疑似参照訳文ベクトルの重心に基づく高速なニューラル最小ベイズリスク復号

出口 祥之　坂井 優介　上垣外 英剛　渡辺 太郎
奈良先端科学技術大学院大学
{deguchi.hiroyuki.db0, sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

概要

翻訳品質の人手評価スコアを予測する COMET モデルを用いた最小ベイズリスク復号法の COMET-MBR が，従来のビーム探索等を用いた最大事後確率復号と比較して，翻訳品質の人手評価における最高性能を達成している．しかし，典型的な最小ベイズリスク復号では，ある仮説文のベイズリスクは全仮説文との間のスコアの期待値計算によって求められるため，翻訳仮説の文数を N とすると $\mathcal{O}(N^2)$ の計算時間を要する．本研究では，COMET-MBR の復号速度を改善するため，COMET モデルが計算する翻訳候補集合の文ベクトルをクラスタリングし，各クラスタの重心表現を用いてスコアを計算することで，期待値計算を近似する．WMT'22 翻訳タスクの英 ↔ 日，英 ↔ 独，英 ↔ 中の 6 言語方向の翻訳実験を行ったところ，従来の MBR 復号より，同等の COMET スコアを維持しながら，ベイズリスクの計算時間が 13.6 倍，うち期待値計算が 26.2 倍高速化することを確認した．

1　はじめに

最小ベイズリスク（minimum Bayes risk; MBR）復号による機械翻訳は，複数文からなる翻訳仮説集合全体から計算されるベイズリスクを最小化する出力文を選択することで，頑健かつ品質の高い翻訳を実現する [14, 5, 16]．特に近年では，翻訳品質の人手評価スコアを予測する COMET モデル [19, 18] を MBR で用いた COMET-MBR が，ビーム探索を用いた従来の最大事後確率復号と比較して，翻訳品質の人手評価における最高性能を達成している [8]．COMET モデルは原文，仮説文，参照訳文をそれぞれ独立に文ベクトルに符号化し，それら 3 つの文ベクトルを用いてスコアを計算する．

しかし，MBR 復号では，ある仮説文のベイズリスクが各参照訳との間で計算されるスコアの期待値計算によって求められ，仮説文集合自身を擬似的に参照訳集合とみなす典型的な MBR 復号は，翻訳仮説の数を N とすると $\mathcal{O}(N^2)$ の計算時間を要する．N は多い時で 1000 を超えることもあり [9]，二乗オーダーの計算時間は MBR 復号の大きな課題となっている．これまでに，MBR の復号速度を改善するため，仮説文を枝刈りする手法が提案されており，軽量な代替評価尺度 [6] や N-best リランカ [8] を用いる手法などが提案されている．ただし，代替評価尺度の種類を適切に選択する必要が生じたり，リランカモデルの訓練・推論コスト等が生じたりする．

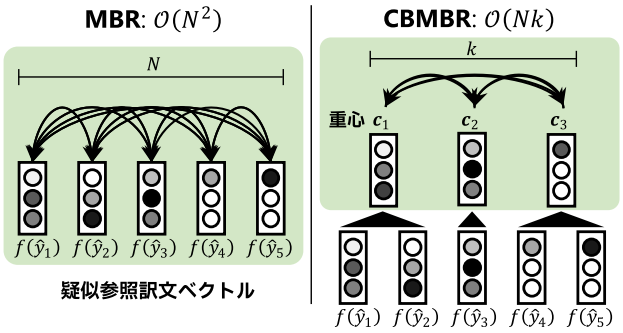

図 1　疑似参照訳文数 $N = 5$，クラスタ数 $k = 3$ における，従来法（MBR）と提案法（CBMBR）の概要図．緑色の背景箇所がベイズリスクの期待値計算を表す．

ここで，COMET モデルは仮説文と参照訳文が意味的に近いときに高いスコアを出力するように学習されていることから，文ベクトル間の距離が文間の類似度を表現できるのではないかと考えた．本研究では，COMET-MBR において，翻訳候補集合中の近い文ベクトル同士を 1 つにまとめてスコアを算出することで MBR の期待値計算を近似し，復号速度を改善する．提案法では，翻訳候補の文ベクトルを $k \ll N$ 個のクラスタにクラスタリングし，N 個の文ベクトルの代わりに各クラスタの k 個の重心ベクトルを用いて COMET スコアを算出することで，復号

時のスコア計算回数を N^2 回から Nk 回に削減する．

WMT'22 翻訳タスクの英 ↔ 日，英 ↔ 独，英 ↔ 中の 6 言語方向の翻訳実験より，従来の MBR 復号と比較して，同等の COMET スコアを維持しながら，ベイズリスクの計算時間が 13.6 倍，うち期待値計算が 26.2 倍速くなることを確認した．

2 背景: COMET-MBR

最小ベイズリスク復号 とりうる入力文と出力文の全体集合をそれぞれ $\mathcal{X}, \mathcal{Y}$ とする．このとき，最大事後確率（maximum-a-posteriori; MAP）復号による出力文 $y^*_{\text{MAP}} \in \mathcal{Y}$ は次のように求められる:

$$y^*_{\text{MAP}} = \operatorname*{argmax}_{y \in \mathcal{Y}} p_\theta(y|x). \tag{1}$$

なお，θ は入力文 $x \in \mathcal{X}$ が与えられたときに y が出力される尤度を計算する翻訳モデルを表す．とりうる全ての $y \in \mathcal{Y}$ について確率を計算することは困難であるため，通常，ビーム探索などによって枝刈りを行い，近似解を求める．一方，最小ベイズリスク（minimum Bayes risk; MBR）復号による出力文 $y^*_{\text{MBR}} \in \mathcal{Y}$ は，次のように効用関数 $u : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ の期待値（期待効用）を最大化する仮説が選択される:

$$y^*_{\text{MBR}} = \operatorname*{argmax}_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{\hat{y} \sim P(y|x)} [u(h, \hat{y})] \tag{2}$$

$$\approx \operatorname*{argmax}_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{\hat{y} \in \hat{\mathcal{Y}}} [u(h, \hat{y})]\,. \tag{3}$$

ただし，$\mathcal{H} = \{h_1, \ldots, h_{|\mathcal{H}|}\} \subset \mathcal{Y}$ は翻訳仮説集合，$P(y|x)$ は入力文 $x \in \mathcal{X}$ が与えられたときに y に翻訳される確率であり，一般には，式 3 のようにサンプリングされた参照訳 $\hat{\mathcal{Y}} = \{\hat{y}_1, \ldots, \hat{y}_{|\hat{\mathcal{Y}}|}\} \subset \mathcal{Y}$ を用いて近似される．翻訳時は未知の文に対する参照訳集合を手に入れるのが困難であるため，典型的な MBR 復号では機械翻訳による翻訳候補を擬似的な参照訳集合とみなし，$\hat{\mathcal{Y}} \coloneqq \mathcal{H}$ とされる．ここで，$N \coloneqq |\mathcal{H}|$ とおくと，MBR 復号の時間計算量は $\mathcal{O}(N^2)$ となる．

COMET COMET は，翻訳品質の直接評価（Direct Assesment; DA）スコアを予測するニューラルモデルを用いた，人手評価との相関係数が高い評価尺度である [19, 18]．原文 $x \in \mathcal{X}$，仮説文 $h \in \mathcal{Y}$，参照訳文 $\hat{y} \in \mathcal{Y}$ の 3 つ組文のそれぞれの文を XLM-RoBERTa (XLM-R) [4] によって D 次元文ベクトルに符号化した後，それらの文ベクトルを出力層に入力することにより，スコアが計算される．$f : \mathcal{X} \cup \mathcal{Y} \to \mathbb{R}^D$ を XLM-R による文ベクトルへの符号化関数，$s : \mathbb{R}^D \times \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ を出力層とすると，COMET スコアは次式のように定式化される．

$$\text{COMET}(x, h, \hat{y}) = s(f(x), f(h), f(\hat{y})). \tag{4}$$

COMET を用いた MBR 復号（COMET-MBR）では，式 3 の効用関数 u が COMET$(\cdot, \cdot, \cdot)$ に置き換えられ，仮説文集合を擬似的に参照訳文集合とみなした上で次のように計算される:

$$y^*_{\text{COMET-MBR}} = \operatorname*{argmax}_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{\hat{y} \in \hat{\mathcal{Y}}} [s(f(x), f(h), f(\hat{y}))]\,. \tag{5}$$

このとき，$\hat{\mathcal{Y}} = \mathcal{H}$ のような，典型的な MBR 復号の場合，はじめに一度だけ仮説文の文ベクトル集合 $\{f(h_1), \ldots, f(h_{|\mathcal{H}|})\}$ を計算し，その中からペアを取り出して関数 s を適用することで，出力層 s のみの N^2 回の計算で全ての仮説文に対するベイズリスクが求まる．

3 提案法: Centroid-Based MBR

提案法 Centroid-Based MBR (CBMBR) は，COMET 内部の文ベクトル空間に着目し，近い文ベクトルから計算されるスコアの期待値を，それらの重心から計算する近似により，復号速度の改善を狙う．提案法は **(1) 文符号化**，**(2) クラスタリング**，**(3) 期待効用計算** の順に従ってベイズリスクを計算することで復号する．

(1) 文符号化 はじめに，従来の COMET-MBR と同様に，原文 x の文ベクトル $f(x) \in \mathbb{R}^D$ と各疑似参照訳文の文ベクトル $\{f(\hat{y}_1), \ldots, f(\hat{y}_{|\hat{\mathcal{Y}}|})\} \subset \mathbb{R}^D$ を計算する．なお，$\hat{\mathcal{Y}} = \mathcal{H}$ である．

(2) クラスタリング 次に，疑似参照訳文ベクトルを $k \ll N$ 個のクラスタにクラスタリングし，各クラスタの重心ベクトル $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k\} \subset \mathbb{R}^D$ を求める．ここで，クラスタリングは推論時に行われるため，復号速度の観点から，提案法では収束の早いクラスタリング法である k-means++ [1] を採用する．k-means++は，はじめに，次のような手順に従って重心の初期値を獲得する．

1. $\{f(\hat{y}_1), \ldots, f(\hat{y}_{|\hat{\mathcal{Y}}|})\}$ からランダムに 1 つ文ベクトルを選択し，1 つめの重心とする．
2. 各文ベクトル $f(\hat{y}_i)$ とそれらの最近傍重心 $\boldsymbol{c}^* = \operatorname{argmin}_{\boldsymbol{c} \in \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k\}} \|f(\hat{y}_i) - \boldsymbol{c}\|_2$ との間の二乗ユークリッド距離 $d^2(\hat{y}_i) = \|f(\hat{y}_i) - \boldsymbol{c}^*\|_2^2$ を求める．
3. 各文ベクトル $f(\hat{y}_i)$ に対して重み $\frac{d^2(\hat{y}_i)}{\sum_{j=1}^{|\hat{\mathcal{Y}}|} d^2(\hat{y}_j)}$ を求め，求めた重みを確率分布とみなして $\hat{\mathcal{Y}}$ か

ら文ベクトルをサンプリングすることで，新たな重心を 1 つ得る．

4. 重心が k 個得られるまで手順 2, 3 を繰り返す．

続いて，獲得した初期値から，一般的な k-means によるクラスタリングを行う．具体的には，次の 1) と 2) を反復的に計算する: 1) 各文ベクトルの所属を，最近傍の重心を代表ベクトルとするクラスタに割り当て，2) 各クラスタに割り当てられた文ベクトルの重心を求めてクラスタ重心を更新する．

(3) 期待効用計算　最後に，式 5 の $f(\hat{y}) \in \mathbb{R}^D$ をクラスタ重心 $\boldsymbol{c} \in \mathbb{R}^D$ に置き換えて期待効用を計算する:

$$y^*_{\text{CBMBR}} = \underset{h \in \mathcal{H}}{\operatorname{argmax}} \underset{\boldsymbol{c} \in \{\boldsymbol{c}_i\}_{i=1}^{k}}{\mathbb{E}} [s(f(x), f(h), \boldsymbol{c})] . \quad (6)$$

従来法では，全ての仮説文の期待効用を計算するために $\mathcal{O}(N^2)$ の時間計算量を要していたのに対し，提案法は $\mathcal{O}(Nk)$ で計算される．

4　実験

実験設定　提案法 CBMBR の有効性を検証するため，英 ↔ 日，英 ↔ 独，英 ↔ 中の 6 言語方向の翻訳実験を行った．評価セットに WMT'22 翻訳タスク [13] を用い，COMET スコアを用いて翻訳品質を評価した．翻訳文の生成には訓練済み多言語翻訳モデルの M2M100 を用いた．MAP 復号にはビーム探索を用い，ビーム幅は 100 とした．MBR 復号にはビーム探索と top-p サンプリング ($p = 0.9$) を用い，それぞれビーム幅を 100 として 100-best 翻訳を生成し，計 200 文を翻訳仮説とした．比較のため，参照訳を用いない品質推定モデルによるリランキング（QE），QE によって 10 候補に枝刈りした後に MBR 復号する QE→MBR [8]，また，参照訳を用いて最大スコアの仮説を選択する品質上限（Oracle）の翻訳品質を評価した．QE モデルには COMET-QE [20] を用いた．提案法の CBMBR について，GPU を用いた他の復号法と速度を比較するため，k-means++は PyTorch で実装した．k-means の反復回数は最大 3 回とし，$k = 8$ とした．文符号化と品質推定については全てバッチサイズ 128 文でミニバッチ化して計算した．復号時間については，全実験の復号時間を処理単位で計測し，処理ごとに 1 文あたりの平均処理時間（ミリ秒）を算出した．

翻訳品質　はじめに，翻訳品質の比較結果を表 1 に示す．表中の平均スコアより，MAP 復号と比較して，MBR 復号および提案法の CBMBR 復号が COMET スコアを +4.3% 改善し，Oracle との差が 7.4%から 3.1%まで縮んだことがわかる．また，QE モデルによるリランキングと翻訳品質を比較すると，CBMBR のほうが 0.5% スコアが高い．QE→MBR と CBMBR を比較すると，言語方向によって最適な手法が入れ替わるものの，平均的に 0.1% 以内の差で収まっている．なお，CBMBR は COMET モデルのみで計算するのに対し，QE→MBR は COMET モデルに加えて COMET-QE モデルを必要とするため，実行時のメモリ使用量が増加する点に注意されたい．

表 1　WMT'22 翻訳タスクにおける各復号法の翻訳品質の比較（COMET%）．

復号法	英日	日英	英独	独英	英中	中英	平均
MAP	79.7	70.3	77.5	79.2	76.7	71.7	75.9
MBR	84.3	74.9	82.8	82.5	81.7	74.9	80.2
QE [20]	83.9	74.4	81.6	81.7	81.1	75.3	79.7
QE→MBR [8]	84.4	75.1	82.5	82.3	81.7	75.6	80.3
CBMBR	84.9	74.6	82.6	82.2	81.9	75.1	80.2
Oracle	87.2	79.1	85.3	85.2	84.8	78.3	83.3

復号速度　次に，各復号法の復号時間の比較を表 2 に示す．表より，復号全体の処理時間は CBMBR が従来の MBR よりも 1.4 倍速いことがわかる．特に，$\mathcal{O}(N^2)$ の計算量を要していたベイズリスクの計算においては，k-means++まで含めて 13.6 倍，期待効用計算のみで比較すると 26.2 倍高速化した．また，QE や QE→MBR においては，品質推定の時間計算量は $\mathcal{O}(N)$ にもかかわらず実際の処理時間が長いが，これは COMET-QE [20] モデルが COMET モデルとは異なり，原文と仮説文の 2 文を結合して符号化するため，品質推定モデルの入力系列長が COMET モデルよりも長いためであると考えられる．

以上より，実験結果をまとめると，CBMBR は従来の MBR と比較して，メモリ使用量が増加することなく，同程度の翻訳品質を維持しながら，ベイズリスクの計算時間を 13.6 倍，うち期待値の計算が 26.2 倍高速化することを確認した．

5　考察

5.1　クラスタ数 k と翻訳品質

クラスタ数 k と翻訳品質の関係を調べるため，$k \in \{1, 4, 8, 12\}$ を変化させたときの翻訳品質を COMET スコアによって評価した．表 3 は開発セット WMT'21 翻訳タスクにおける実験結果を示す．

表 2 WMT'22 翻訳タスクにおける各復号法の 1 文あたりの平均処理時間（ミリ秒）.

	MBR	QE	QE→MBR	CBMBR
原文符号化	3.7	–	3.3	3.6
仮説文符号化	360.0	–	22.8	353.3
品質推定	–	584.5	555.0	–
ベイズリスク計算				
k-means++	–	–	–	5.3
期待効用計算	149.2	–	3.0	5.7
合計	513.0	584.5	584.1	367.9

表 3 開発セット WMT'21 翻訳タスクにおけるクラスタ数 k と翻訳品質の比較.

k	英日	日英	英独	独英	英中	中英	平均
1	84.1	68.5	80.2	83.2	78.0	73.5	77.9
4	85.0	69.5	80.7	83.3	78.6	74.2	78.6
8	85.1	69.9	80.9	83.4	78.8	74.5	78.8
12	85.1	70.0	80.9	83.5	78.8	74.5	78.8

$k = 1$ のときは時間計算量が $\mathcal{O}(N)$ となる一方で，$k = 4$ のときと比べて COMET スコアが 0.7%低い．また，k を 4, 8, 12 と増やすにつれ品質も改善しているが，$k = 8$ から $k = 12$ への増加では品質にほとんど差がないことがわかる．一方で，CBMBR の時間計算量は $\mathcal{O}(Nk)$ であるため，k を大きくするにつれて，復号時間は線形に増加する．そのため，4 節の翻訳実験では $k = 8$ を採用した．

5.2 類似文間の文ベクトル距離

提案法が成り立つための条件である，“COMET の文ベクトル空間では類似文同士のベクトル間距離が近くなる” という仮定を確かめるため，文間意味的類似度（Semantic Textual Similarity; STS）タスクを用いて，類似文間の文ベクトル距離を分析した．

評価データには STS-Benchmark (STS-B) [2] を用い，正解類似度との Pearson の相関係数 r を算出した．実験結果を表 4 に示す．表中の “Pearson $r \times 100$” は相関係数 r を 100 倍したスコアを示す．表より，COMET は，文ベクトル空間に対して対照学習のような教師あり学習を行っていないにもかかわらず，73.6 と強い相関を示している．また，COMET は XLM-RoBERTa (XLM-R) [4] に対して翻訳品質を予測するよう追加学習したモデルであるが，XLM-R のスコア 31.6 と比較すると，COMET の追加学習によって間接的に，意味的類似文を近いベクトルに写像するよう学習していることがわかる．さらに，対訳文間の文ベクトル類似度を高める学習を行った [7] より 0.9%スコアが高いことも確認した．

表 4 文間意味的類似度タスク STS-B の開発セットと評価セットにおける評価結果.

	Pearson $r \times 100$	
	開発	評価
XLM-R_{large} [4]	39.1	31.6
LaBSE [7]	72.9	72.7
$ME5_{large}$ [22]	87.4	83.6
COMET [18]	78.2	73.6

以上より，COMET は文ベクトル空間に対して対照学習のような明示的な学習を加えていないにもかかわらず，STS-B タスクにおいて正解類似度と強い正の相関を示し，翻訳品質を予測する学習を通して，副次的に類似文のベクトル間距離を近づけるように学習していることが確認できた．

6 関連研究

MBR 復号は統計的自動音声認識 [10] や統計的機械翻訳 [14] などで有効性が示され，近年ではニューラル機械翻訳への応用が進んでいる [5, 16]．また，今後，ブラックボックスなテキスト生成システムが増えた場合，複数システムからよりよい生成文を選択するという設定において特に MBR が有用であるといえる [11]．ただし，MBR の復号速度については依然として課題である．復号速度を改善するために翻訳仮説を枝刈りする手法がいくつか提案されており，翻訳仮説の軽量な代替評価尺度を用いた枝刈り [6]，QE モデルによる枝刈り [8]，少量の疑似参照訳から始めて仮説を枝刈りしながら段階的に疑似参照訳を増やしていく手法 [3] が挙げられる．本研究は先行研究と異なり，仮説側を枝刈りすることなく，疑似参照訳側を枝刈りした手法とみなせる．

7 おわりに

本研究では，COMET モデルを用いた最小ベイズリスク復号 COMET-MBR において，全ての疑似参照訳を用いる代わりに，疑似参照訳の文ベクトルのクラスタ重心表現のみを用いてベイズリスクを計算することにより，復号速度を改善した．実験より，WMT'22 翻訳タスクの英 ↔ 日，英 ↔ 独，英 ↔ 中の 6 言語方向の翻訳において，従来の MBR 復号と比較して，COMET スコアが低下することなく，ベイズリスクの計算時間が 13.6 倍，うち期待値の計算が 26.2 倍速くなることを確認した．今後は，ニューラルモデルを用いた他の品質評価尺度である BLUERT [21] 等に向けて提案法を適用したい．

謝辞

本研究の一部は JSPS 科研費 JP22J11279, JP22KJ2286 の助成を受けたものである．ここに謝意を表する．

参考文献

[1] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In **Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms**, SODA '07, p. 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics.

[2] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)**, pp. 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[3] Julius Cheng and Andreas Vlachos. Faster minimum Bayes risk decoding with confidence-based pruning. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 12473–12480, Singapore, December 2023. Association for Computational Linguistics.

[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[5] Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 4506–4520, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[6] Bryan Eikema and Wilker Aziz. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 10978–10993, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[7] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[8] Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. Quality-aware decoding for neural machine translation. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1396–1412, Seattle, United States, July 2022. Association for Computational Linguistics.

[9] Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 9198–9209, Singapore, December 2023. Association for Computational Linguistics.

[10] Vaibhava Goel and William J Byrne. Minimum bayes-risk automatic speech recognition. **Computer Speech & Language**, Vol. 14, No. 2, pp. 115–135, 2000.

[11] Ikumi Ito, Takumi Ito, Jun Suzuki, and Kentaro Inui. Investigating the effectiveness of multiple expert models collaboration. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 14393–14404, Singapore, December 2023. Association for Computational Linguistics.

[12] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics.

[13] Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. Findings of the 2022 conference on machine translation (WMT22). In **Proceedings of the Seventh Conference on Machine Translation (WMT)**, pp. 1–45, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[14] Shankar Kumar and William Byrne. Minimum Bayes-risk decoding for statistical machine translation. In **Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004**, pp. 169–176, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[16] Mathias Müller and Rico Sennrich. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 259–272, Online, August 2021. Association for Computational Linguistics.

[17] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised learning of sentence embeddings using compositional n-gram features. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 528–540, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[18] Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In **Proceedings of the Seventh Conference on Machine Translation (WMT)**, pp. 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[19] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[20] Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In **Proceedings of the Seventh Conference on Machine Translation (WMT)**, pp. 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[21] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

[22] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2022.

A　参考情報

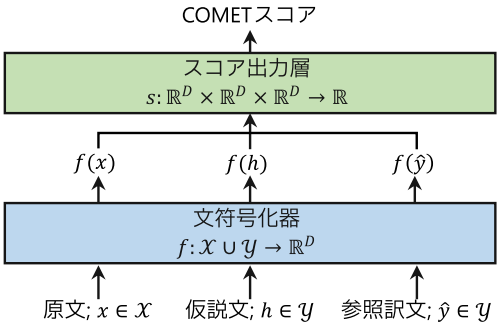

図 2　COMET モデルの概要図.

表 5　実験設定の詳細.

項目	設定
モデル	
翻訳文生成	M2M100 (615M パラメータ)[1)]
COMET	`Unbabel/wmt22-comet-da`
	MBR，品質評価の両方に使用
COMET-QE	`Unbabel/wmt22-cometkiwi-da`
演算	
GPU	NVIDIA A6000 1 基
浮動小数点精度	単精度
その他	
文符号化	表 4 および 9 において，文符号化方法が明示的に定義されていない XLM-R [4] などのモデルでは，COMET に合わせて，最終層の中間表現の平均ベクトルを文ベクトルとした.

表 6　開発セット WMT'21 翻訳タスクにおける翻訳品質（COMET スコア）の比較.

復号法	英日	日英	英独	独英	英中	中英	平均
MAP	80.4	65.3	75.7	80.4	74.1	71.3	74.5
MBR	84.6	68.9	80.9	83.7	78.7	73.9	78.5
QE	83.7	69.7	80.0	83.1	77.8	74.5	78.1
QE→MBR	84.0	69.8	80.6	83.6	78.2	74.7	78.5
CBMBR（k-means 反復回数: 3）							
$k = 1$	84.1	68.5	80.2	83.2	78.0	73.5	77.9
$k = 4$	85.0	69.5	80.7	83.3	78.6	74.2	78.6
$k = 8$	85.1	69.9	80.9	83.4	78.8	74.5	78.8
$k = 12$	85.1	70.0	80.9	83.5	78.8	74.5	78.8
CBMBR ($k = 8$)							
反復回数: 1	85.1	69.8	80.9	83.4	78.8	74.5	78.7
反復回数: 2	85.1	69.9	80.9	83.4	78.8	74.5	78.8
反復回数: 3	85.1	69.9	80.9	83.4	78.8	74.5	78.8
反復回数: 4	85.1	70.0	80.9	83.4	78.8	74.5	78.8
反復回数: 5	85.1	70.0	80.9	83.4	78.8	74.5	78.8
Oracle	86.2	72.6	82.4	85.8	80.5	76.6	80.7

1)　https://dl.fbaipublicfiles.com/flores101/pretrained_models/flores101_mm100_615M.tar.gz

表 7　開発セット WMT'21 翻訳タスクにおける 1 文あたりの平均処理時間の比較（ミリ秒）.

		QE→	CBMBR			
	MBR	QE	MBR	$k = 4$	$k = 8$	$k = 12$
原文符号化	4.2	–	3.5	4.3	4.3	4.3
仮説文符号化	486.7	–	29.3	493.3	490.0	500.0
品質推定	–	795.9	796.5	–	–	–
ベイズリスク計算						
k-means++	–	–	–	2.8	5.1	7.2
期待効用計算	142.7	–	3.0	2.8	5.6	9.0
合計	633.5	795.9	832.4	503.2	505.0	520.5

表 8　開発セット WMT'21 翻訳タスクにおける，クラスタ数と反復回数を変化させたときの 1 文あたりの平均クラスタリング時間（ミリ秒）.

反復回数	$k = 4$	$k = 8$	$k = 12$
1	1.6	3.2	4.8
2	2.2	3.9	6.2
3	2.8	5.1	7.2
4	3.4	6.0	8.8
5	3.9	6.9	9.6

表 9　文間意味的類似度タスク STS-B の開発セットと評価セットにおける評価結果．表 4 に加えて単言語モデルの評価結果を含めた．なお，"多言語モデル" は "✓" が付いているモデルが多言語に対応していることを示す.

		Pearson $r \times 100$	
	多言語モデル	開発	評価
$\text{RoBERTa}_{\text{large}}$ [15]	–	53.7	43.0
fastText [12]	–	65.3	53.6
Sent2Vec [17]	–	78.7	75.5
$\text{XLM-R}_{\text{large}}$ [4]	✓	39.1	31.6
LaBSE [7]	✓	72.9	72.7
$\text{ME5}_{\text{large}}$ [22]	✓	87.4	83.6
COMET [18]	✓	78.2	73.6

参考情報として，COMET モデル [19, 18] の概要図を図 2 に示す．また，実験設定の詳細（表 5），開発セット WMT'21 翻訳タスクにおける翻訳品質と復号速度の詳細な比較（表 6, 7, 8），および，単言語モデルまで含めた文間意味的類似度タスクの比較結果（表 9）をそれぞれ記す.