
BLEU	▼ 0.042	▼ 0.029	0.931
NIST	▼ 0.114	▼ 0.074	0.911
RIBES	**0.632**	**0.716**	**0.949**

▼はマイナス

BLEU, NIST は中英では良いが、日英・英日では人手と相関なし。

RIBES はこれらよりはましだが、NTCIR-7 の時より悪い。

NTCIR-9 特許翻訳オーガナイザによるメタ評価

NTCIR-9 のメタ評価の結果には、意外なところがあります。

日本語は、語順が比較的自由だと言われています。

「私はごはんを食べます」⇔「ごはんを私は食べます」

このことは、語順を重視する RIBES にとって不利のようです。

ところが、日本語の語順を測定している英日翻訳の方が、英語の語順を測定している日英翻訳よりも、人間の評価との相関が高いのです。

どんな自動評価法が望ましいか？

磯崎の考える望ましさの基準。

- 人間の評価との相関が高い。
- 言語に依存したリソースが不要。
- 計算が軽い。
- 必要な参照訳が少ない。
- 翻訳のどこが悪いかがわかりやすい。
- SMT のパラメタ・チューニングに使える。

日英・英日翻訳では

BLEU: 人×リ○軽○参×悪×チ○
RIBES:人○リ○軽○参○悪○チ？

どんな自動評価法が望ましいか？

人間の評価に近づけようとすれば、言い換え辞書等に頼らざるを得ない。

しかし、辞書を作り、メンテナンスしていくのは大変。

TESLA 論文 WMT-2010 W10-1754 は、以下の 3 つに分けている。

Heavyweight Linguistic Approaches RTE P09-1034, ULC WMT-2008 W08-0332

Lightweight Linguistic Approaches METEOR W05-0909 , MaxSim P08-1007

Non-linguistic Approaches BLEU P02-1040, TER AMTA-2006-snover

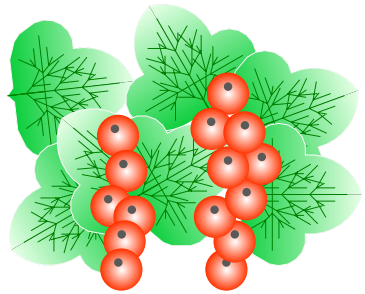
Dreyer&Marcu の考える望ましさの基準

N12-1017

- 直観的である。
- 自動で計算できる。
- 安定していて、再現可能である。
- 人間か自動翻訳か RBMT か SMT かハイブリッドかによらず、うまく動く。
- その評価法に合わせてチューニングができる。
- 翻訳エンジンのどこに問題があるのかを判定する手助けとなる。

彼らの HyTER は、上記をすべて満たすそうだ。

Outline



LRscore （イギリス・エジンバラ大学）

http://homepages.inf.ed.ac.uk/abmayne/

RIBES の論文をそろそろ投稿しようと思っていたら、Birch ら Machine Translation 2010 が、磯崎より少し前に中英翻訳評価で τ を利用していることが判明。

ただし、単語アラインメントは自分で計算して与えないといけない。

LRscore (Lexical Reordering score) はその改良

WMT-2010
W10-1749, D11-1079, P11-1103

BLEU と以下の量の補間

- K: BP つき Kendall の τ
- H: BP つきハミング距離

たとえば、LR-HB1 は、BP つきハミング距離と BLEU1 の補間
LR-KB4 は、BP つき Kendall の τ と BLEU4 の補間

FRS (Google)

Talbot 他 WMT-2011 W11-2102, Katz-Brown 他 D11-1017, 賀沢他 NLP11-D5-4

FRS: Fuzzy Reordering Score, $\mathrm{FRS} = 1 - \dfrac{\text{チャンク境界の数}}{\text{単語境界の数}}$

(cf. METEOR WMT-2011 W11-2107, W05-0909 の fragmentation penalty)

SMT のチューニング中なら、原文と翻訳の単語対応が完全にわかる。

原文	he caught a cold	because	he got soaked in the rain		
SMT	he got soaked in the rain	because	he caught a cold		

		BLEU	RIBES	FRS
原文	彼は雨に濡れたので、風邪を引いた。			
参照訳	He caught a cold because he got soaked in the rain.			
RBMT	He caught a cold because he had gotten wet in the rain.	0.53	0.93	1.00?
SMT	He got soaked in the rain because he caught a cold.	0.74	0.38	0.80?

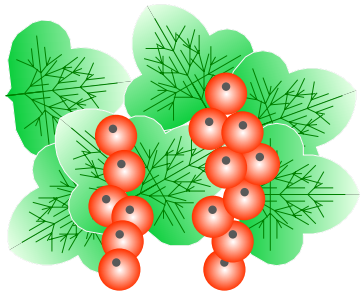
FRS (Google)

賀沢ら NLP11-D5-4 は、英日翻訳で**文レベルのメタ評価**を行っている。

Metric	Correlation
FRS	0.508
Tau	0.505
BLEU	0.409
FRS+Tau	0.546
FRS+BLEU	0.560
(FRS+Tau)/2 + BLEU	0.588

Table 2: Sentence-level correlation of evaluation metrics to human judgement score

Outline



翻訳自動評価に関する外国での研究

BLEU に対する疑問は、**外国でも**昔からある。

統計的機械翻訳ワークショップ (Workshop on Statistical Machine Translation, WMT) では、欧⇔英翻訳で様々な翻訳自動評価法を比較。

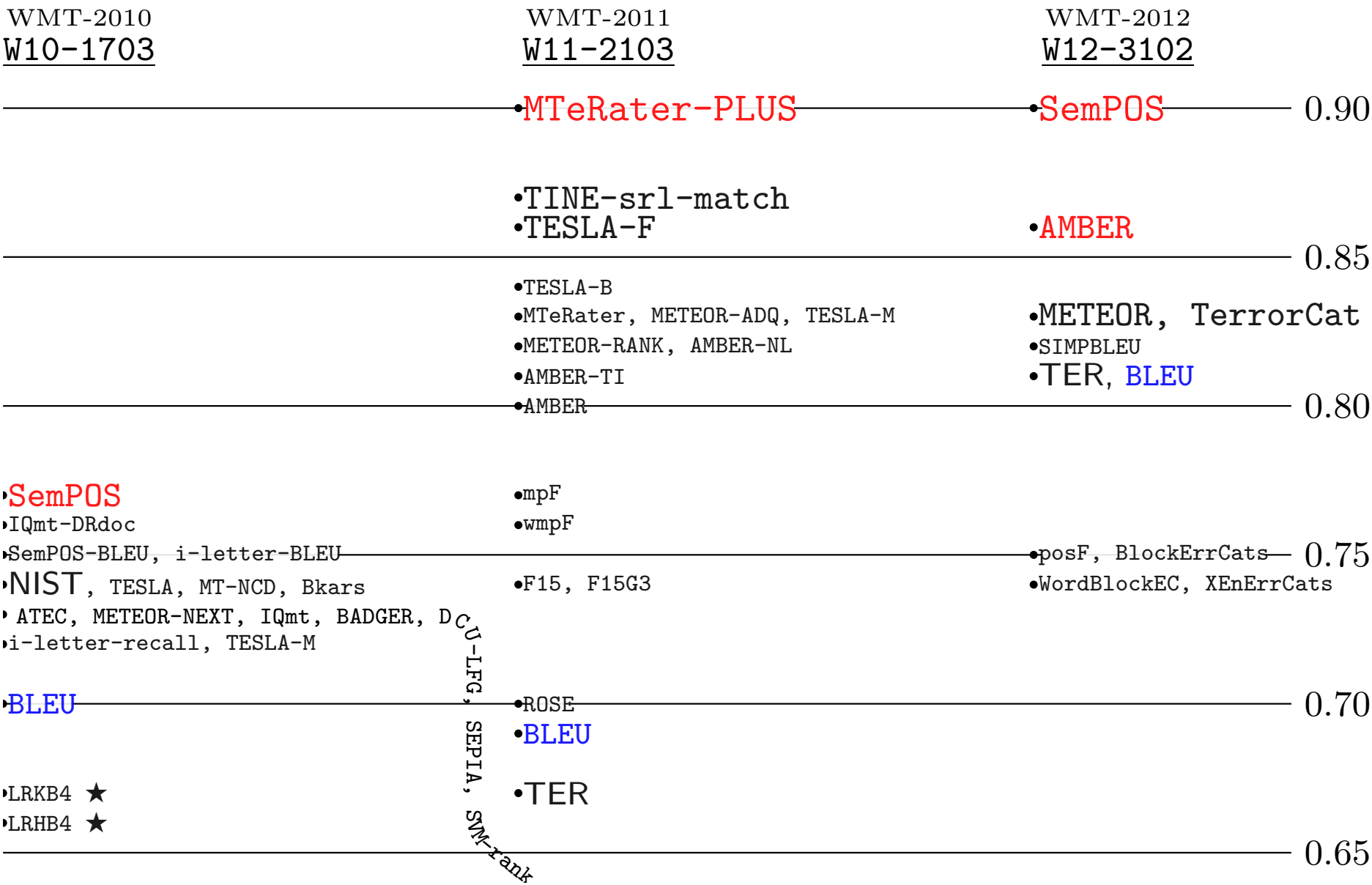
英語以外のヨーロッパ言語としては、ドイツ語、フランス語、スペイン語、チェコ語が使われる。

（チェコ語はしばしばデータ不足などで、表からはずされる。）

以下では 2010 年以降の WMT で上位だった手法を紹介する。

WMT10〜12 翻訳自動評価タスクの推移（欧英翻訳）

人手評価とのシステムレベルの順位相関（Spearman）



SemPOS (チェコ・プラハ・カレル大学)

Macháček and Bojar PBML-2009, P10-2016, W10-1705 (WMT-2010), W11-2108 (WMT-2011)

SemPOS = Semantic Part-of-Speech に基づき、**内容語の重なり**で評価

ただし、単語を t-lemma と呼ばれる標準形に変換して比較

t-lemma は、深い構文解析が必要で重いので、浅い処理で近似

MTeRater-Plus （コロンビア大学, ETS）

（ETS は TOEFL などの英語のテストをしている機関）

WMT-2011
W11-2111

MTeRater 人間が書いたエッセイの自動採点用に考案された ETS e-rator® の素性を使って機械学習 **参照訳不要**
grammar, usage, mechanics, style, organization, development, lexical complexity, vocabulary usage の誤りを採点。

- grammar: sentence fragments, verb form errors, pronoun errors
- usage: articles, prepositions, collocations
- mechanics: spelling, punctuation, capitalization

MTeRater-Plus MTeRater の素性に、BLEU, TERp, METEOR などの素性を加えたもの。

AMBER （カナダ国立研究機構）

Chen 他 WMT-2011 W11-2105, WMT-2012 W12-3104
AMBER: A Modified BLEU, Enhanced Ranking Metric

n-gram の適合率・再現率の関数で定義されたスコアに、
いろんなペナルティを掛けたもの。

Strict Brevity Penalty (SBP) Chiang D08-1064
オリジナルの Brevity Penalty は、個々の訳文ではなく、全訳文の長さの和しか問題にしていない。そのせいで、長すぎる訳文が短かすぎる訳文を助けてしまうことがある。これを禁止。

Strict Redundancy Penalty (SRP) 長すぎる訳文へのペナルティ。

Chunk Penalty (CKP) 訳文が途切れることに対するペナルティ。
METEOR の fragmentation penalty と同じ

正規化した順位相関係数も導入している。

Normalized Spearman's Correlation Penalty (NSCP)

Normalized Kendall's Correlation Penalty (NKCP)

たくさんパラメタがあるので、downhill simplex 法でチューニング。

Philipp Koehn の SMT 本で Simplex 法と呼ばれているものだが、線形計画法ではない。

SPEDE (スタンフォード大学)

Wang & Manning D12-1090

セグメントレベルの自動評価でベスト（だが、$\tau = 0.25$ は低すぎる）

重みつき編集距離を計算するのに確率的有限状態機械 (pFSM) を使う。

最近の評価手法は、線形回帰や SVR で人間の評価を真似

(SVR: SVM による回帰)

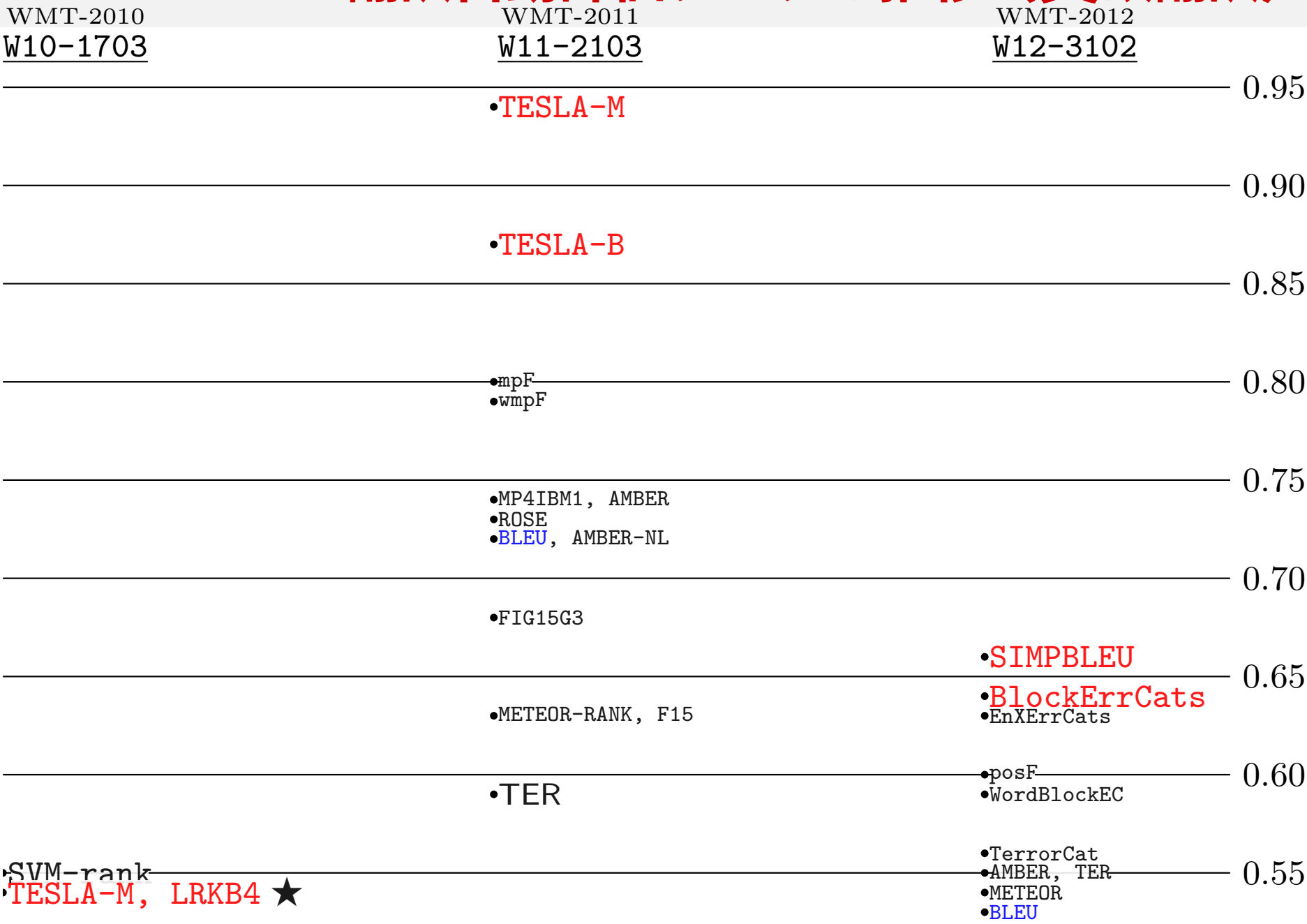
Padó ら（P09-1034）の RTER (← RTE + Regression)

しかし、**素性を抽出するのに時間**がかかり、使いづらい。

RTER に比べてずっと軽いことを確認した。

pPDA (probabilistic Pushdown Automaton) に拡張。

WMT10〜12 翻訳自動評価タスクの推移（英欧翻訳）



なぜ、英語以外での自動評価は悪いのか？

個別の言語による事情が十分考慮できていない。

- 複数の言語でうまくいく方法を見つけるのが難しい
- 研究者の知識不足
- ツールの精度不足
- 言語リソースの不足

TESLA （NUS:シンガポール国立大学）

Liu 他: D11-1035 (WMT-2011), W11-2106 (WMT-2011), W10-1754 (WMT-2010) http://nlp.comp.nus.edu.sg/software

TESLA: Translation Evaluation of Sentences with Linear-programming-based Analysis

WMT11 Evaluation task で、人間の評価との順位相関 ρ が、英独、英西、英仏の平均でベスト

WordNet の同義語を利用

機械訳と参照訳の n-gram のマッチングにおいて、n-gram に重みを与えられるように、**n-gram のマッチングを線形計画法で解く**のがポイント

TESLA （NUS:シンガポール国立大学）

Liu 他: D11-1035, W11-2106 (WMT-2011), W10-1754 (WMT-2010)

n-gram の類似度は、マッチングの良さを表す。

しかし、n-gram の重要性は異なるので、n-gram の重みを考える。

この重みを線形計画法で解く。

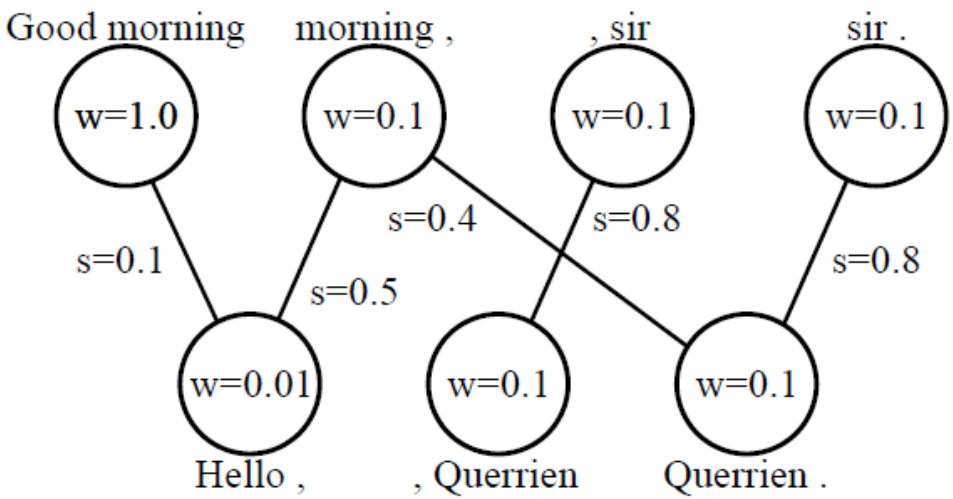
目的関数は、類似度の加重和で、最大化する。

各 n-gram につながる枝の重みの総和には上限を与える（たとえば 1.0）

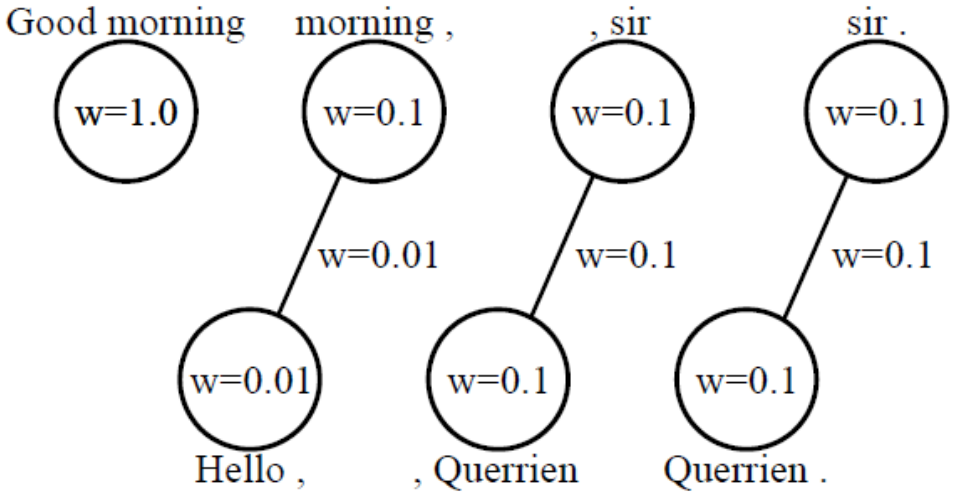
TESLA-M は、n ユニグラムの F 値の平均

TESLA(-F) は SVM-rank でトレーニング

TESLA （NUS:シンガポール国立大学）



(a) The matching problem



(b) The solution

「Good morning , sir .」と
「Hello , Querrien .」のマッチング

ノードは n グラムで、重みは出現頻度。

含まれる機能語ごとに 0.1 倍。

エッジの重みは n グラムの類似度。

基本形・語義・品詞の一致で計算。

SIMPBLEU （シェフィールド大学）

WMT12 のオーガナイザの論文 WMT-2010 W12-3102 は、SIMPBLEU のポインタとして、ROSE の論文 (Song and Cohn WMT-2011 W11-2113) を引用。

しかし、著者本人に確認したところ、ROSE と SIMPBLEU は別のもので、SIMPBLEU の論文はまだ出版されていない。

TerrorCat (チューリッヒ大学他), BlockErrCats, etc.

(DFKI=ドイツ人工知能研究所)

Fishel 他 WMT-2012 W12-3105, Popović WMT-2012 W12-3106

翻訳誤りを分類し、誤りのカテゴリごとに、翻訳品質に与える影響の深刻さを調べておく。

同じ誤りでも、言語によって深刻さは違う。

- 英語では語順の誤りが重要
- チェコ語やドイツ語では、屈折の誤りが重要

Zeman の Addictor PBML-2011 や Popovic の Hjerson J11-4002 を使って自動的に誤りを解析する

WMT11 Tunable Metrics Task

WMT-2011
W11-2103

目的：**MT のチューニングに使える**評価法を作ろう (invitation-only, Urdu-English)
評価関数を取り換えやすい Z-MERT PBML-2009 を利用
結果：BLEU より明らかによいものはなし

Tunable Metrics Task
1324–1484 comparisons/system

System	≥**others**	>others
BLEU •	**0.79**	0.28
BLEU-SINGLE •	**0.77**	0.27
CMU-METEOR •	**0.76**	0.27
RWTH-CDER	**0.76**	0.26
CU-SEMPOS-BLEU •	**0.74**	0.29
STANFORD-DCP •	**0.73**	0.27
NUS-TESLA-F	**0.68**	0.28
SHEFFIELD-ROSE	**0.05**	0.00

- indicates a **win**: no other system combination is statistically significantly better at p-level≤0.10 in pairwise comparison.

WMT12 Quality Estimation Task

WMT-2012
W12-3102

目的：**参照訳を用いずに**、機械学習技術により翻訳の品質を推定したい

英語からスペイン語への翻訳結果を利用

とくに良かったのは

SDLLW_MP5bestDeltaAvg, SDLLW_SVM 米国・SDL Language Weaver Soricut 他 WMT-2012 W12-3118

UU_bltk スウェーデン・ウプサラ大学 WMT-2012 W12-3112
Tree kernel を使用

SDLLW_M5P..., SDLLW_SVM (SDL Language Weaver)

Soricut 他 WMT-2012 W12-3118

後処理の大変さを予測できるように、M5P（Quinlan の M5 もどき）や SVM で学習。

以下の 3 種類の素性を利用

ベースライン素性 タスクで定義されている 17 素性。
原言語・目的言語でのトークン数、言語モデルの確率、低頻度・高頻度の n-gram の割合、句読点の数など。

デコーダ素性 Moses デコーダの内部コスト (distortion cost, word penalty cost, language-model cost, etc.)

独自素性 原言語側の未知語数、言語モデルのパープレキシティ、BLEU4/BP、１対１単語アラインメントの数・割合など。

UU_bltk スウェーデン・ウプサラ大学

WMT-2012
W12-3112

「参照訳あり」で tree kernel (TK) を使うのはもっともな話。
ここでは「**参照訳なし**」の翻訳品質の推定に TK を入れた SVR を利用。
いずれにせよ、いかにも計算時間がかかりそう。

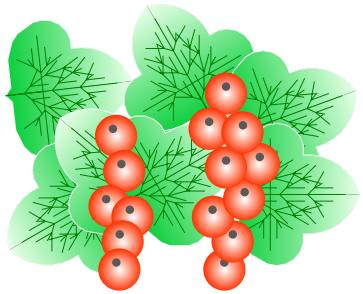
					Cross-validation				*Test set*			
	Features	T	C	d	Δ	ρ	MAE	RMS	Δ	ρ	MAE	RMS
UU_best	99 explicit + TK	0.05	4	2	0.506	0.566	0.550	0.692	0.56	0.62	0.64	0.79
(a)	99 explicit + TK	0.03	8	3	0.502	0.564	0.552	0.700	0.56	0.61	0.63	0.78
(b)	17 explicit + TK	0.05	4	2	0.462	0.530	0.568	0.714	0.57	0.61	0.65	0.79
UU_bltk	17 explicit + TK	0.03	8	3	0.466	0.534	0.566	0.712	0.58	0.61	0.64	0.79
(c)	99 explicit	0	8	2	0.492	0.560	0.554	0.700	0.56	0.59	0.65	0.80
(d)	17 explicit	0	8	2	0.422	0.466	0.598	0.748	0.52	0.55	0.70	0.83
(e)	TK only	–	4	–	0.364	0.392	0.632	0.782	0.51	0.51	0.70	0.85

T: Tree kernel weight　C: Training error/margin trade-off　d: Degree of polynomial kernel
Δ: DeltaAvg score　ρ: Spearman rank correlation　MAE: Mean Average Error
RMS: Root Mean Square Error　TK: Tree kernels

Table 1: Experimental results

TK 単独はよくない。他の素性が基本で、少しだけ TK を混ぜている。

Outline



PORT （カナダ国立研究機構）

Chen 他 P12-1098 （AMBER と同じグループ）

PORT: Precision-Order-Recall Tunable metric

PORT はリソースがいらず、計算が速く、人間の評価との相関もよいだけでなく、**チューニングに適している。**

PORT は Q_{mean} と **新しい順位相関係数 ν** の α 乗の調和平均

Q_{mean} は、n グラム適合率の平均 P_a と n グラム再現率の平均 R_a にペナルティをかけ二乗平均

$$Q_{\text{mean}} = \sqrt{\left((P_a \times \text{SBP})^2 + (R_a \times \text{SRP})^2\right)/2}$$

SBP: Strict Brevity Penalty
SRP: Strict Redundancy Penalty

Chen 他 P12-1098 どうして新しい順位相関係数を作ったのか？

Spearman の ρ は、長距離移動のペナルティが大きすぎる。

磯崎も τ と ρ の比較実験で、ρ の方がきついことには同意

ρ の計算式は、移動距離の 2 乗和を正規化する形。

$$\rho = 1 - \frac{\sum_{i=1}^{n} d_i^2}{{}_{n+1}C_3}$$

ならば、2 乗しなければいい、という発想らしい。

$$\nu_1 = 1 - \frac{\sum_{i=1}^{n} |d_i|}{{}_{n+1}C_2}$$

これだけではだめだったのか、ν は ν_2 という別の係数との調和平均。

PORT （カナダ国立研究機構）

Chen 他 P12-1098 チューニングの結果

PORT が BLEU に勝っている。

	PORT win	BLEU win	equal	total
zh-en	**19**	18	12	49
small	**38.8%**	36.7%	24.5%	
zh-en	**69**	46	36	151
large	**45.7%**	30.5%	23.8%	
fr-en	14	**17**	12	43
Hans	32.6%	**39.5%**	27.9%	
de-en	**34**	17	6	57
WMT	**59.7%**	29.8%	10.5%	
All	**136**	98	66	300
	45.3%	32.7%	22.0%	

Table 5: Human preference for outputs from PORT-tuned *vs*. BLEU-tuned system.

PORT （カナダ国立研究機構）

Chen 他 P12-1098 **新しい順位相関係数 ν の効果**はあったのか？

Tune	BLEU	METEOR	1-TER
BLEU	25.1	53.7	36.4
PORT(ν)	**25.3**	**54.4**	**37.0**
PORT(ρ)	25.1	54.2	36.3
PORT(τ)	25.1	54.0	36.0

Table 9: Comparison of the ordering measure: replacing ν with ρ or τ in PORT.

Task	Tune	ordering measures		
		ρ	τ	ν
NIST06	BLEU	0.979	0.926	0.915
	PORT	0.979	**0.928**	**0.917**
NIST08	BLEU	0.980	0.926	0.916
	PORT	**0.981**	**0.929**	**0.918**
CTB	BLEU	0.973	0.860	0.847
	PORT	**0.975**	**0.866**	**0.853**

Table 10: Ordering scores (ρ, τ and ν) for test sets NIST 2006, 2008 and CTB.

微妙

AM-FM （シンガポール科学技術研究庁インフォコム研究所）

Banchs&Li P11-2027 **参照文が不要**

AM Adequacy-oriented component of the metric

Dumais らの **CL-LSI** (Cross-Language Latent Semantic Indexing) を利用し、**原文と訳文のコサイン類似度**を求める。

FM Fluency-oriented component of the metric

n-gram 言語モデルによる確率

AM-FM AM と FM の重みつき調和平均 $\text{AM-FM} = \dfrac{\text{AM} \times \text{FM}}{\alpha\text{AM} + (1-\alpha)\text{FM}}$

Metric	Adequacy	Fluency	H Mean
BLEU	**0.4232**	**0.4670**	**0.4516**
NIST	0.3178	0.3490	0.3396
Meteor	0.4048	0.3920	0.4065
AM-FM	0.3719	0.4558	0.4170

Table 2: Pearson's correlation coefficients (computed at the system level) between automatic metrics and human-generated scores

WMT07 の欧⇔英翻訳

参照訳なしの割に健闘

HyTER (米・SDL Language Weaver)

Dreyer&Marcu: N12-1017

同義表現をボトムアップにアノーテーションできるツールを作成

これにより、等価な表現の組み合わせによってできる**膨大な参照訳の集合**を表すことができる

HyTER は、このネットワークによって HTER AMTA-2006-Snover を真似る。

HTER とは、機械訳と参照訳を**人間が**見て作った新しい参照訳を基準にして計算した TER (Translation Edit Rate)

距離計算は、有限状態オートマトン (FSA) により効率的に計算できる。

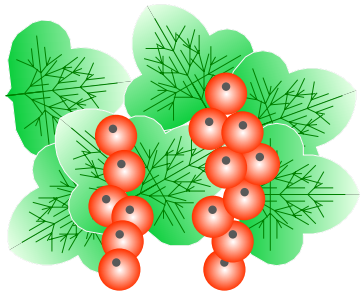
HyTER (米・SDL Language Weaver)

Dreyer&Marcu: N12-1017
HTER を基準としたメタ評価の結果

Arabic-English									
Size	Likert	Meteor 1	Meteor 4	BLEU 1	BLEU 4	TERp 1	TERp 4	HyTER U (r5)	HyTER SPU (r5)
1	.653	.529	.541	.512	.675	.452	.547	.643 (.661)	.647 (.655)
2	.645	.614	.636	.544	.706	.599	.649	.733 (.741)	.735 (.732)
4	.739	.782	.804	.710	.803	.782	.803	.827 (.840)	.831 (.838)
8	.741	.809	.822	.757	.818	.796	.833	.827 (.828)	.830 (.825)
16	.868	.840	.885	.815	.887	.824	.862	.888 (.890)	.893 (.894)
32	.938	.945	.957	.920	.948	.930	.947	.938 (.935)	.940 (.936)
64	.970	.973	.979	.964	.973	.966	.968	.964 (.960)	.966 (.961)
Chinese-English									
Size	Likert	Meteor 1	Meteor 4	BLEU 1	BLEU 4	TERp 1	TERp 4	HyTER U (r5)	HyTER SPU (r5)
1	.713	.495	.557	.464	.608	.569	.594	.708 (.721)	.668 (.681)
2	.706	.623	.673	.569	.655	.639	.651	.713 (.716)	.702 (.701)
4	.800	.628	.750	.593	.734	.651	.726	.822 (.825)	.820 (.814)
8	.810	.745	.778	.783	.808	.754	.754	.852 (.856)	.854 (.845)
16	.881	.821	.887	.811	.884	.826	.844	.912 (.914)	.914 (.908)
32	.915	.873	.918	.911	.930	.851	.911	.943 (.942)	.941 (.937)
64	.950	.971	.976	.979	.973	.952	.970	.962 (.958)	.958 (.957)

Table 3: Document-level correlations of various scores to HTER. Meteor, BLEU and TERp are shown with 1 and 4 references each, HyTER is shown with the two combination methods (U and SPU), and with reordering (r5).

Outline



意味論に基づく翻訳評価 (スペイン・カタルーニャ工科大)

Giménez&Màrquez WMT-2007 W07-0738, I08-1042, WMT-2008 W08-0332, WMT-2009 W09-0440

2010 年より前だが、意味論に基づく評価としてしばしば参照されるので、参考のため入れておく。

文を以下のような LE = Linguistic Elements の bag とみなす。

語形、品詞タグ、係り受け関係、構文的な句、固有表現、意味役割

DRS = Discource Representation Theory （談話表示理論）を利用

MEANT （香港科技大学）

Lo&Wu P11-1023, L10-1521, SSST-2011 W11-1002, SSST-2012 W12-4206

意味役割付与 (SRL) のマッチングによる採点

トレーニング・コーパスを使わない。

文レベルでの人手評価との相関がよい。

正しく翻訳された意味フレームの再現率・適合率から、F 値を計算

部分的に正しく翻訳された意味フレームにも少し点を与えるために、式が複雑になっている。

Sagan （アルゼンチン・コルドバ国立大学）

Castillo&Estrella WMT-2012 W12-3103

テキスト含意認識 (Recognizing Textural Entailment, RTE) 技術を利用

RTE とは？

Text (T) In 1963 Lee Harvey Oswald murdered JFK.

Hypothesis (H) JFK died in 1963.

のようなものが与えられて、(T) が (H) を含意するかどうか判断する課題。

意味的テキスト類似度 (STS = Semantic Textual Similarity) に基づく

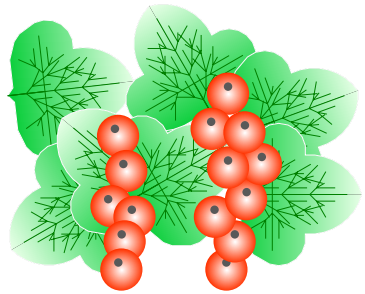
翻訳以外の文章の自動評価タスク・ワークショップ

TAC AESOP task (Automatically Evaluating Summaries of Peers) 要約の自動評価 2009, 2010, 2011, 2012

HOO pilot shared task (Helping Our Own) 英語を母国語としない人のテキストの誤りの自動修正 2011, 2012

ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization 2005

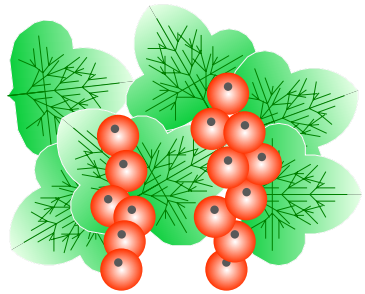
Outline



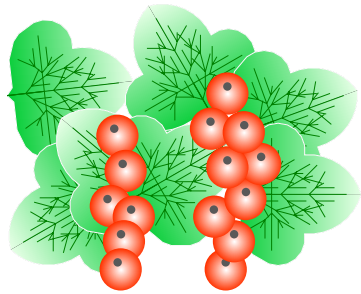
まとめ

- 翻訳自動評価は、翻訳品質の向上に不可欠である。
- 細かい訳語の違いに敏感な BLEU や NIST は、英日・日英翻訳では、人間による評価と無相関。
- 現在の英日・日英翻訳の SMT 出力は、語順がめちゃくちゃで、個々の訳語の良し悪しが目立たない。
- RIBES は、訳語の違いを重視せず、語順を重視することで、人間の自動評価との相関を高くしている。
- 新しい自動評価法の中には、順位相関係数を取り込んだものもある。
- 「**参照訳が不要な**」あるいは「**パラメタ・チューニングに使える**」自動評価法の模索が始まっている。

Outline



RIBESに関する参考資料



RIBES に関する文献（査読あり国際会議 その 1）

国際的には EMNLP-2010 で発表したが、当時は RIBES という名前がついていなかった。その前の WMT-2010 で、Head Finalization による preordering の効果を Kendall の τ で評価している。

W10-1736 **Isozaki, Sudoh, Tsukada, Duh: Head Finalization: A Simple Reordering Rule for SOV Languages, WMT-2010, pp.244-251.**

D10-1092 **Isozaki, Hirao, Duh, Sudoh, Tsukada: Automatic Evaluation of Translation Quality for Distant Language Pairs, EMNLP-2010, pp.944-052.**

— 以下は D10-1092 を引用している文献 —

I11-1004 Wu, Sudoh, Duh, Tsuakada, Nagata: Extracting Pre-ordering Rules from Predicate-Argument Structures, IJCNLP-2011, pp.29–37.

I11-1153 Duh, Sudoh, Wu, Tsukada, Nagata: Generalized Minimum Bayes Risk System Combination, IJCNLP-2011, pp.1356–1360.

MT Summit XIII Sudoh, Wu, Duh, Tsukada, Nagata: *Post*-ordering in Statistical Machine Translation, MT Summit-2011, pp.316–323.

D11-1017 Katz-Brown, Petrov, McDonald, Och, Talbot, Ichikawa, Seno, Kazawa: Training a Parser for Machine Translation Reordering, EMNLP-2011, pp.183–192.

RIBES に関する文献（査読あり国際会議 その 2）

W11-2102 Talbot, Kazawa, Ichikawa, Katz-Brown, Seno, Och: A Lightweight Evaluation Framework for Machine Translation Reordering, Workshop on Statistical Machine Translation, pp.12–21.

W11-2105 Chen, Kuhn: AMBER: A Modified BLEU Enhanced Ranking Metric, WMT-2011, pp.71–77.

D12-1077 Neubig, Watanabe, Mori: Inducing a Discriminative Parser to Optimize Machine Translation Reordering, EMNLP-2012, pp.843–853.

P12-1001 Duh, Sudoh, Wu, Tsukada, Nagata: Learning to Translate with Multiple Objectives, ACL-2012, pp.1–10.

P12-1098 PORT: a Precision-Order-Recall MT Evaluation Metric for Tuning, ACL-2012, pp.930–939.

P12-2061 Goto, Utiyama, Sumita: Post-ordering by Parsing for Japanese-English Statistical Machine Translation, ACL-2012, pp.311-316.

P12-3022 Wu, Matsuzaki, Tsujii: Akamon: An Open Source toolkit for Tree/Forest-Based Statistical Machine Translation, ACL-2012, pp.127–132.

W12-3104 Chen, Kuhn, Foster: Improving AMBER, an MT Evaluation Metric, WMT-2012, pp.59–63.

W12-4207 Dan, Sudoh, Wu, Duh, Tsukada, Nagata: Head Finalization Reordering for Chinese-to-Japanese Machine Translation. Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6), pp.57–66, 2012.

RIBES に関する文献（言語処理学会年次大会）

日本語では、2011 年の年次大会で RIBES という名前で発表した。前年の年次大会で、Head Finalization による preordering の評価を Kendall の τ で行っている。

NLP10-B4-2 **磯崎: 英日翻訳における語順について, 言語処理学会年次大会, 2010 年**

NLP11-D5-2 **平尾・磯崎・Duh・須藤・塚田・永田: RIBES:順位相関に基づく翻訳の自動評価法、言語処理学会年次大会, 2011 年.（優秀発表賞）**

— 以下は引用 —

NLP11-D5-4 賀沢・Talbot・妹尾・Katz-Brown・市川・Och: 英日機械翻訳における語順評価の有効性、言語処理学会年次大会, 2012 年

NLP12-E1-1 星野・宮尾:日本語単語分割が統計的機械翻訳に与える影響の評価、言語処理学会年次大会, 2012 年

NLP12-E1-7 呉・須藤・Duh・永田:An Improvement to the Predicate-Argument Structure Based Pre-ordering for Statistical Machine Translation、言語処理学会年次大会, 2012 年

NLP12-E2-8 松本・村上・徳久:機械翻訳における人手評価と自動評価の考察、言語処理学会年次大会, 2012 年

NLP12-E2-9 後藤・Lu・Chow・隅田・Tsou:特許翻訳における機械翻訳システムの評価—NTCIR-9 特許機械翻訳タスクでの分析—, 言語処理学会年次大会, 2012 年

NLP12-P3-18 韓・須藤・呉・Duh・塚田・永田: Syntactic Based Reordering Rules for Chinese-to-Japanese Machine Translation 言語処理学会年次大会, 2012 年.

RIBES に関する文献（NTCIR）

NTCIR-9 では、評価方法として BLEU と並び RIBES が採用されたため、とくに英日・日英翻訳の論文で利用・引用されている。

NTCIR9-GotoI Goto, Lu, Chow, Sumita, Tsou: Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop, pp.559–578, 2012.

NTCIR9-SudohK Sudoh, Duh, Tsukada, Nagata, Wu, Matsuzaki, Tsujii: NTT-UT Statistical Machine Translation in NTCIR-9 PatentMT, pp.585–592, 2012.

NTCIR9-LeeY Lee, Xiang, Zhao, Franz, Roukos, Al-Onaizan: IBM Chinese-to-English PatentMT System for NTCIR-9, pp.606–613, 2012.

NTCIR9-OhioT Oshio, Mitsuhashi, Kakita: Use of the Japio Technical Field Dictionaries for NTCIR-PatentMT, pp.614–617, 2012.

NTCIR9-EharaT Ehara:Machine translation system for patent documents combining rule-based translation and statistical post-editing applied to the PatentMT Task, pp.623–628, 2012.

NTCIR9-KondoS Kondo, Komachi, Matsumoto, Sudoh, Duh, Tsukada:Learning of Linear Ordering Problems and its Application to J-E Patent Translation in NTCIR-9 PatentMT, pp.641–645, 2012.

NTCIR9-MurakamiJ Murakami, Tokuhisa: Statistical Machine Translation with Rule based Machine Translation, pp.646–651, 2012.

NTCIR9-NaH Na, Li, Kim, Lee:POSTECH's Statistical Machine Translation Systems for NTCIR-9 PatentMT Task (English-to-Japanese), pp.652–656, 2012.

NTCIR9-WuX Wu, Matsuzaki, Tsujii: SMT Systems in the University of Tokyo for NTCIR-9 PatentMT, pp.666-672, 2012.