人物埋め込み空間の内挿性と制御性を兼ね備えた応答生成モデル

[1] 奈良先端科学技術大学院大学

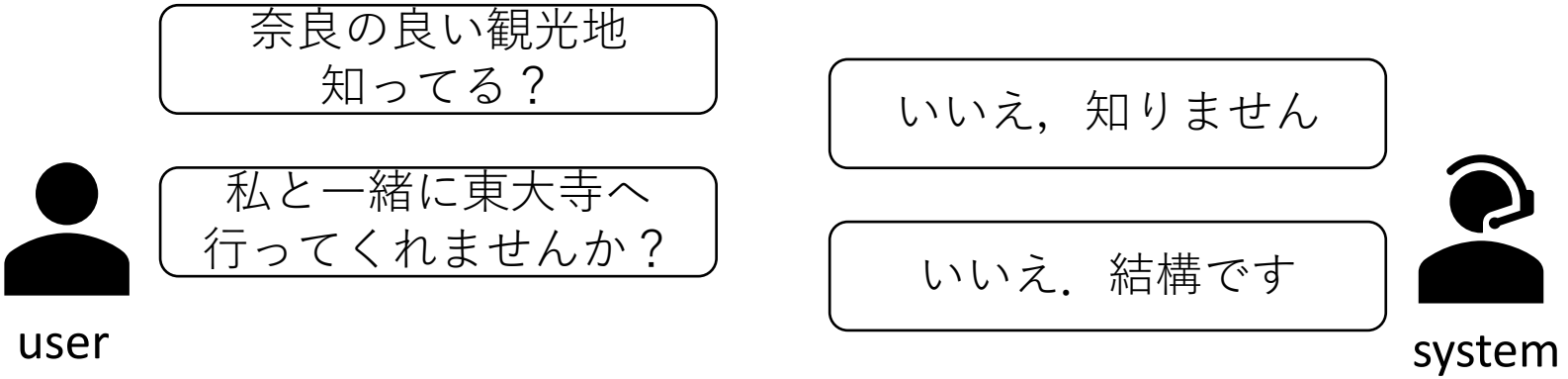
[2] NTT コミュニケーション科学基礎研究所

安川 浩貴[1] 水上 雅博[2] 品川 政太朗[1] 杉山 弘晃[2] 須藤 克仁[1] 中村 哲[1]

人物情報を反映した応答生成モデルの貢献

人物情報を含んだ一貫性のある対話の利点は？

Dull responseを返す場合……あまり魅力的ではない

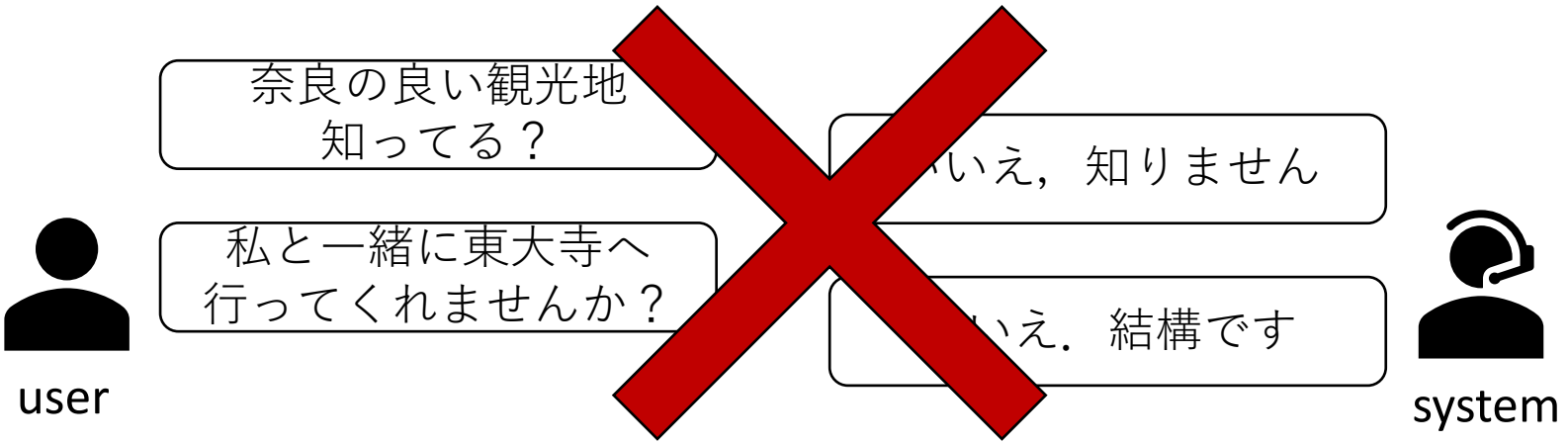


Dull responseを返す雑談対話システムとのやり取りの例

人物情報を反映した応答生成モデルの貢献

人物情報を含んだ一貫性のある対話の利点は？

Dull responseを返す場合……あまり魅力的ではない



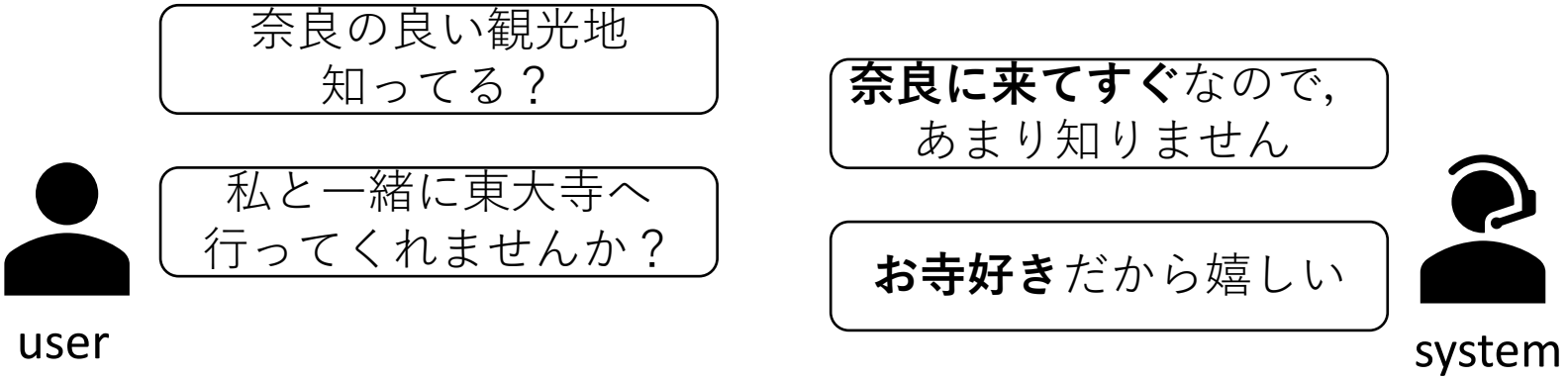
Dull responseを返す雑談対話システムとのやり取りの例

人物情報を反映した応答生成モデルの貢献

人物情報を含んだ一貫性のある対話の利点

— 対話システムの**信頼性**を高め，**魅力的**にする [Miyazaki et al., 2021]

— 生成文が対話相手へ与える**魅力度**を高める [Zhang et al., 2018]



人物情報を考慮した雑談対話システムとのやり取りの例

→対話相手を楽しませることができる！

人物情報を反映した応答生成モデルの実現方法

用いる人物情報によって分けられ，それぞれ利点がある

1．ユーザ識別子：人物を識別する識別子を使用

→ユーザ識別子から人物を表現する埋め込みを学習

Pros. 埋め込みの演算による新たな人物の作成が可能

Cons. 狙った特徴を持つ人物の埋め込みを作ることは難しい

2．ペルソナ文：人物を表現する文を使用

→ ペルソナ文を用いて応答文へ反映したい人物の特徴を指定

Pros. 任意の性質への変更が容易

Cons. データ数が少なく，構造化された埋め込みの学習が難しい

本研究の目的

本研究では...

内挿性と**制御性**のある応答生成モデルを実現したい

⇓

ユーザ識別子とペルソナ文の双方を扱える
応答生成モデルの提案

本研究の目的

本研究では…

内挿性と制御性のある応答生成モデルを実現したい

⇓

ユーザ識別子とペルソナ文の双方を扱える
応答生成モデルの提案

Aさん　　Bさん　　AさんとBさんの間の人

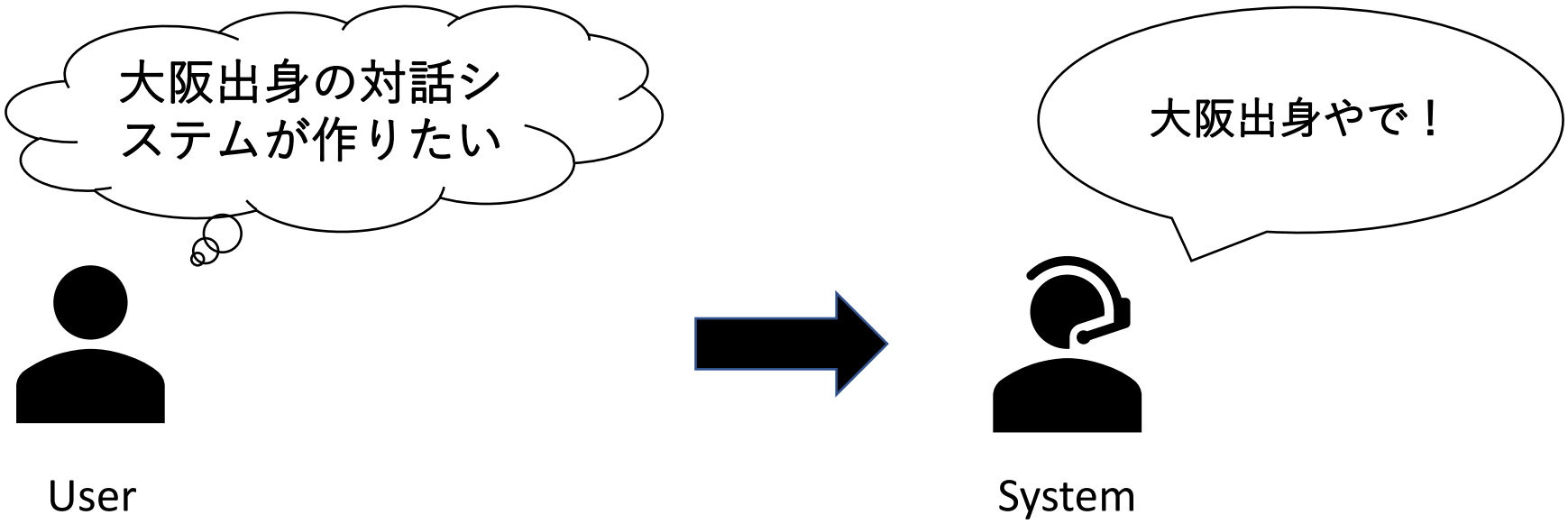
本研究の目的

本研究では...

内挿性と**制御性**のある応答生成モデルを実現したい

⇓

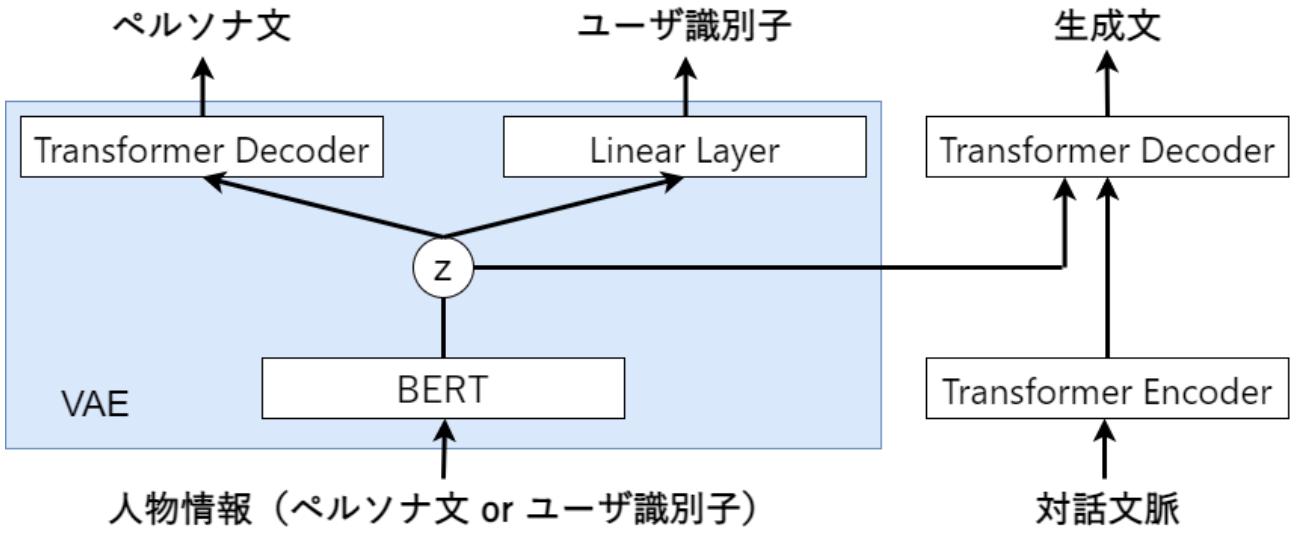
ユーザ識別子とペルソナ文の双方を扱える
応答生成モデルの提案



アプローチ

具体的に

- ユーザ識別子とペルソナ文の双方を扱えるVAEを用いてある人物を表現する人物埋め込みを計算
  → 構造化された埋め込み空間の構築
  → 制御が容易な人物埋め込みの実現



本研究の概観

人物情報を反映した応答文生成モデルの実現

応答文に紐づく**人物埋め込み**の学習が重要である

本研究の概観

人物情報を反映した応答文生成モデルの実現

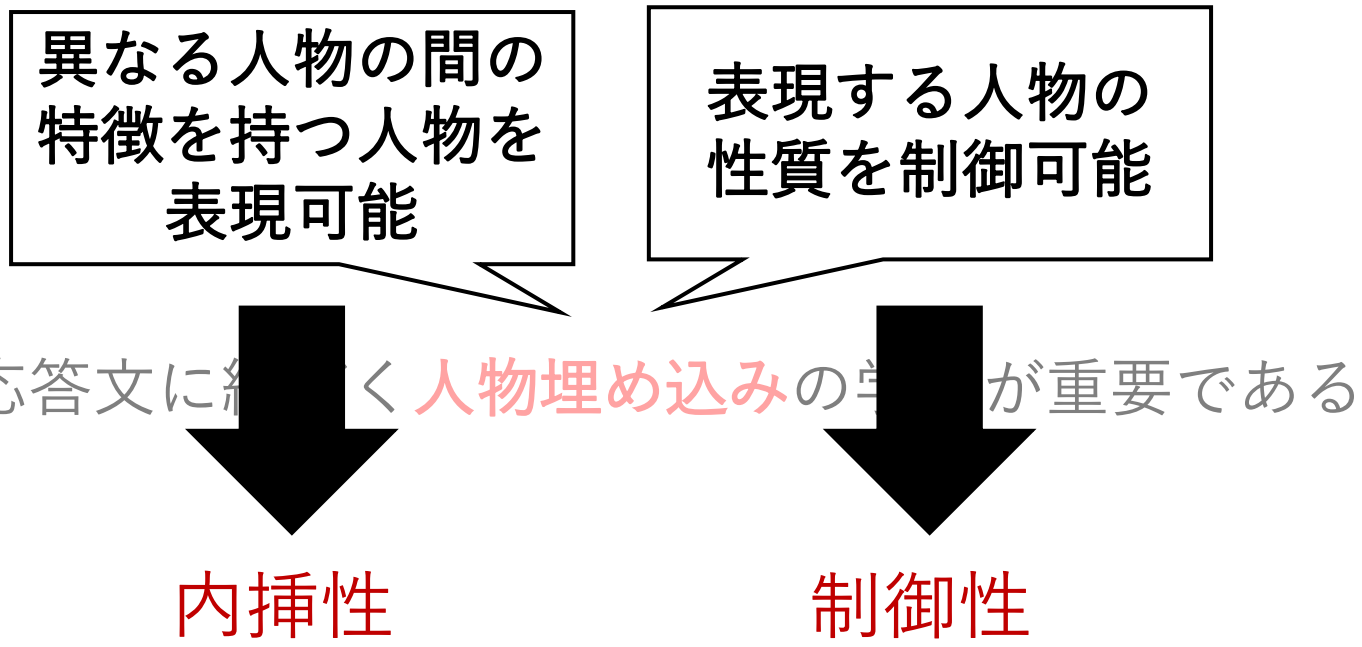
異なる人物の間の特徴を持つ人物を表現可能

表現する人物の性質を制御可能

応答文に紐づく**人物埋め込み**の学習が重要である

本研究の概観

人物情報を反映した応答文生成モデルの実現

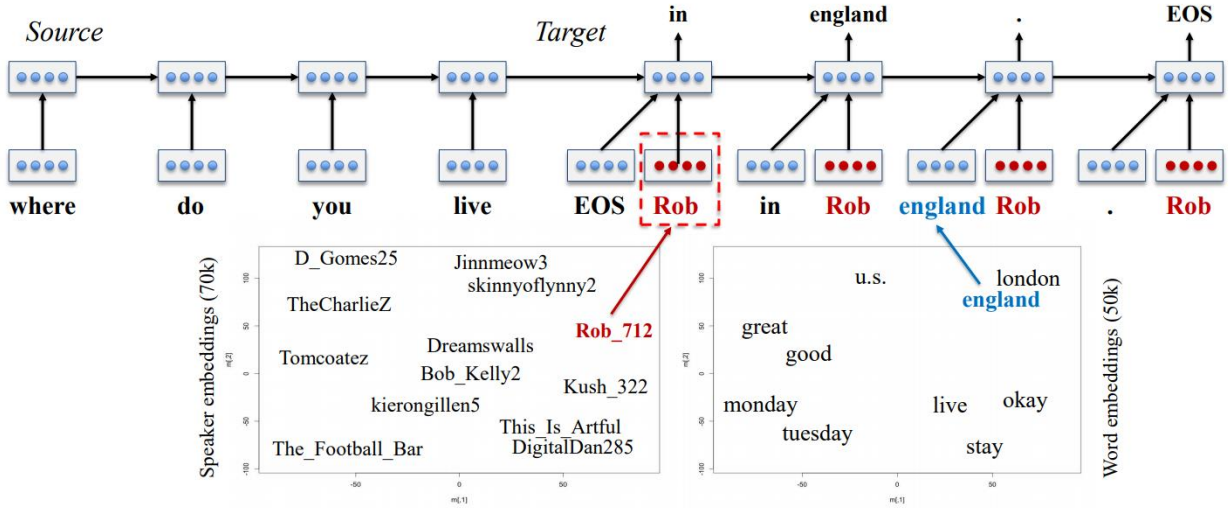


先行研究 ― ユーザ識別子 ―

埋め込み表現を用いるモデル [Li et al., 2016]

・対話とユーザ識別子を含んだデータセット　例）Twitter

・ユーザを表現する埋め込み表現を学習し，応答生成に用いる



→埋め込み同士の演算により，
　演算に用いた人物の特徴に応じた新たな人物を作成可能

→狙った特徴の人物を表現するのは難しい

先行研究　― ペルソナ文 ―

PERSONA-CHAT [Zhang et al., 2018]

・ユーザの情報を記述した文章 (**ペルソナ文**)

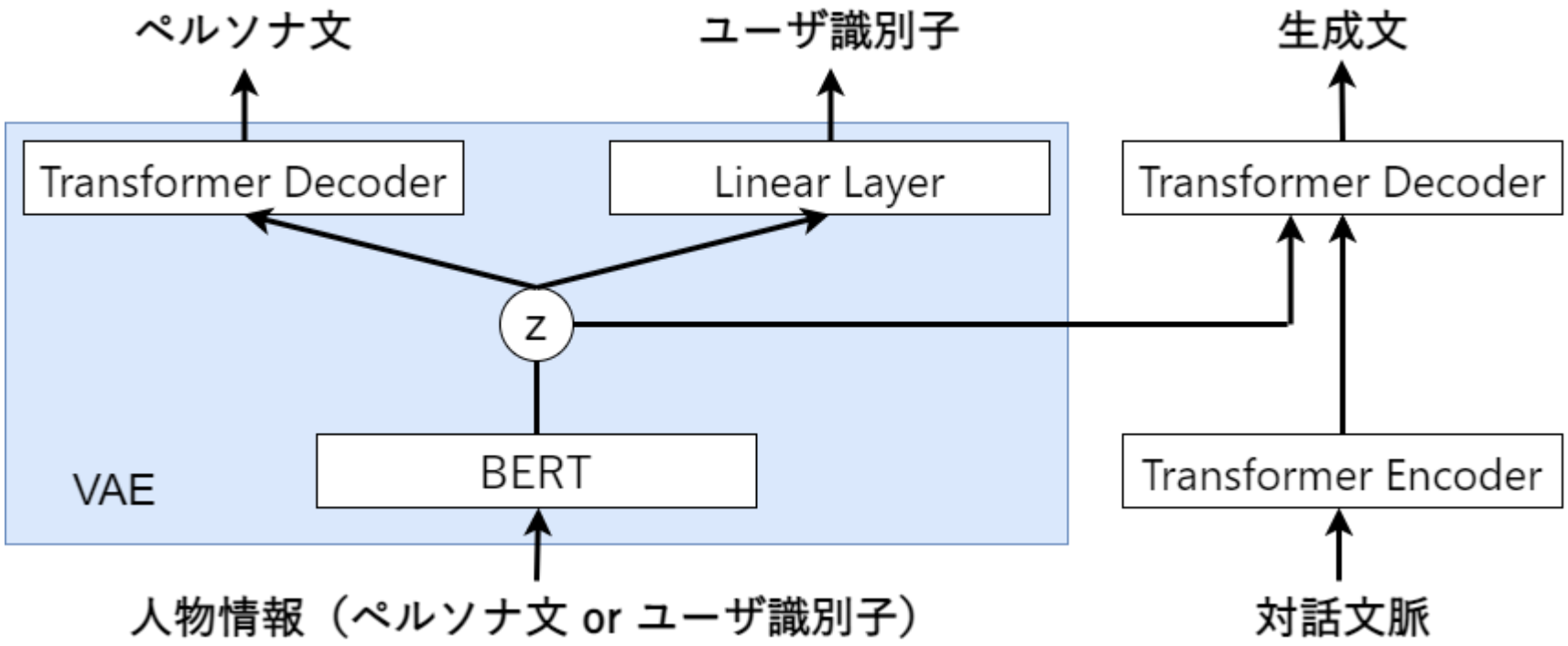
・ペルソナ文に基づいて行われた**対話**

人物1	人物2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
対話	
[人物1] I like to go skiing. How about you?	
[人物2] I once went skiing with my four children.	

→応答文に反映させる人物の性質を容易に指定可能

→データ数が少なく，人物情報が限られている

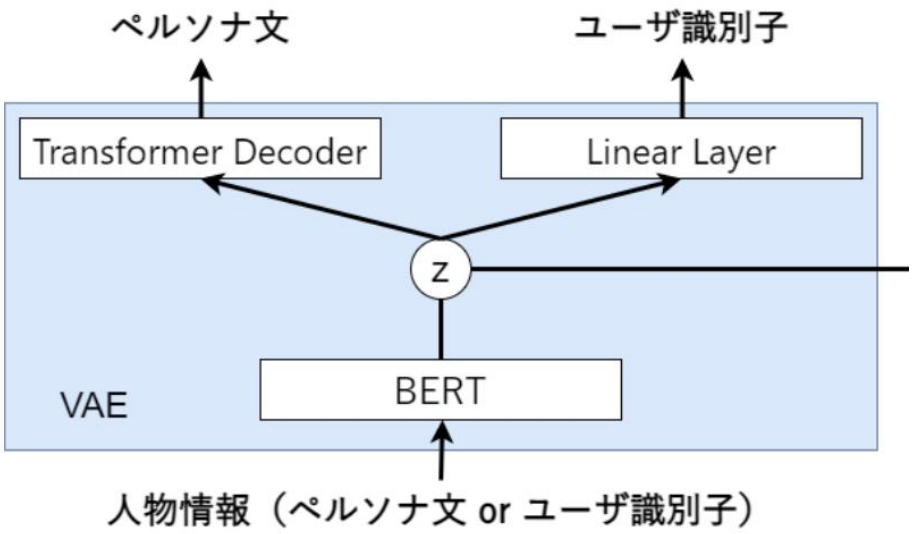
提案手法 ― モデル全体図 ―

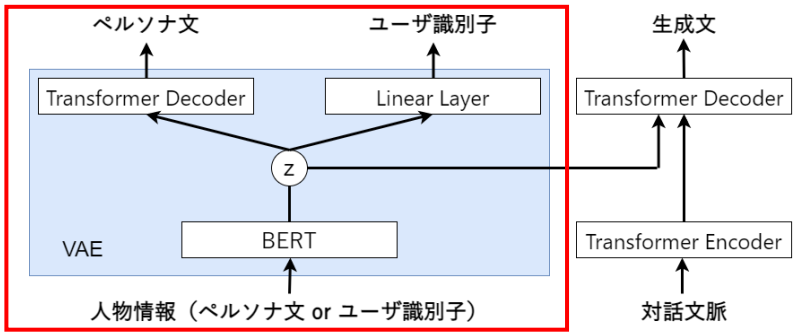


人物埋め込みzを獲得するVAE（左）と
応答文を生成する応答生成モデル（右）

提案手法 ― VAEによる人物埋め込みモデル ―

与えられた人物情報から人物埋め込みを計算



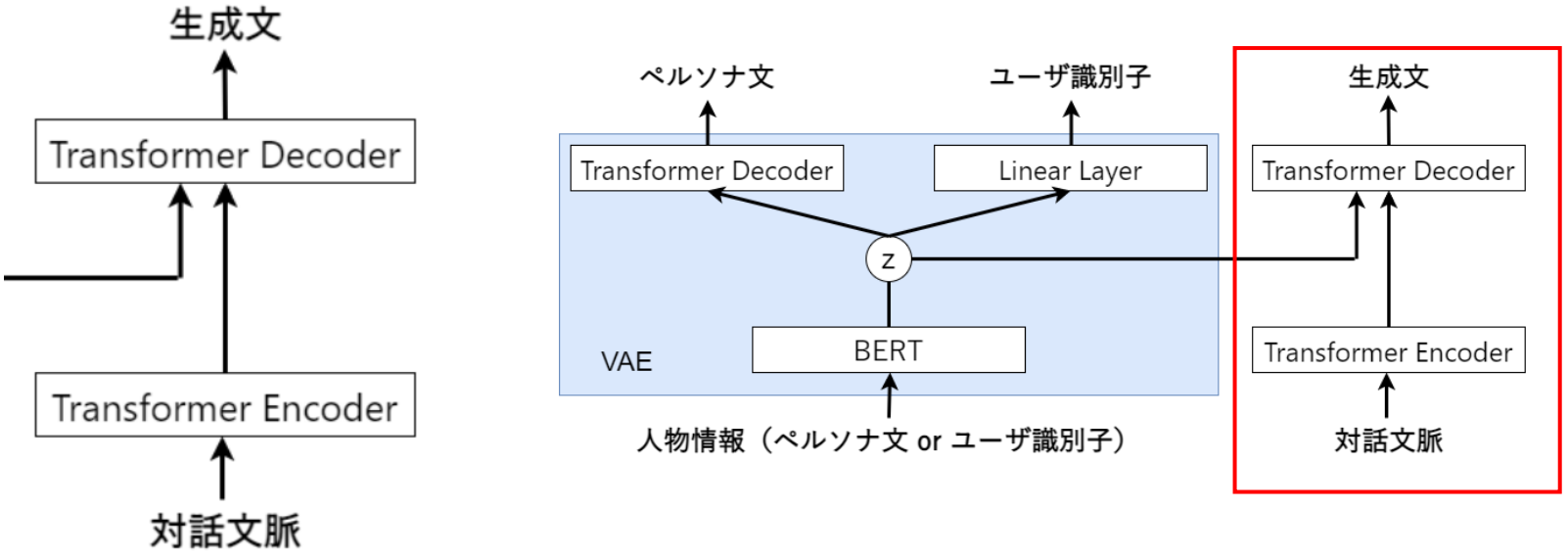


・Encoder：人物情報に応じた人物埋め込みzを計算

・Decoder：人物埋め込みzから人物情報を予測

提案手法 ― 応答生成モデル ―

人物埋め込みzと対話文脈を受け取り，応答生成を行う



・Transformer Encoder-decoder[Vaswani et al., 2017]ベース
→人物埋め込みzと対話文脈から人物の特徴を反映した応答文を生成

実験設定

○データセット

・JPersonaChat [Sugiyama et al., 2023]
　→ペルソナ文5文が人物情報として付属
・趣味雑談コーパス[Sugiyama et al., 2020]
　→ユーザ識別子が人物情報として付属

→混合して使用

○実験

実験1. 生成文が人物情報を反映できているか
実験2. ペルソナ文から得られた人物埋め込みは内挿性を持つか
実験3. 生成文は人物埋め込みに紐づいているか
実験4. ペルソナ文から得られた生成文は内挿性を持つか

実験 1
― 生成文が人物情報を反映できているか ―

評価方法

・提案モデル及びベースラインモデルが正解データへ与える**Perplexity（PPL）**を使用

・**PPL**が低い（正解データへ与える生起確率が高い）ほど生成文が人物情報を反映できている

↓

正解データに近い文を生成できたかをベースラインと比較

※ベースライン

一般的なTransformer Encoder-decoder モデル

結果 1
― 生成文が人物情報を反映できているか ―

正解データに近い文を生成できたかをベースラインと比較

評価対象	PPL（↓）
ベースライン	24.157
提案モデル	21.899

提案モデルはベースラインよりも低い**PPL**の値を獲得

→人物情報を反映した応答生成において優れている

実験 2
― ペルソナ文から得られた人物埋め込みは内挿性を持つか ―

異なる人物の間の特徴を持つ人物（**中間人物**）を
表現するように人物埋め込みを学習できているのか検証

中間人物

ある人物（**source**）のペルソナ文を他の人物（**target**）の
ペルソナ文とk文入れ替えて作成

kを大きくするに従い**中間人物**の埋め込みが**target**に
近づけば，人物埋め込みは内挿性を持つと言える

中間人物の作成

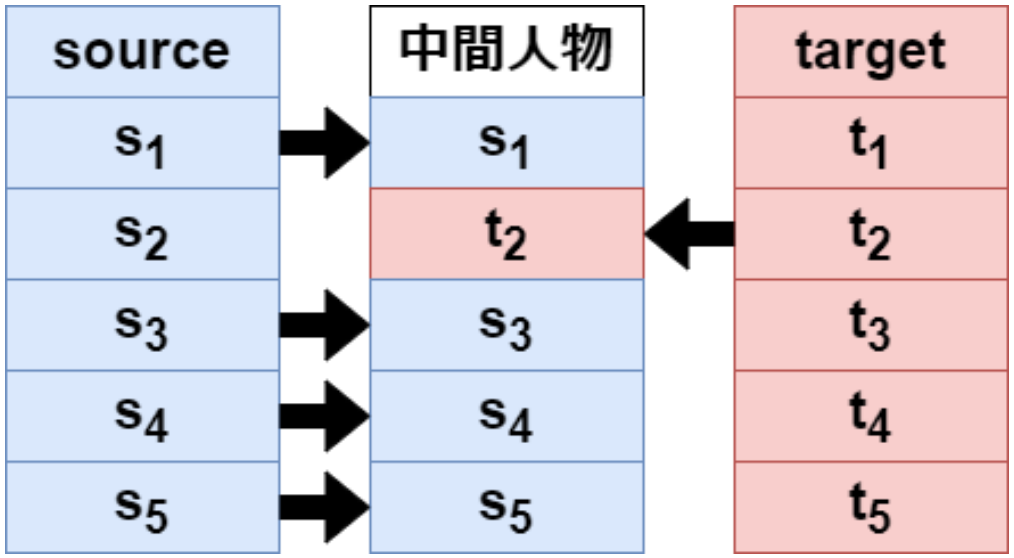
sourceのペルソナ文5文の内，k文を**target**と入れ替えて得られるペルソナ文5文を持つ人物を**中間人物**とする

source	中間人物	target
s_1	—	t_1
s_2	—	t_2
s_3	—	t_3
s_4	—	t_4
s_5	—	t_5

中間人物の作成

sourceのペルソナ文5文の内，k文を**target**と入れ替えて得られるペルソナ文5文を持つ人物を**中間人物**とする

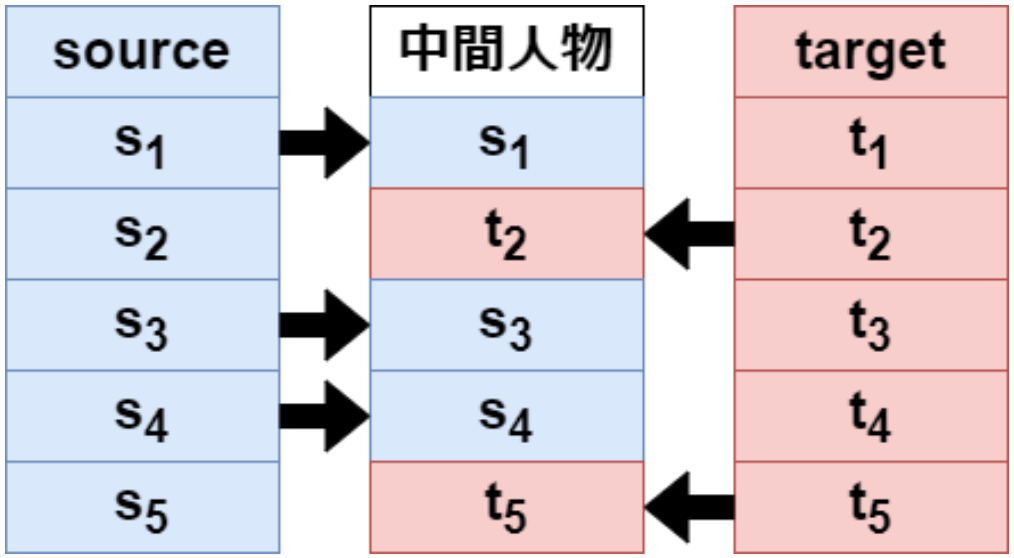
k=**1**の場合



中間人物の作成

sourceのペルソナ文5文の内，k文を**target**と入れ替えて得られるペルソナ文5文を持つ人物を**中間人物**とする

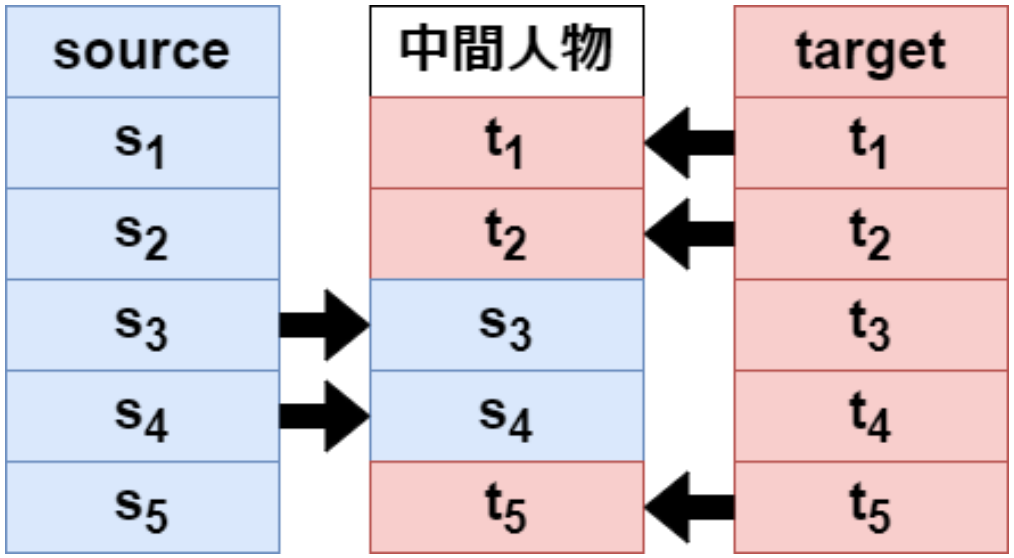
k=**2**の場合



中間人物の作成

sourceのペルソナ文5文の内，k文を**target**と入れ替えて得られるペルソナ文5文を持つ人物を**中間人物**とする

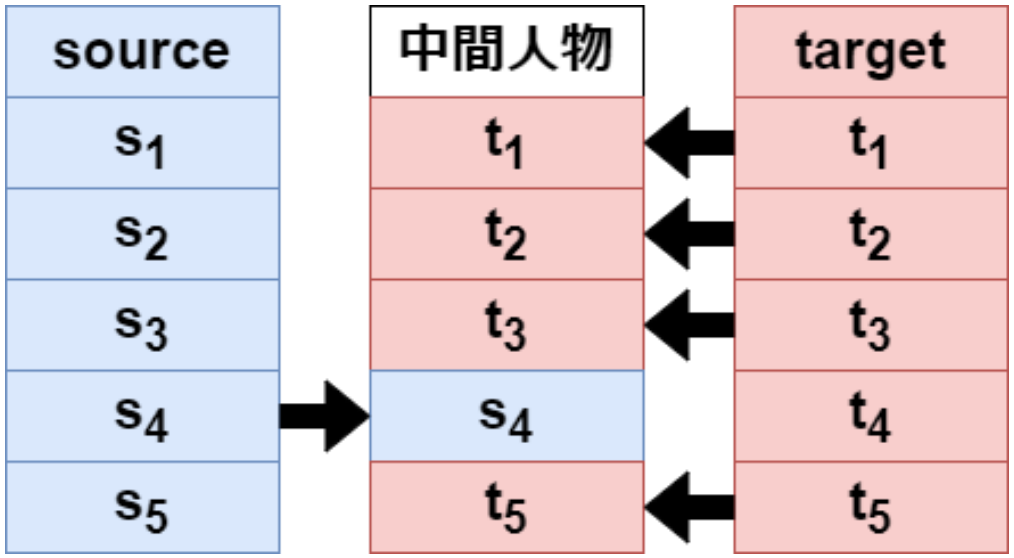
k=**3**の場合



中間人物の作成

sourceのペルソナ文5文の内，k文を**target**と入れ替えて得られるペルソナ文5文を持つ人物を**中間人物**とする

k=**4**の場合



実験 2
― ペルソナ文から得られた人物埋め込みは内挿性を持つか ―

中間人物と**target**の近さの評価

→人物埋め込みの**KL Divergence**（**KL Div**）で計測

・kが大きくなるに従い,
中間人物と**target**の埋め込み間の**KL Div**が小さくなる

↓

異なる人物の間にいる人物を，その近さに応じて表現可能

＝内挿性がある

結果 2
— ペルソナ文から得られた人物埋め込みは内挿性を持つか —

kが大きくなると中間人物埋め込みが**target**へ近づく

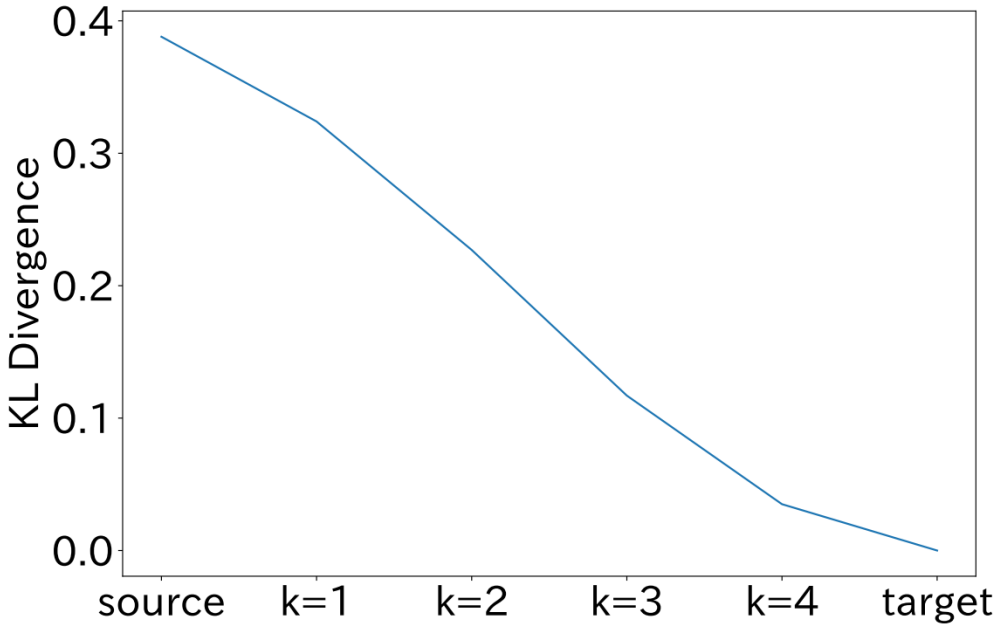


図
targetの人物埋め込みとの**KL Div**

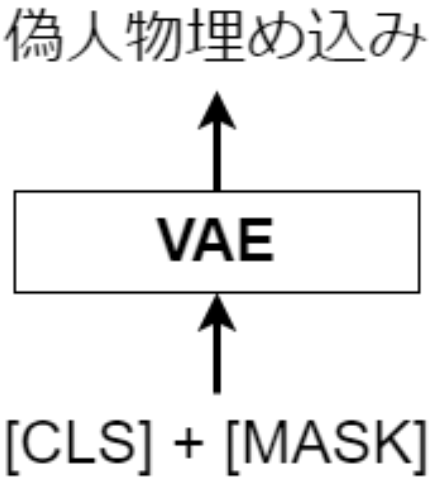
ペルソナ文から得られた人物埋め込みは内挿性を持つように学習できた

実験 3
― 生成文は人物埋め込みに紐づいているか ―

無関係な人物埋め込みを用いた場合，**PPL**が悪化する

実験内容

1. VAEへ「[CLS]+[MASK] 」を入力し，応答文に紐づいた人物埋め込みと異なる埋め込み（**偽人物埋め込み**）を計算

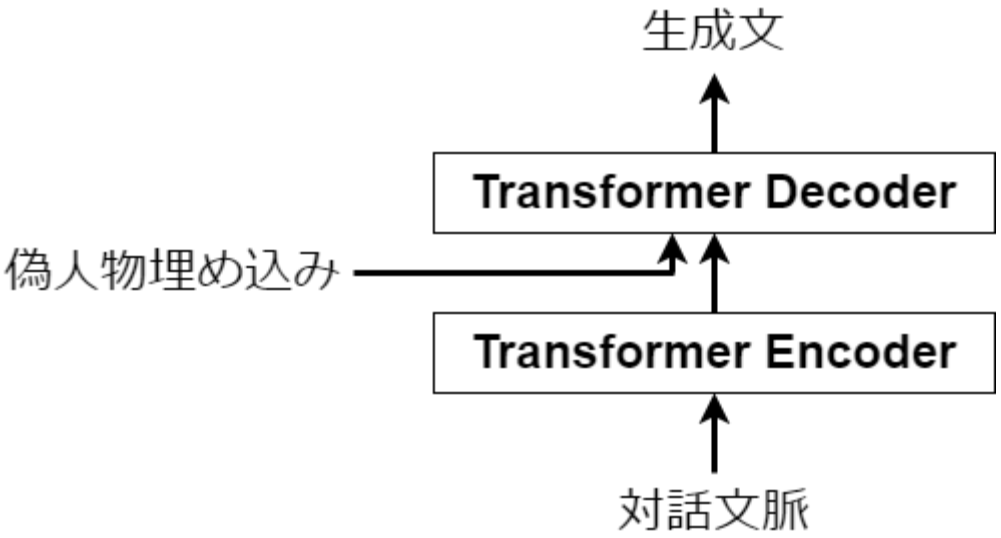


実験 3
― 生成文は人物埋め込みに紐づいているか ―

無関係な人物埋め込みを用いた場合，**PPL**が悪化する

実験内容

2. 正解データの応答文に紐づく人物埋め込みの代わりに**偽人物埋め込み**を用いて応答文生成を行い，**PPL**を計算



実験 3
― 生成文は人物埋め込みに紐づいているか ―

無関係な人物埋め込みを用いた場合，**PPL**が悪化する

実験内容

3. 応答文に対応する人物埋め込みを用いた場合と，**偽人物埋め込み**を用いた場合の**PPL**を比較

結果 3
― 生成文は人物埋め込みに紐づいているか ―

評価対象	PPL（↓）
偽人物埋め込み	24.014
元々の埋め込み	21.899

偽人物埋め込みを使用した場合，PPLは高くなり悪化した

→生成文は人物埋め込みに紐づいている

実験 4
― ペルソナ文から得られた生成文は内挿性を持つか ―

異なる人物の間の特徴を持つ人物（**中間人物**）の特徴を
反映した応答文を生成できるのかを検証

中間人物（実験2と同様）
ある人物（**source**）のペルソナ文を他の人物（**target**）の
ペルソナ文とk文入れ替えて作成

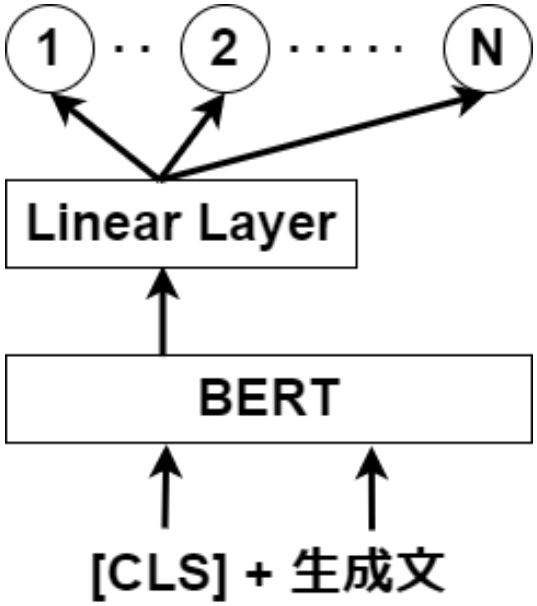
kを大きくするに従い生成文の**target**らしさが段階的に増加

→**中間人物**の特徴を反映した応答文を生成できると言える

実験 4
― ペルソナ文から得られた人物埋め込みは内挿性を持つか ―

生成文の**target**らしさの評価

→文に紐づいた人物を予測する分類器で計測



・[CLS]トークンと応答文を**BERT**へ入力
・応答文に紐づいた人物に関する
多値分類

学習結果

・精度　　：**48.71%**
・**chance rate**：**0.55%**

→生成文がどの人物によるものなのかを捉えられる

結果 4
― ペルソナ文から得られた生成文は内挿性を持つか ―

kが大きくなると生成文の**target**らしさが増加する

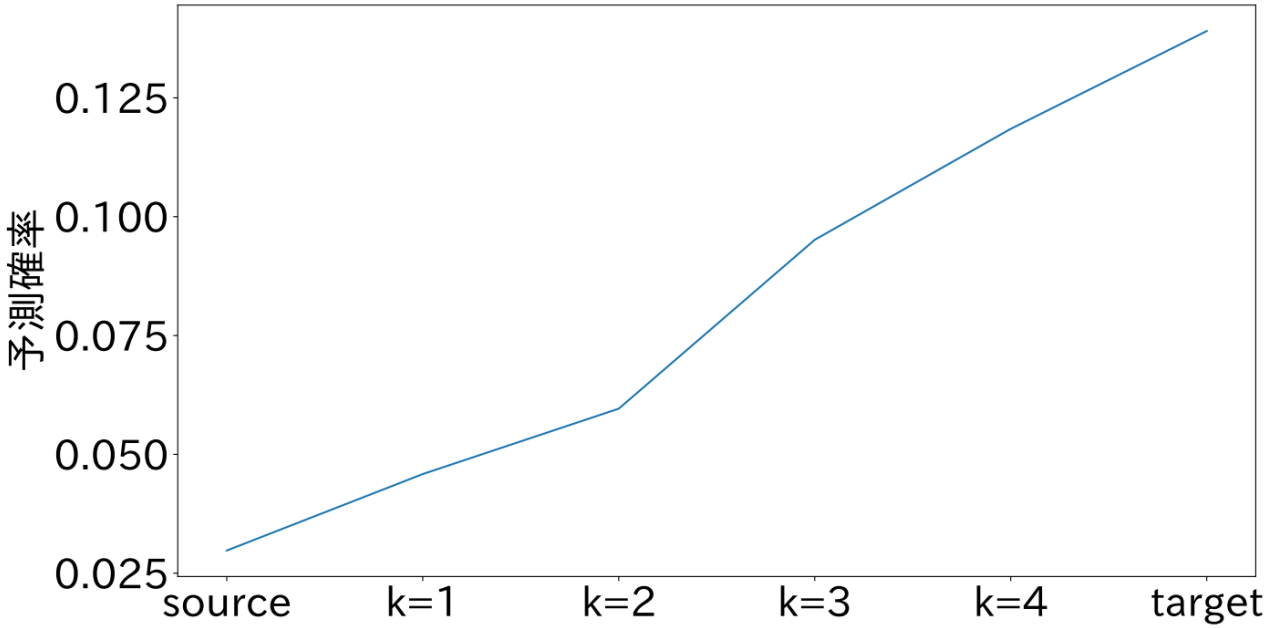


図
生成文が**target**と
予測される確率

中間人物の特徴を反映した応答文生成が可能である

まとめ

結果
ペルソナ文と**ユーザ識別子**の双方を扱えるVAEを用いて
内挿性と**制御性**のある人物埋め込みの学習ができた

今後の課題

・より詳細な内挿性の分析

例）「辛いもの好き」な人物と「甘いもの好き」な人物の中間人物は？

・別のデータセットを用いた規模のより大きな学習

→TwitterなどのSNSから収集したユーザ識別子の付属したデータセット

参考文献

[Li et al., 2016] Jiwei Li et al. A persona-based neural conversation model. In Proc. ACL, pages 994–1003, 2016.

[Zhang et al., 2018] Saizheng Zhang et al., Personalizing Dialogue Agents: I have a dog, do you have pets too? In Proc. ACL, pages 2204–2213, 2018.

[Devlin et al., 2019] Jacob Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. IJCNLP, pages 4171–4186, 2019.

[Vaswani et al., 2017] Ashish Vaswani et al. Attention is all you need. In Proc. NeurIPS, volume 30, 2017.

[Wu et al., 2023] Shih-Lun Wu and Yi-Hsuan Yang. MuseMorphose: Full-song and fine-grained piano music style transfer with one Transformer VAE. TASLP, 2023.

[Fu et al., 2019] Hao Fu et al. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In Proc. NAACL, pages 240–250, 2019.

[Kingma et al., 2016] Durk P Kingma et al. Improved variational inference with inverse autoregressive flow. In Proc. NeurIPS, volume 29, 2016.

[Sugiyama et al., 2023] Hiroaki Sugiyama et al. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. In Proc. SLT, pages 685–691, 2023.

[Sugiyama et al., 2020]杉山 弘晃 et al. Transformer encoder-decoder モ デルによる趣味雑談システムの構築. 人工知能学 会研究会資料 言語・音声理解と対話処理研究会, 90:24, 2020.

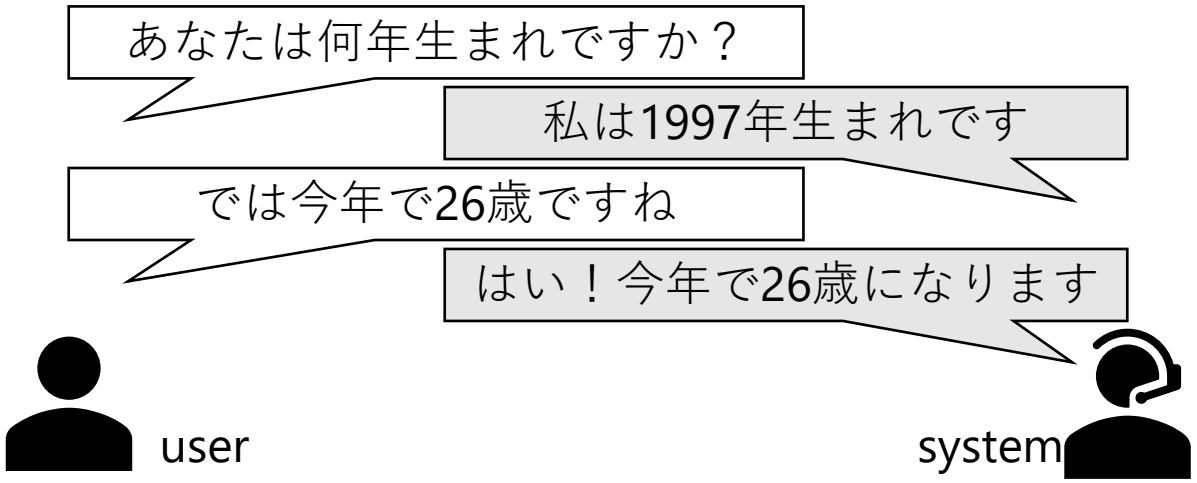
人物情報を反映した応答生成モデルの概説

与えられた人物情報に基づき応答文を生成するモデル

　人物情報：ペルソナ文，ユーザ識別子

機械学習による実現

→「人物情報及び対話文脈」と「応答文」の対応関係を学習する



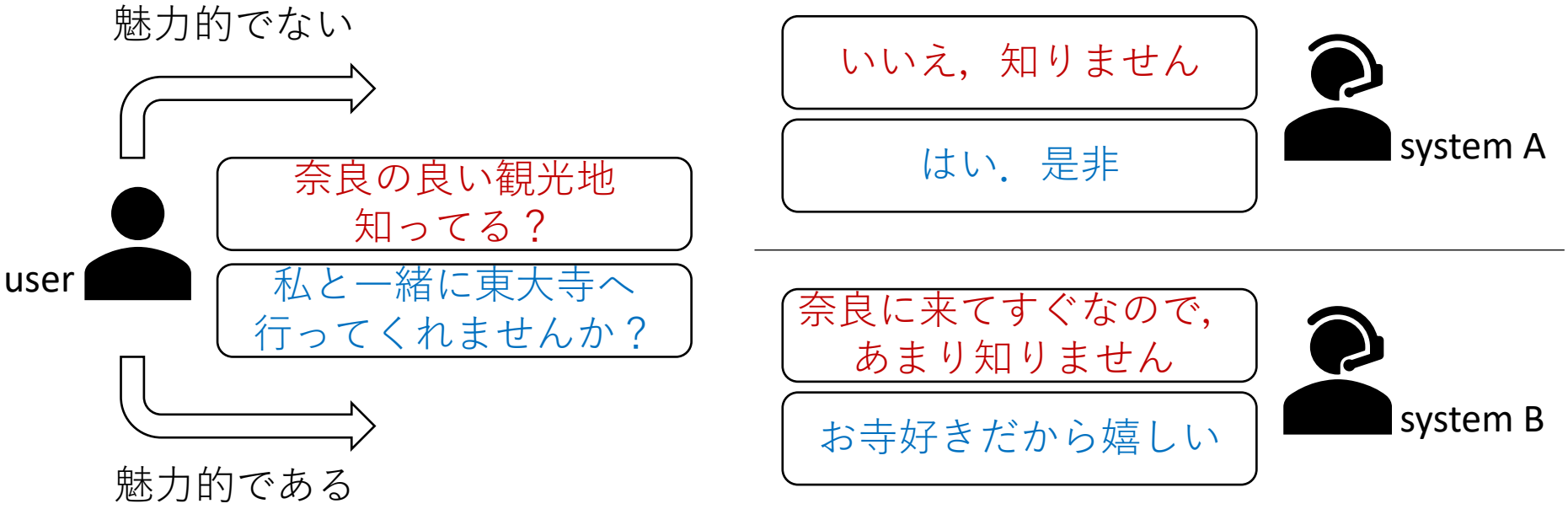
「年齢が26歳」という情報が与えられた対話システムの対話例

人物情報を反映した応答生成モデルの貢献

人物情報を含んだ一貫性のある対話を実現する利点

— 対話システムの**信頼性**を高め，**魅力的**にする [Miyazaki et al., 2021]

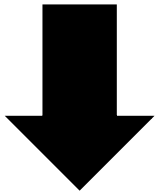
— 生成文が対話相手へ与える**魅力度**を高める [Zhang et al., 2018]



人物情報を含まない一般的な発話とユーザ情報を含んだ発話の例

人物情報を反映した応答生成モデルの実現方法

用いる人物情報によって分けられ，それぞれ利点がある



		内挿性	制御性
ユーザ識別子	→	○	×
ペルソナ文	→	×	○

本研究では…

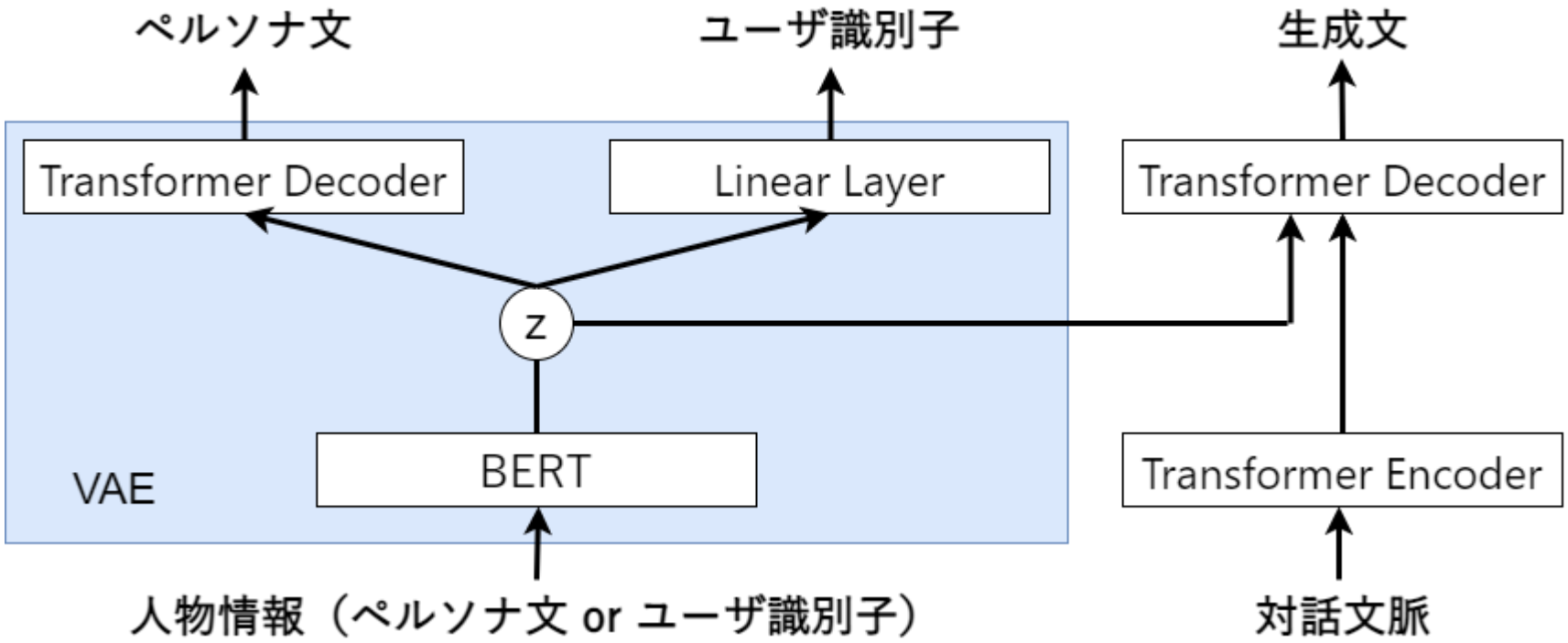
内挿性と**制御性**のある応答生成モデルを実現したい

⇓

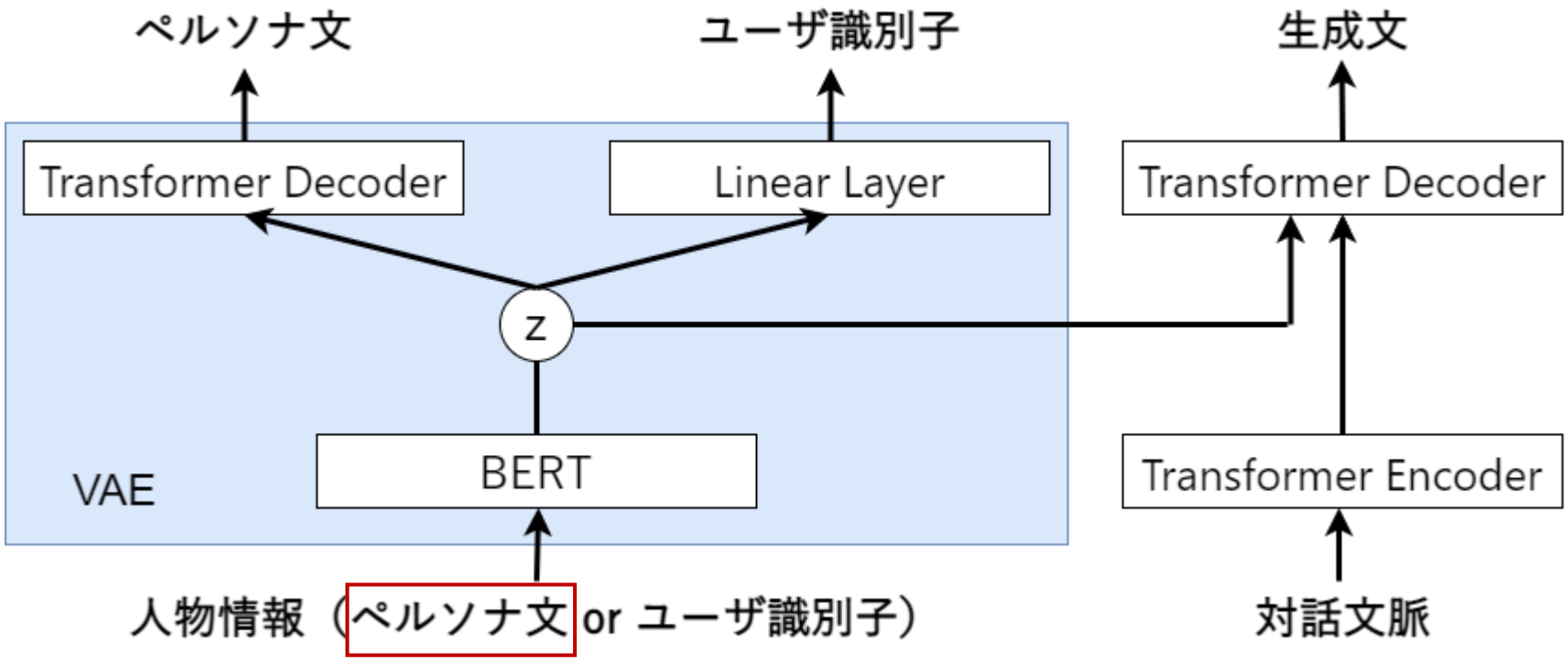
ユーザ識別子とペルソナ文の双方を扱える
応答生成モデルの提案

アプローチ

ペルソナ文とユーザ識別子の二つを利用可能なVAEを用いて
内挿性と制御性のある人物埋め込みを学習する



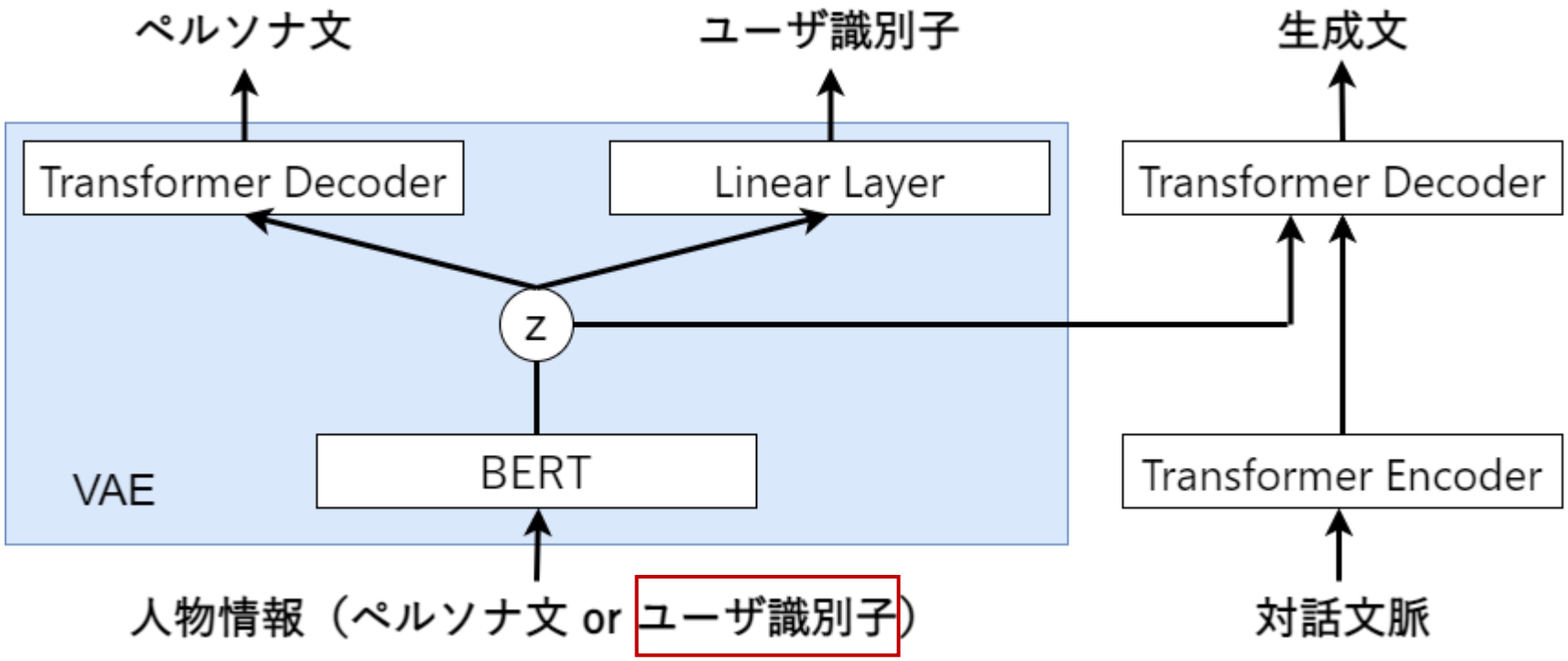
アプローチ



ペルソナ文：対話中の人物の特徴を自然言語で記述した文

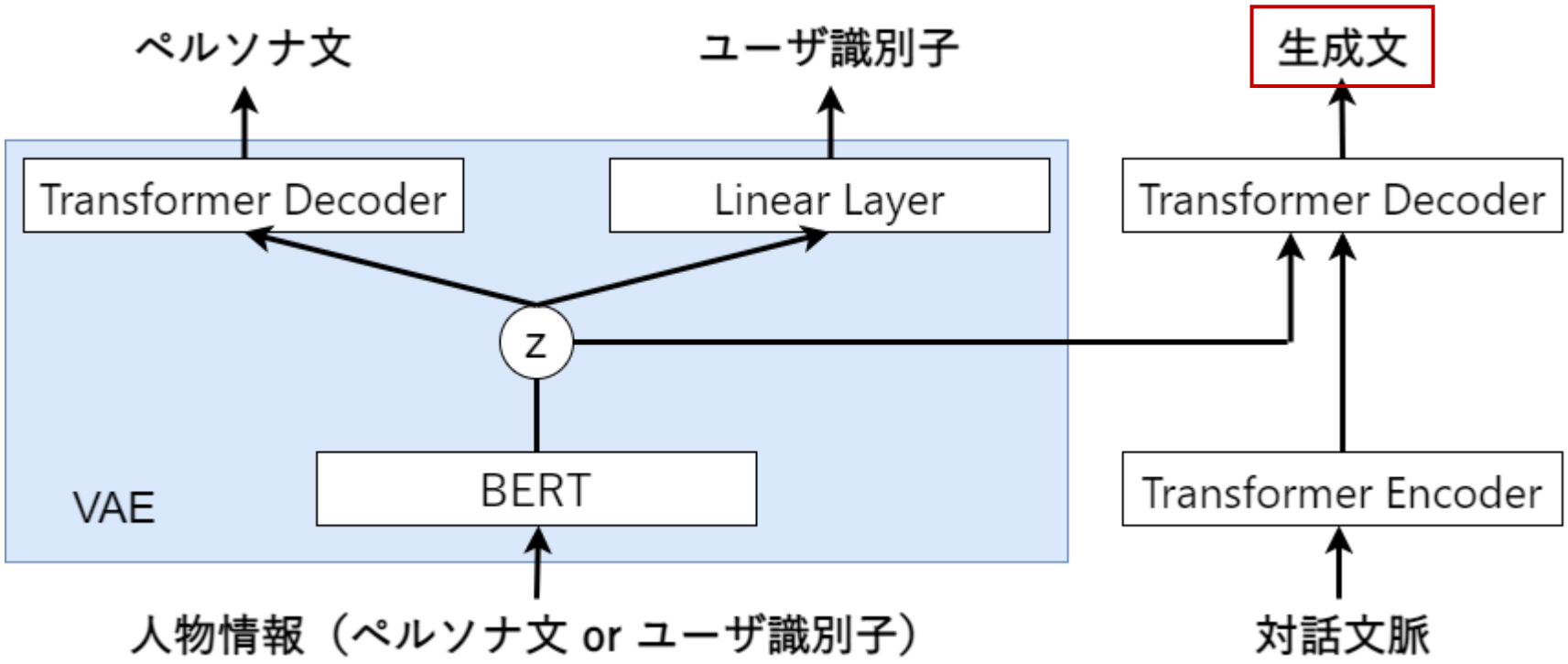
例）「私は大阪出身です」

アプローチ



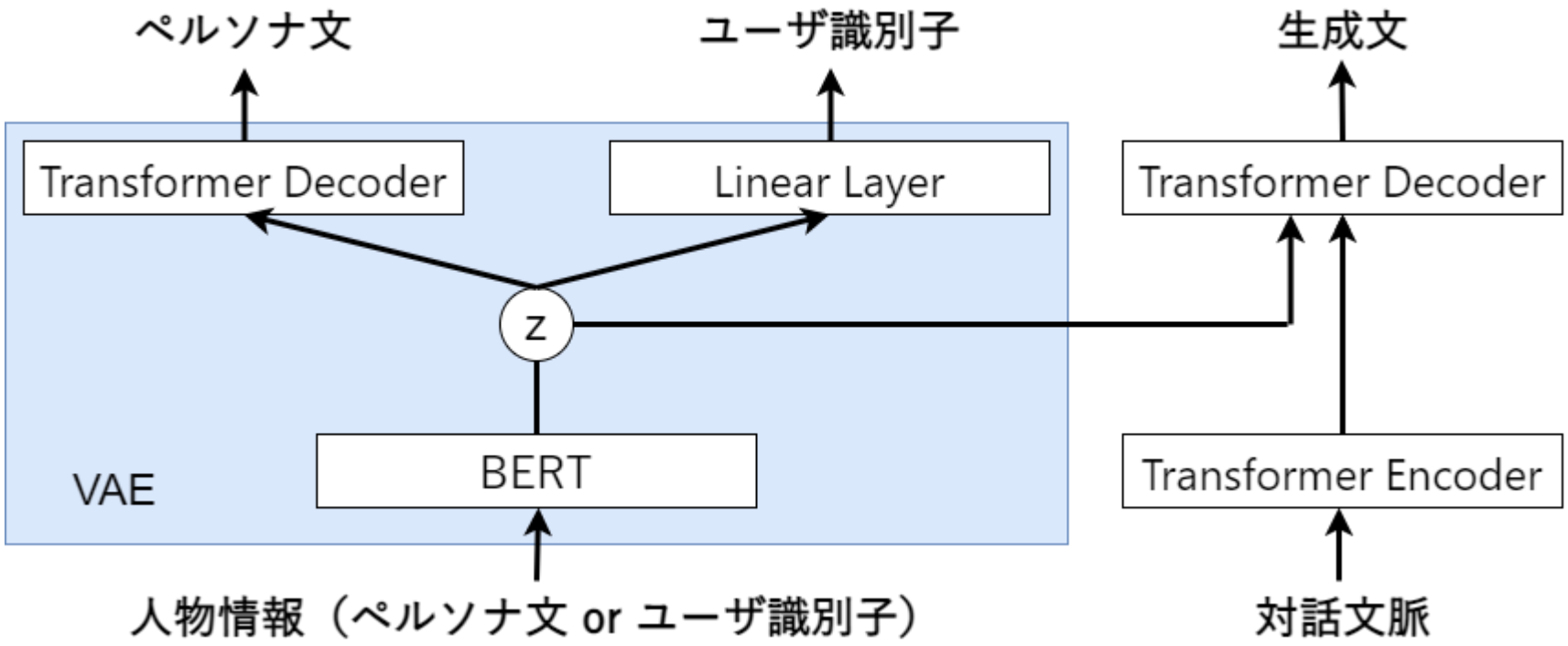
ユーザ識別子：対話に登場する人物に割り振られた識別子

アプローチ



人物埋め込みと対話文脈を用いた応答文生成

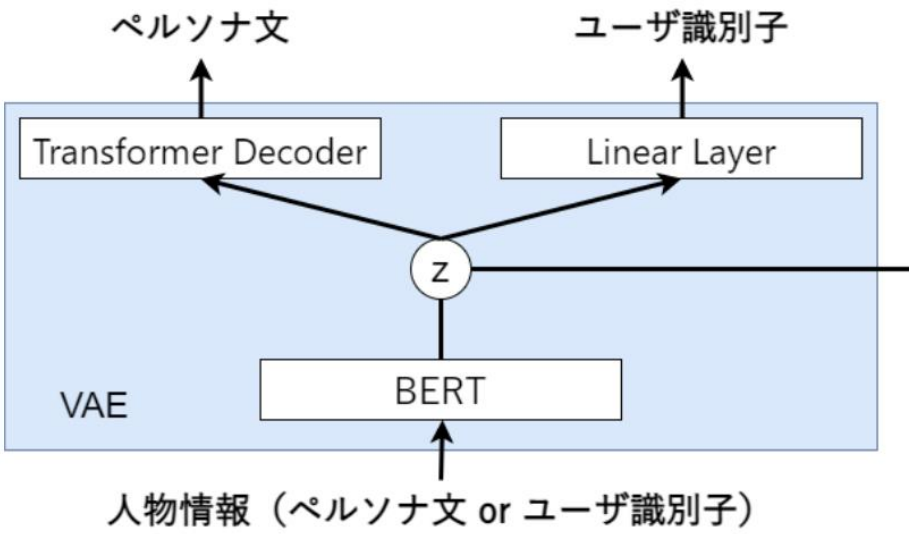
提案手法 ― モデル全体図 ―

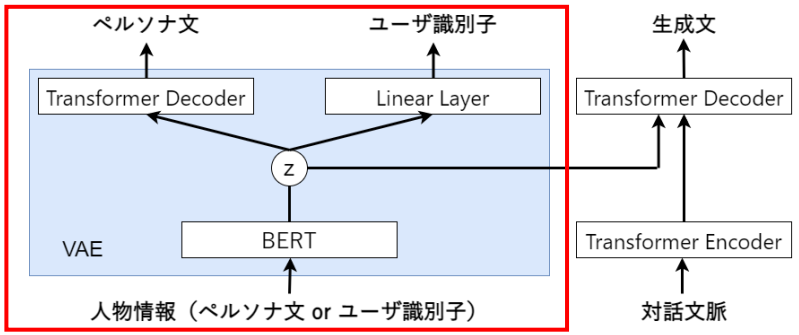


人物埋め込みzを獲得するVAE（左）と
応答文を生成する応答生成モデル（右）

提案手法 ― VAEによる人物埋め込みモデル ―

与えられた人物情報から人物埋め込みを計算



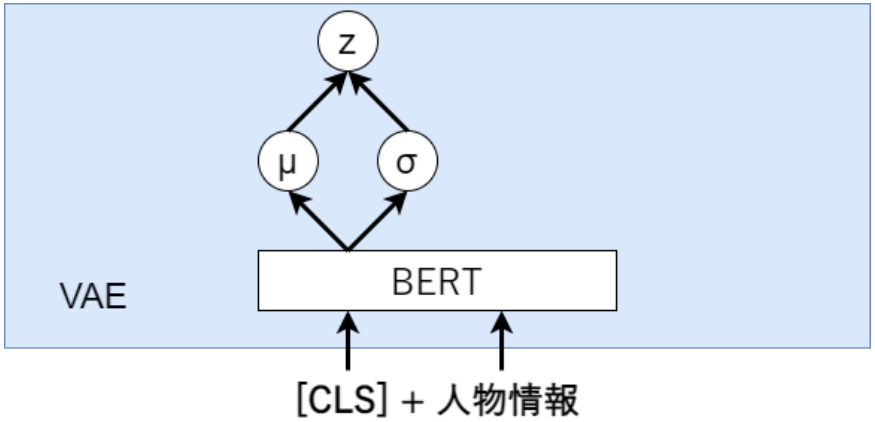


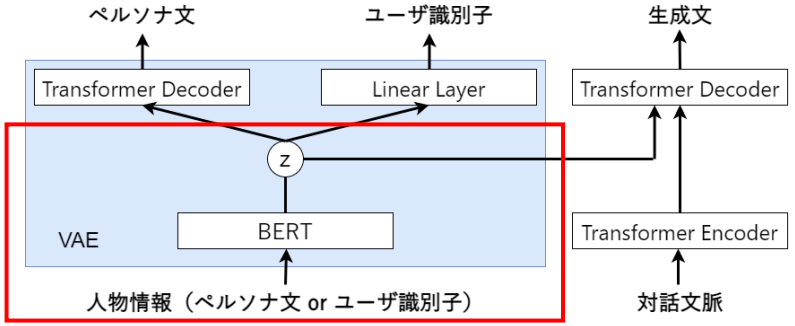
・Encoder：人物情報に応じた人物埋め込みzを計算

・Decoder：人物埋め込みzから人物情報を予測

提案手法 ― VAEによる人物埋め込みモデル ―

人物情報に応じた人物埋め込みzを計算（Encoder）



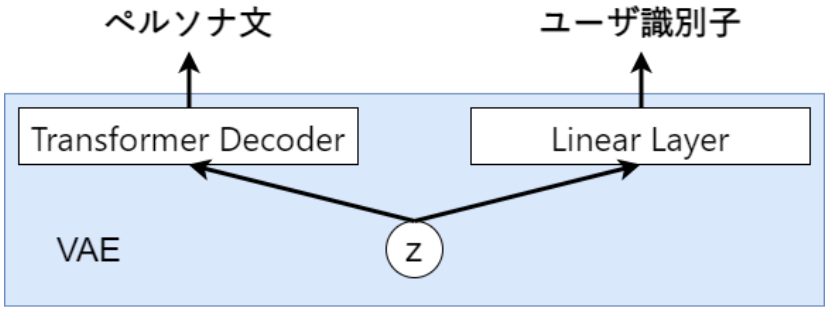


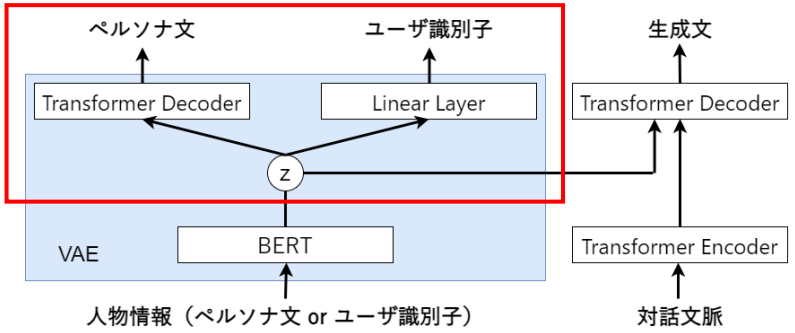
- **BERT**[Devlin et al., 2019] ベース
- [CLS]トークンと人物情報を連結したものを入力

→ ユーザ識別子 or ペルソナ文

提案手法 ― VAEによる人物埋め込みモデル ―

人物埋め込みzから人物情報を予測（Decoder）





- **Transformer Decoder**, **Linear Layer**
- 人物埋め込みzから人物情報を予測

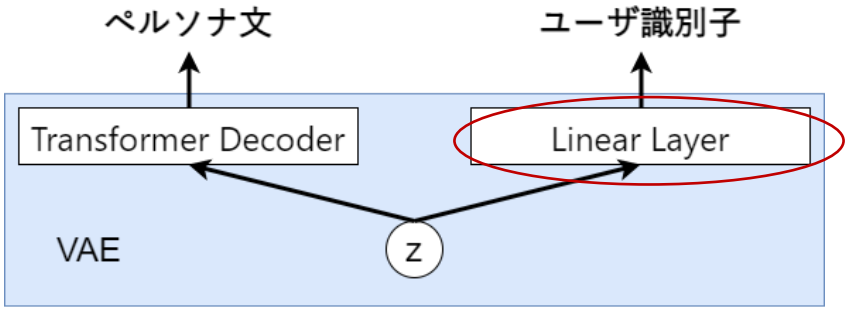
  → ユーザ識別子：zからユーザ識別子を予測

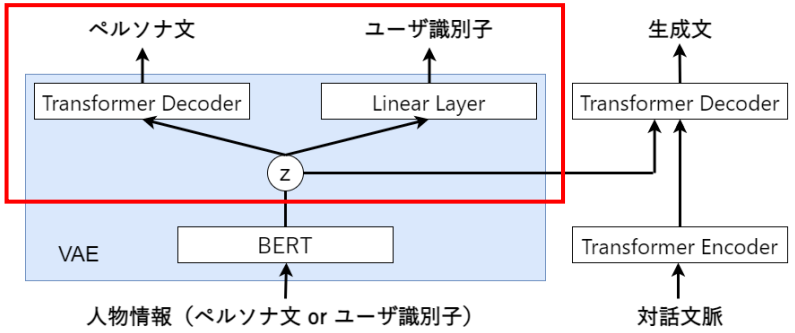
  → ペルソナ文：zからペルソナ文を復元

  ＋疑似的に付与したユーザ識別子の予測

提案手法 ― VAEによる人物埋め込みモデル ―

人物埋め込みzから人物情報を予測（Decoder）





- **Transformer Decoder**, **Linear Layer**
- 人物埋め込みzから人物情報を予測

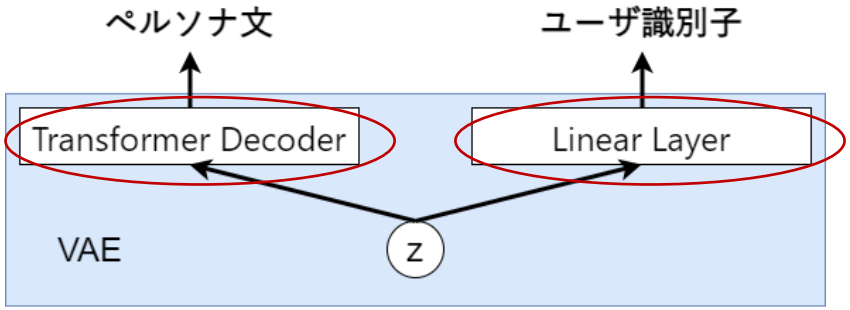
  → **ユーザ識別子**：zからユーザ識別子を予測

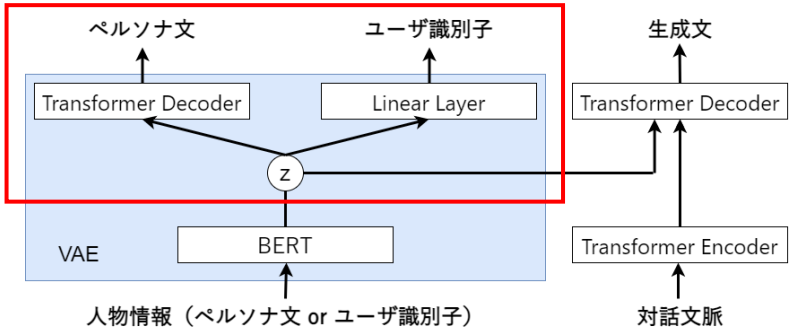
  → ペルソナ文：zからペルソナ文を復元

  ＋疑似的に付与したユーザ識別子の予測

提案手法 ― VAEによる人物埋め込みモデル ―

人物埋め込みzから人物情報を予測（Decoder）





- **Transformer Decoder**, **Linear Layer**
- 人物埋め込みzから人物情報を予測

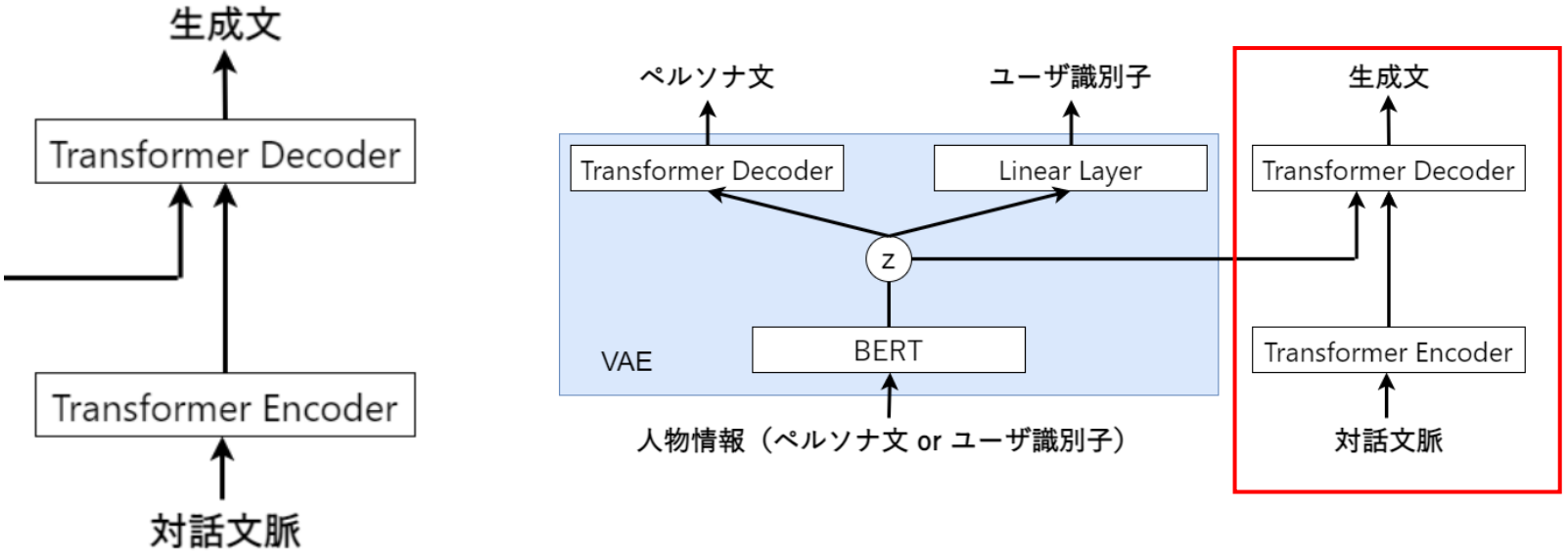
  → ユーザ識別子：zからユーザ識別子を予測

  → ペルソナ文：zからペルソナ文を復元

  ＋疑似的に付与したユーザ識別子の予測

提案手法 ― 応答生成モデル ―

人物埋め込みzと対話文脈を受け取り，応答生成を行う



・Transformer Encoder-decoder[Vaswani et al., 2017]ベース
→人物埋め込みzと対話文脈から人物の特徴を反映した応答文を生成

アプローチ

本研究では…

内挿性と**制御性**のある応答生成モデルを実現したい

⇓

ユーザ識別子とペルソナ文の双方を扱える
応答生成モデルの提案

具体的に

- ユーザ識別子とペルソナ文の双方を扱えるVAEを用いてある人物を表現する人物埋め込みを計算
  → 構造化された埋め込み空間の構築
  → 制御が容易な人物埋め込みの実現

提案手法 ― 損失関数 ―

・VAEによる人物埋め込みモデル

$$\mathcal{L}_{\mathbf{VAE}} = -\mathbb{E}[\log p_\theta(P|z_p)] + \beta KL(q_\phi(z_p|P)||p(z_p))$$

→ 与えられた**人物情報の復元度合い**に基づく損失関数（第一項）

→ Encoderの出力と**事前分布との差異**に基づく損失関数（第二項）

※ Cyclical Annealing[Fu et al., 2019]とKL項へ最低値を導入する手法[Kingma et al., 2016]を採用し第二項を $\beta_{cyc} \max(KL(q_\phi(z_p|P)||p(z_p)), \lambda)$ に変更

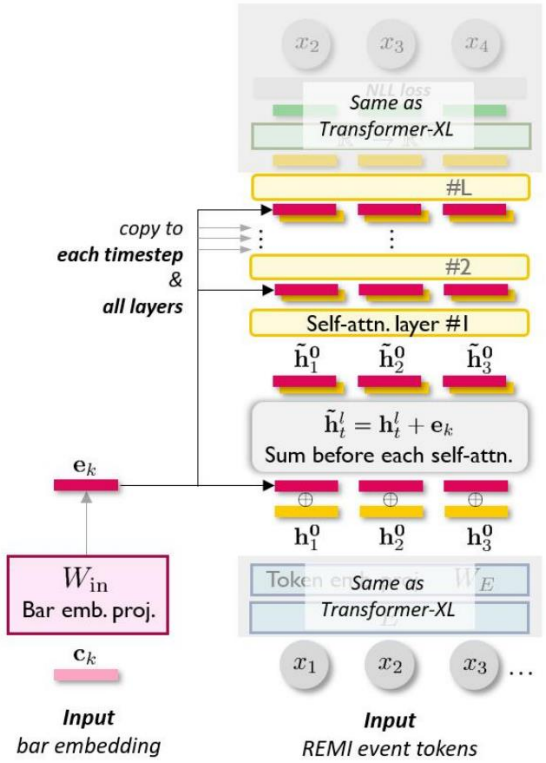
・応答生成モデル

$$\mathcal{L}_{\mathbf{D}} = -\mathbb{E}[\log(p_\theta(Y|X, z_p))]$$

→ **応答生成の精度**に基づく損失関数

In-attention構造

In-attention構造[Wu et al., 2021]



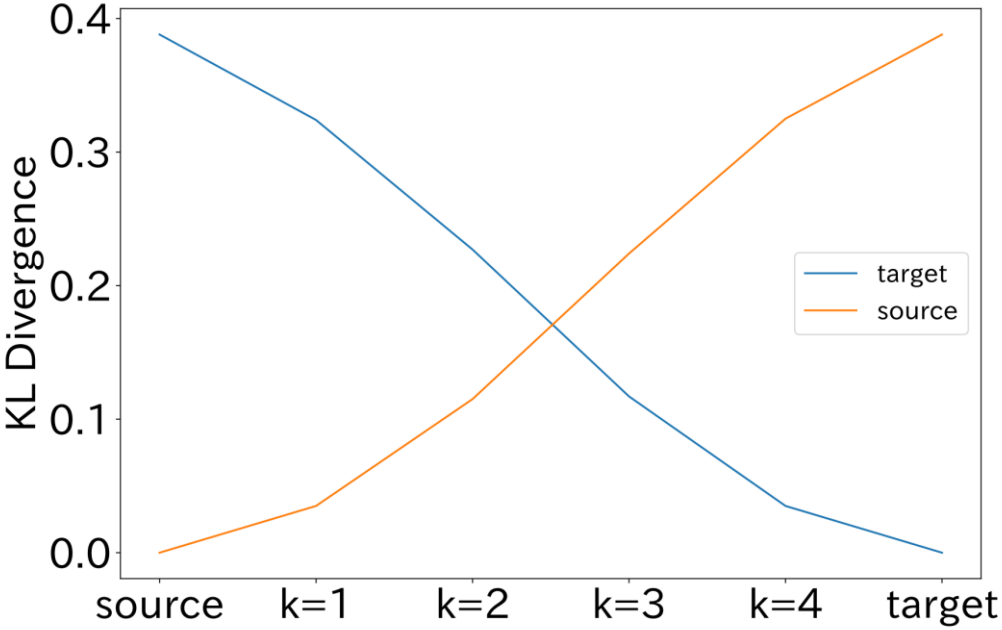
Decoderにおいて，最終ブロックを除く各ブロックのSelf-attention層へ，前層の隠れ状態と潜在変数zを加算したものを入力する．全てのタイムステップtで同様の操作を行う

$$\tilde{\boldsymbol{h}}_t^l = \boldsymbol{h}_t^l + \boldsymbol{e}_k\,, \quad \forall l \in \{0, \ldots, L-1\}\,;$$
$$\boldsymbol{h}_t^{l+1} = \text{Transformer\_self-attention\_layer}(\tilde{\boldsymbol{h}}_t^l)$$

In-attention構造の図．図はWuらの論文より引用

結果 2
— ペルソナ文から得られた人物埋め込みは内挿性を持つか —

実験結果（補足）



kが大きくなるとKL Divが段階的に上昇，低下

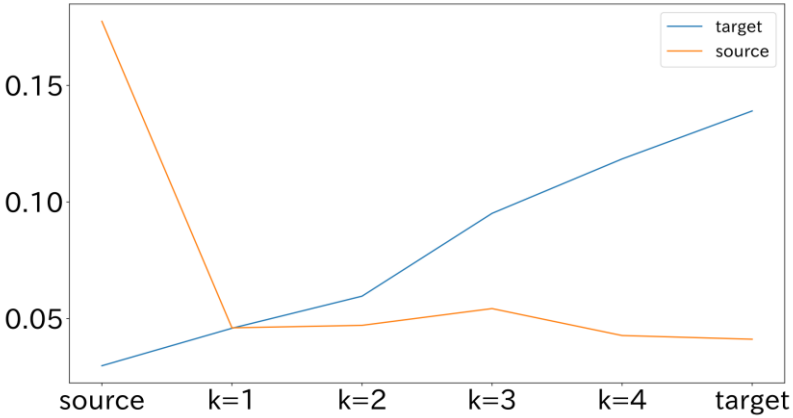
↓

ペルソナ文から得られた人物埋め込みは内挿性を持つように学習できた

図：**source**，**target**と**中間の人物**へそれぞれ与えられる人物埋め込み間の**KL Div**．評価のため，**中間**の人物の代わりに**source**及び**target**を用いた場合の**KL Div**の値もプロットした

結果 4
— ペルソナ文から得られた生成文は内挿性を持つか —

実験結果（補足）



図：**中間の人物**の人物埋め込みを用いて生成された応答文が**source**，**target**と予測される確率．評価のため，**source**及び**target**を用いた場合の値もプロットした

sourceの予測確率の遷移が**target**を左右反転した形にはならなかった

・分類を行う生成文は対話文脈と人物埋め込みを用いて生成される

・**source**は正解の応答文に紐づいた人物，**target**は正解の応答文に登場しない人物

→左端と右端の値が一致しない理由として，**source**について対話の中で言及されており，応答生成に用いられる人物の情報量が異なるということが考えられる

→k=1で急激に減少して横這いになったことも，この辺りの情報量の違いが影響しているのではないか