A Contrastive Framework for Learning Sentence Representations from Pairwise and Triple-wise Perspective in Angular Space

Yuhao Zhang, Hongji Zhu, Yongliang Wang,
Nan Xu, Xiaobo Li, BinQiang Zhao

ACL'22

最先端NLP勉強会 2022

名大 武田・笹野研 D2 山田 康輔

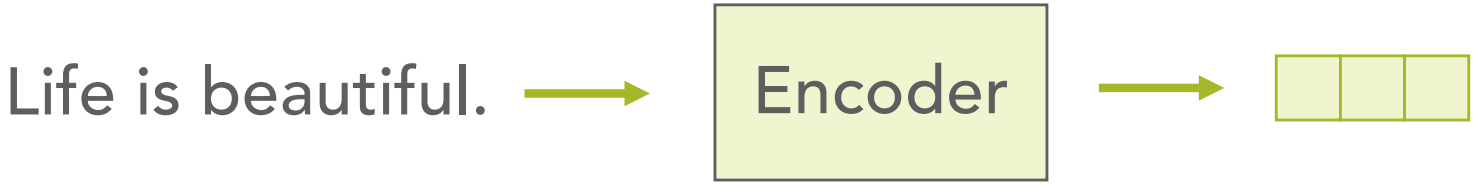
論文URL: https://aclanthology.org/2022.acl-long.336/

概要

- 情報検索などに利用する文埋め込みを構成する研究
- 現在，対照学習 (Contrastive learning) を用いて，BERTをファインチューニングする手法が主流
  - これらの多くは，正例・負例の選択に着目しており，目的関数にはあまり注意が払われていない
- 文ペアの識別力を高め，含意関係をモデル化するArcSCEを提案
- 文埋め込み評価のベンチマークであるSTSやSentEvalで従来手法より優れた性能を発揮

導入: 文埋め込みと評価タスク

- 文埋め込み: 文をベクトルとして表現



- 文埋め込みの代表的な評価タスク
  - Semantic Textual Similarity (STS) タスク [1]
    - 文埋め込みペアのコサイン類似度を文ペアの意味類似度とし，人手でつけたラベルとスピアマン相関係数を算出して評価
  - SentEval [2]
    - 感情極性分類など7タスクに対し，文埋め込みを入力とするロジスティック回帰分類器を学習し，性能を評価

[1] Reimers et al.: Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity (COLING'16)
[2] Cer et al.: SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation (SemEval'17)

導入: BERTから獲得した文埋め込み

- BERTから単純に獲得した文埋め込みは文ペア類似度を測るには不向き [3]

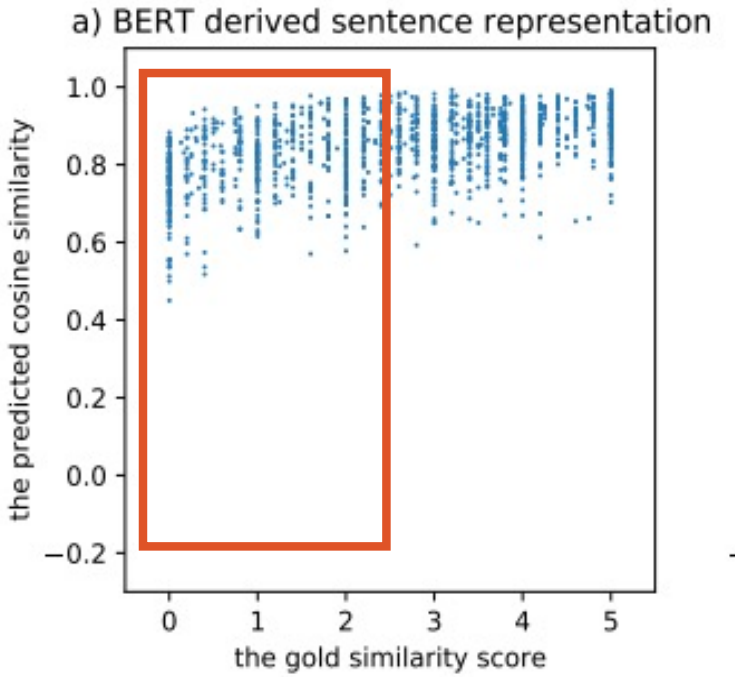
※InferSentやUSEの概要は付録へ

STS12-16,b, SICK-Rの平均スコア

Glove	InferSent	USE	$BERT_{ave}$	$BERT_{cls}$
61.32	65.01	71.22	54.81	29.19

- 多くの文ペアの類似度が高い [4]
  - 類似していない文ペアでさえ高い類似度

$BERT_{ave}$による STSデータセットの スコア別類似度



[3] Reimers and Gurevych: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (EMNLP'19)
[4] Yan et al.: ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer (ACL'21)

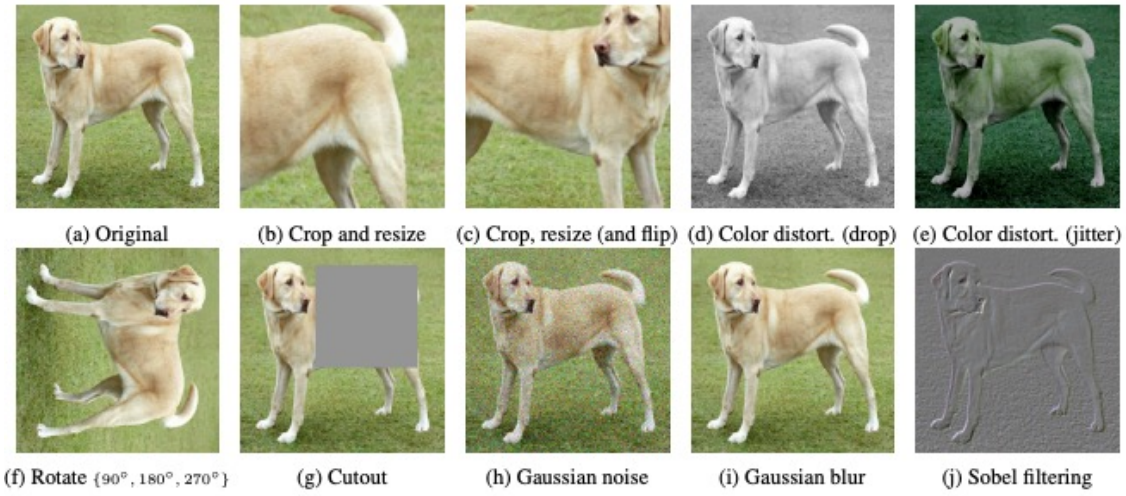
導入: BERT以降の文埋め込み

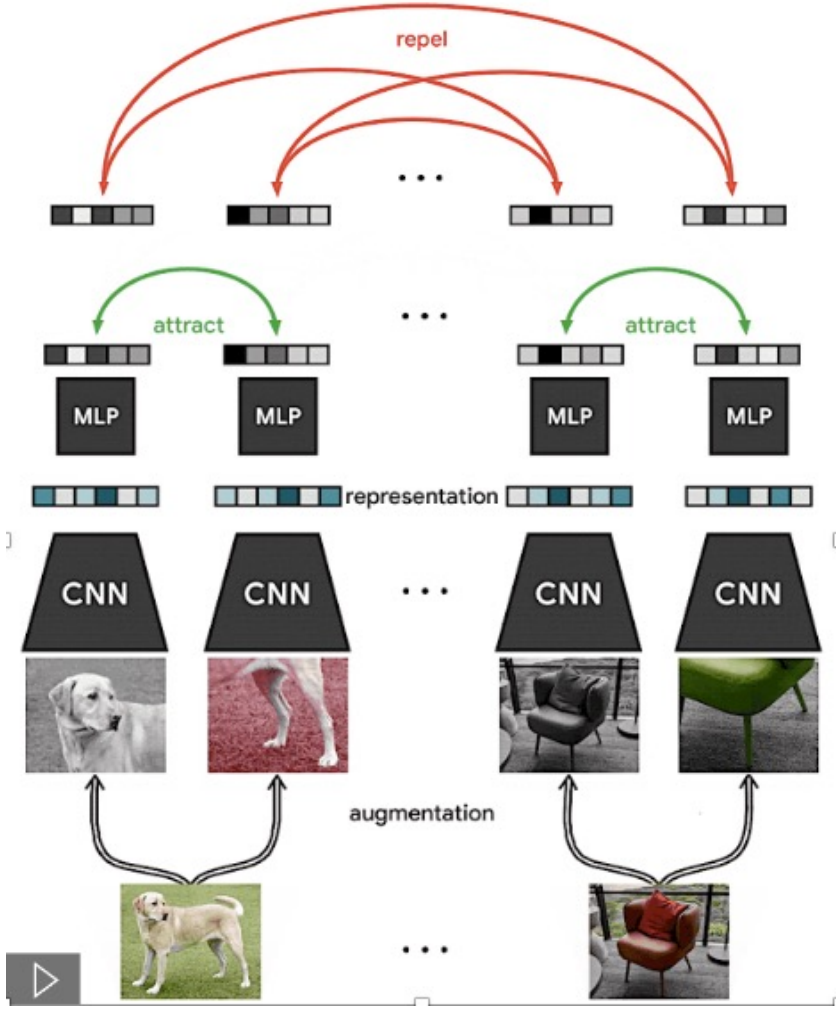
- Supervised
  - Sentence-BERT: Siamese networkでNLIを学習
- Post-processing
  - BERT-flow: ガウス過程潜在変数モデル
  - BERT-whitening: 埋め込みの白色化
- Unsupervised
  - IS-BERT: 文埋め込みとn-gram埋め込みの相互情報量を最大化
  - BERT-CT: パラメータをシェアしない2つの同じモデルを用いて, 同じ文に対する文埋め込み同士の内積を最大化
  - ConSERT: 元の文に対して単語や特徴量を削除したものなどを正例として対照学習
  - SimCSE-unsup: 同じ文に対して異なるdropoutマスクを適用して作成した2つの文埋め込みを正例ペアとして対照学習

残りはスライド最後の付録へ

導入: 対照学習 (Contrastive Learning)

- 似たデータは似た埋め込みに,
  異なるデータは異なる埋め込みに
- SimCLR [5]
  - 画像分類タスク
  - ResNetベースモデル
  - データ拡張戦略
    - CropとColorの組み合わせが最良



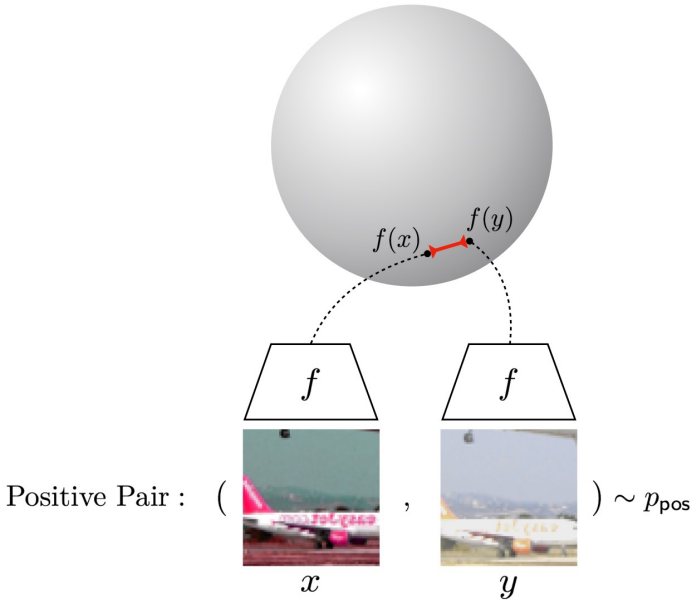


図は[6]より

[5] Chen et al.: A Simple Framework for Contrastive Learning of Visual Representations (ICML'20)
[6] Google AI Blog: Advancing Self-Supervised and Semi-Supervised Learning with SimCLR

導入: 対照学習に関連した2つの指標 [7]

- 対照学習の気持ち
  - 似ているデータは近くに配置 / 関係ないデータ遠くに配置
- 2つの指標
  - アライメント: 正例ペア同士は近くにマップしてほしい
  - 一様性: データ全体が一様に分布してほしい

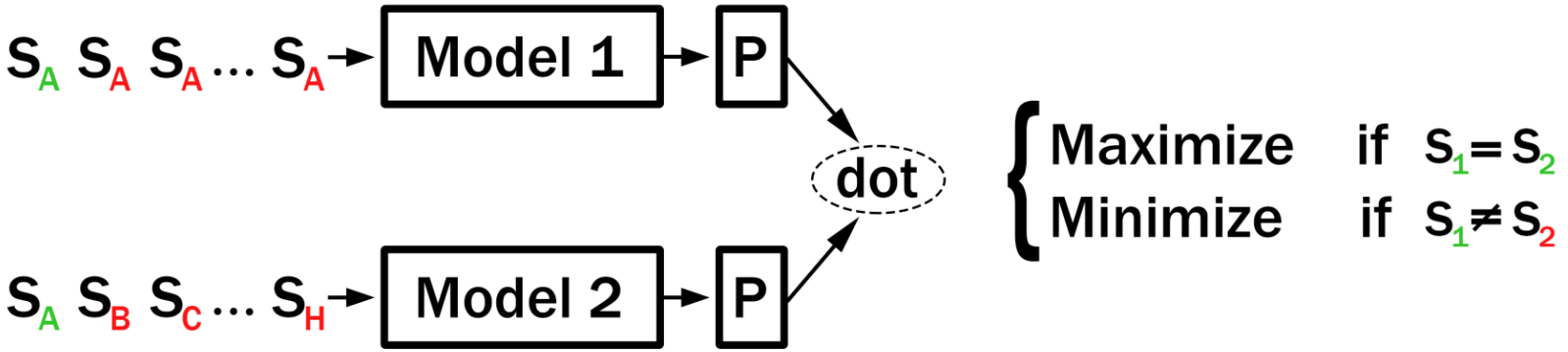


$$\ell_{\text{align}} = \mathop{\mathbb{E}}_{(x,x^+)\sim p_{\text{pos}}} \left\| f(x) - f\left(x^+\right) \right\|^2 \qquad \ell_{\text{uniform}} = \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d}{\sim} p_{\text{data}}} e^{-2\|f(x)-f(y)\|^2}$$

[7] Wang and Isola: Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere (ICML'20)

関連研究: BERT-CT (Contrastive Tension) [8]

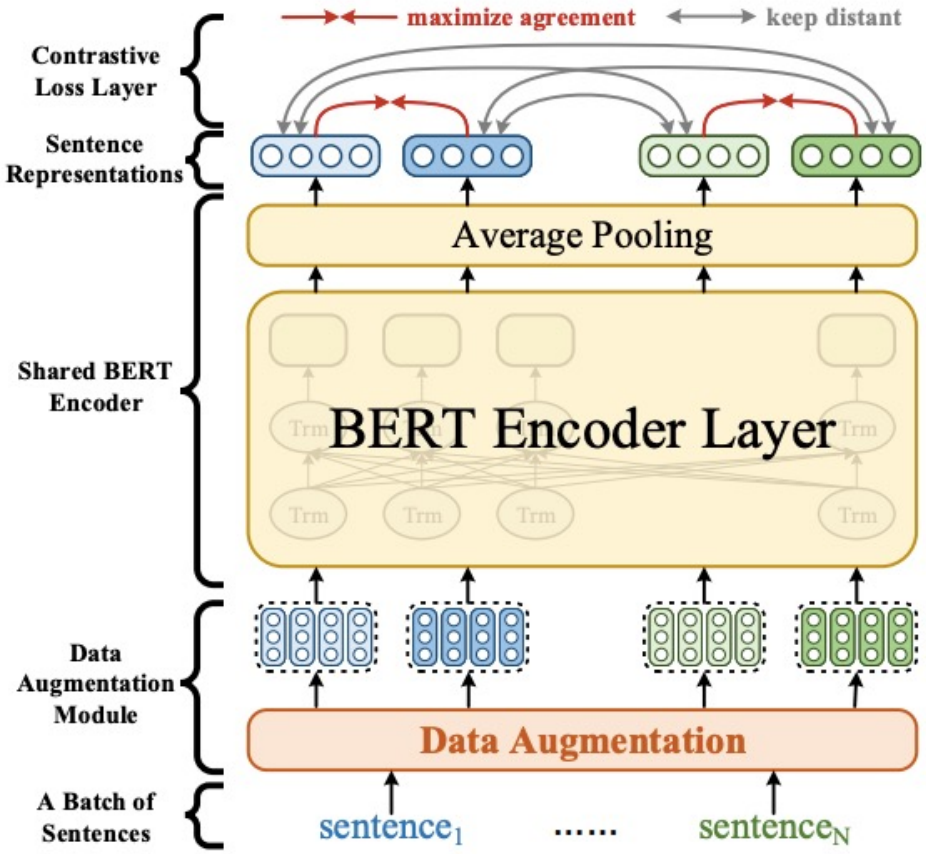
- BERTの構成性を修正するためにContrasitve Tensionを適用
  - 事前学習済みBERTは語彙に関する意味は学習済みであり，モデルに新たな情報は学習する必要はない
  - 独立した2つのエンコーダを利用し，同一文ならば内積を最大，異なる文ならば内積を最小となるように学習
  - 事前学習に使用したテキストで学習

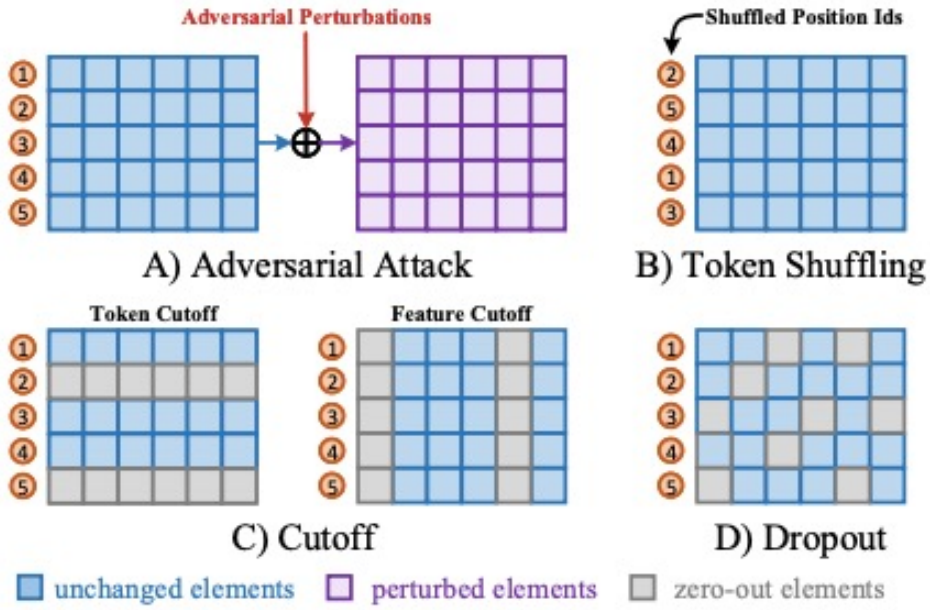


[8] Carlsson et al.: Semantic Re-tuning with Contrastive Tension (ICLR'21)

関連研究: ConSERT [4]

- 元の文に対してデータ拡張したものを正例として対照学習
  - 4種のデータ拡張
    A) Adversarial Attack, B) Token Shuffling, C) Cutoff, D) Dropout
  - STSのラベルなしテキストにて学習

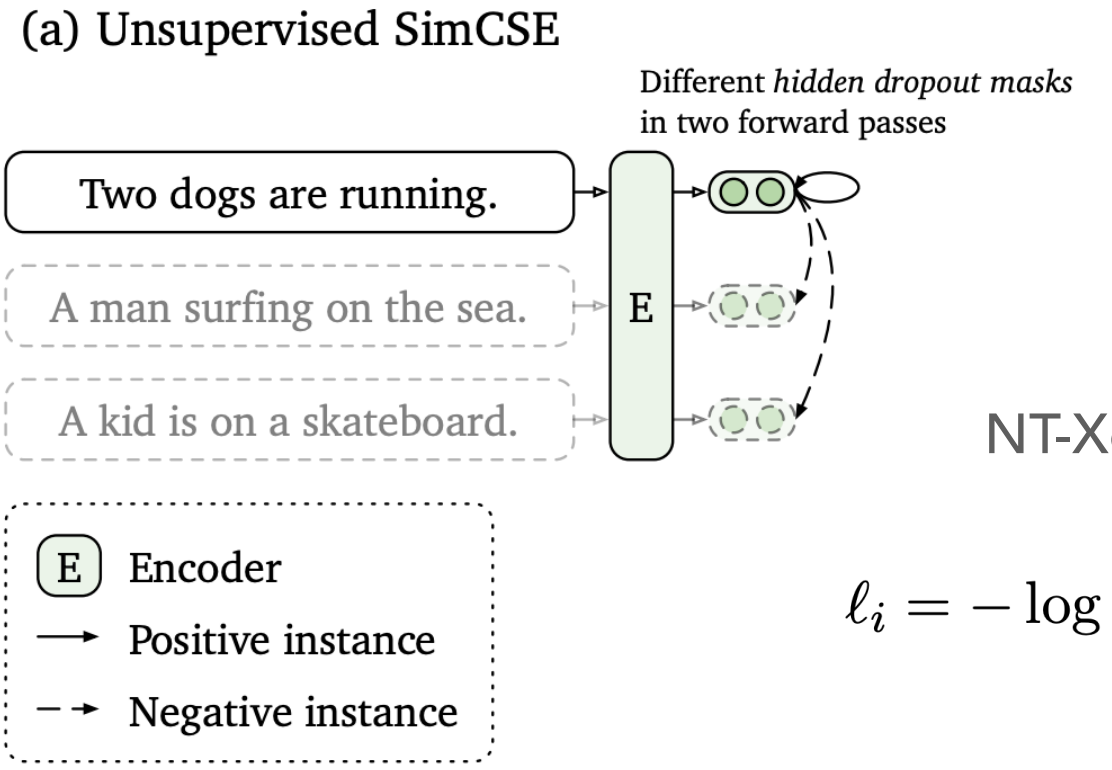




[4] Yan et al.: ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer (ACL'21)

関連研究: SimCSE-unsup [9]

- 同じ文に対して異なるdropout maskを適用して作成した2つの文埋め込みを正例ペアとして対照学習
  - 負例は同じミニバッチ内の異なる文の埋め込み
  - Wikipediaのデータで学習



NT-Xent損失関数

$$\ell_i = -\log \frac{e^{\mathrm{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z_i'})/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z_j'})/\tau}}$$

[9] Gao et al.: SimCSE: Simple Contrastive Learning of Sentence Embeddings (EMNLP'21)

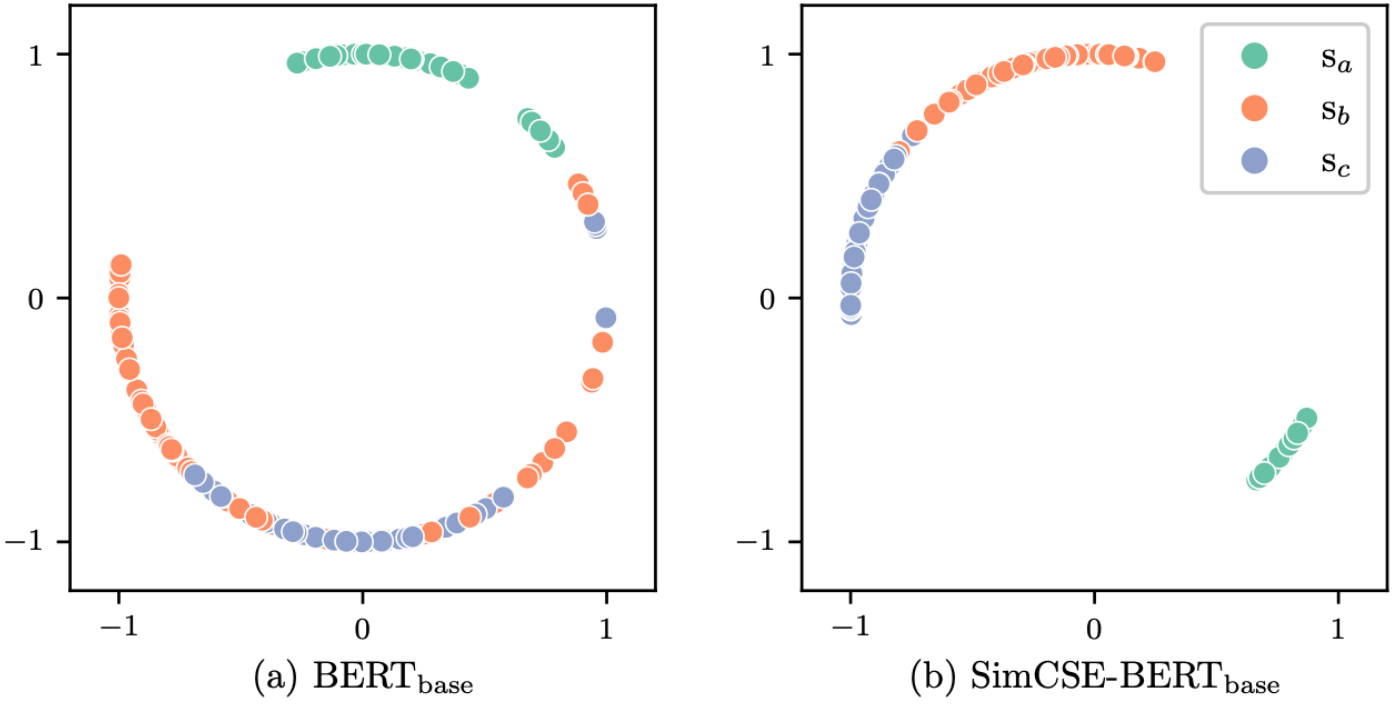
背景: BERTやSimCSEによる文埋め込み

- 異なるdropout maskによる文埋め込みの可視化

  S_a: He was born in Nazareth-Palestine, but immigrated to Lebanon with his parents and then to Jordan where he completed his primary education.

  S_b: He was born in Nazareth-Palestine, but immigrated to Lebanon with his parents.

  S_c: He was born in Nazareth-Palestine, but immigrated to Lebanon.



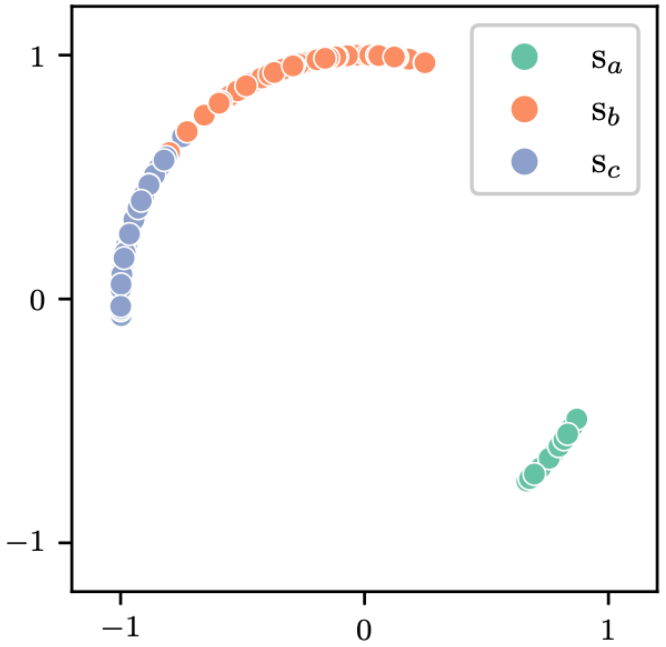
(a) $\text{BERT}_{\text{base}}$ (b) $\text{SimCSE-BERT}_{\text{base}}$

(a) BERTは文意を識別できていない

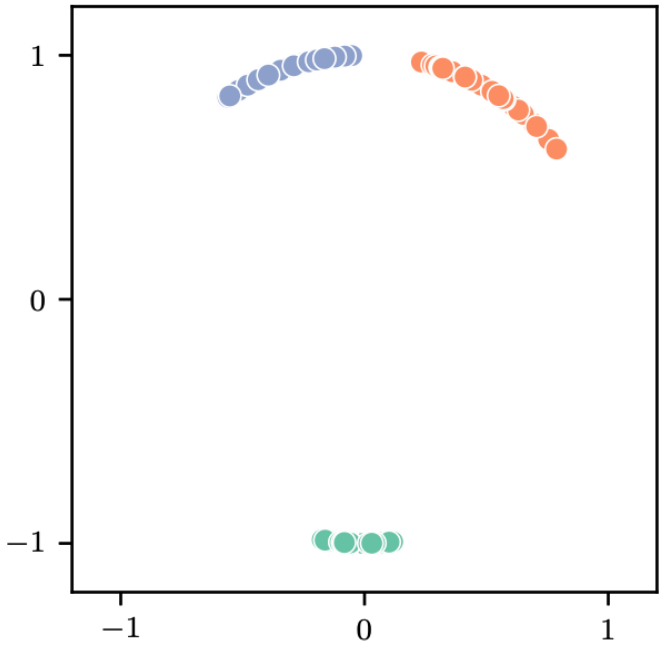
(b) SimCSEでさえS_bとS_cを完全に分離できていない

背景: SimCSEの課題とArcCSEのアプローチ

- SimCSEの課題
  - Softmaxベースの損失は識別力を獲得するには不十分
  - 文の関係をペアごとでモデル化しているのみで，類似の程度 (文間の意味的な順序) は十分に考慮できていない
- ArcCSEのアプローチ
  1. 角度ベースで決定境界にマージンを導入する損失関数を提案
  2. 文の3つ組に対し，含意関係を捉える自己教師あり学習法を提案



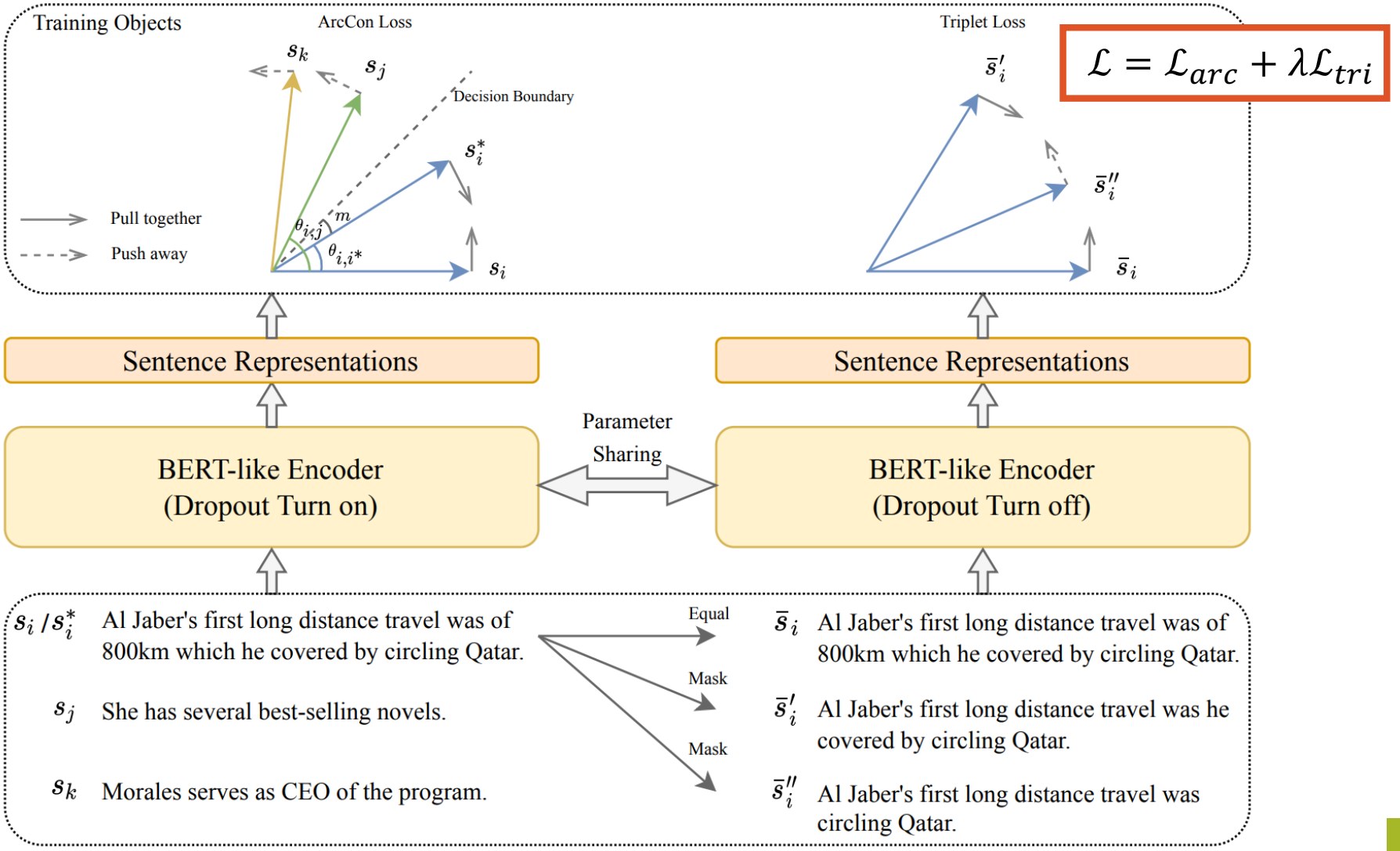
(b) SimCSE-BERT$_{\text{base}}$



(c) ArcCSE-BERT$_{\text{base}}$

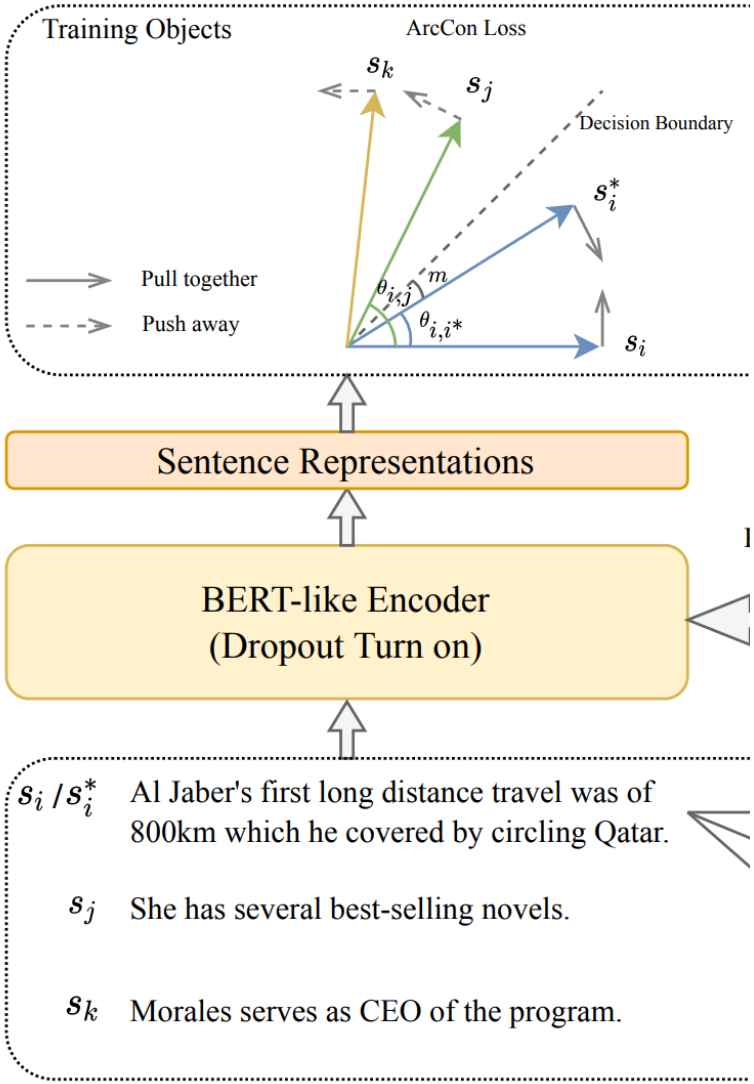
手法: ArcCSEの概要

1. 角度ベースで決定境界にマージンを導入する損失関数を提案
2. 文の3つ組に対し，含意関係を捉える自己教師あり学習法を提案



手法: ArcCon Loss

1. 角度ベースで決定境界にマージンを導入する損失関数を提案
2. 文の3つ組に対し，含意関係を捉える自己教師あり学習法を提案



- ベースはSimCSE
  - 正例・負例の作り方は同じ
  - 損失関数をNT-Xent LossからAngular Margin Contrastive Loss (ArcCon Loss)
    - ArcFace [10] にインスパイア

$$\mathcal{L}_{NT-Xent} = -\log \frac{e^{sim(h_i, h_i^*)/\tau}}{\sum_{j=1}^{n} e^{sim(h_i, h_j)/\tau}}$$

$$sim(h_i, h_j) = cos\theta_{i,j} = \frac{{h_i}^T h_j}{||h_i|| * ||h_j||}$$

$$\mathcal{L}_{NT-Xent} = -\log \frac{e^{\cos(\theta_{i,i^*})/\tau}}{\sum_{j=1}^{n} e^{\cos(\theta_{i,j})/\tau}}$$

マージン

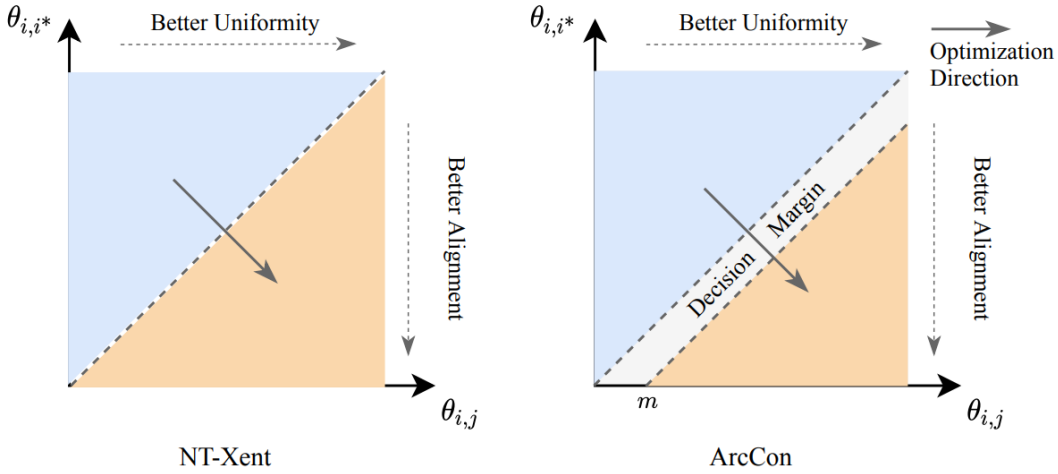
$$\mathcal{L}_{arc} = -\log \frac{e^{\cos(\theta_{i,i^*}+m)/\tau}}{e^{\cos(\theta_{i,i^*}+m)/\tau} + \sum_{j \neq i} e^{\cos(\theta_{i,j})/\tau}}$$

[10] Geng et al.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition (EMNLP'21)

手法: ArcCon Loss

1. 角度ベースで決定境界にマージンを導入する損失関数を提案
2. 文の3つ組に対し，含意関係を捉える自己教師あり学習法を提案

- 決定境界: $\theta_{i,i^*} + m = \theta_{i,j}$
- マージンを入れることで，対照学習をより機能させる



θ_{i,i^*}が小 → 正例ペアのアライメント向上
$\theta_{i,j}$ が大 → 一様性向上

- ベースはSimCSE
  - 正例・負例の作り方は同じ
  - 損失関数をNT-Xent LossからAngular Margin Contrastive Loss (ArcCon Loss)
    - ArcFace [10] にインスパイア

$$\mathcal{L}_{NT-Xent} = -\log \frac{e^{sim(h_i,h_i^*)/\tau}}{\sum_{j=1}^{n} e^{sim(h_i,h_j)/\tau}}$$

$$sim(h_i, h_j) = cos\theta_{i,j} = \frac{{h_i}^T h_j}{||h_i|| * ||h_j||}$$

$$\mathcal{L}_{NT-Xent} = -\log \frac{e^{\cos(\theta_{i,i^*})/\tau}}{\sum_{j=1}^{n} e^{\cos(\theta_{i,j})/\tau}}$$

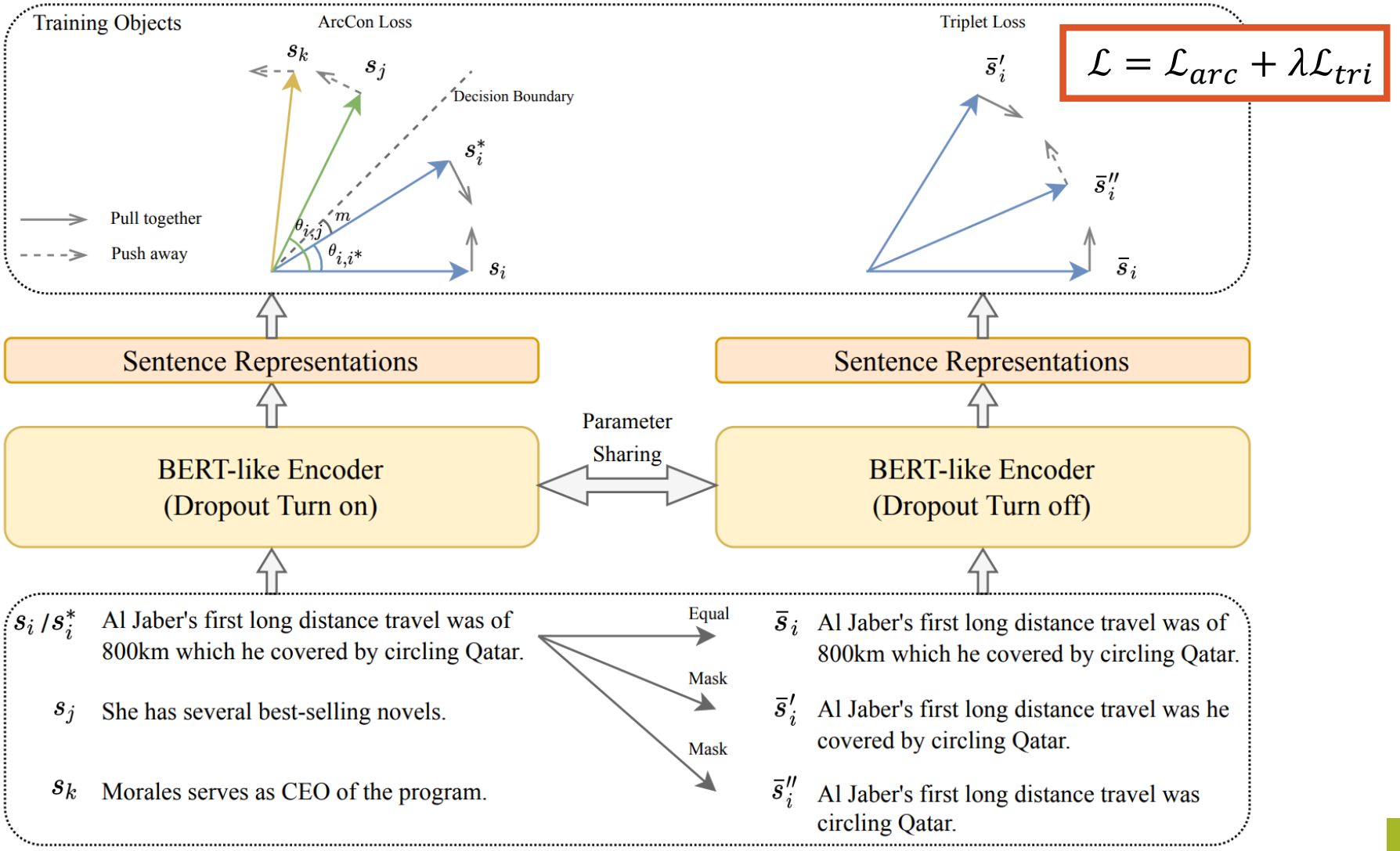
マージン

$$\mathcal{L}_{arc} = -\log \frac{e^{\cos(\theta_{i,i^*}+m)/\tau}}{e^{\cos(\theta_{i,i^*}+m)/\tau} + \sum_{j \neq i} e^{\cos(\theta_{i,j})/\tau}}$$

[10] Geng et al.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition (EMNLP'21)

手法: ArcCSEの概要

1. 角度ベースで決定境界にマージンを導入する損失関数を提案
2. 文の3つ組に対し，含意関係を捉える自己教師あり学習法を提案



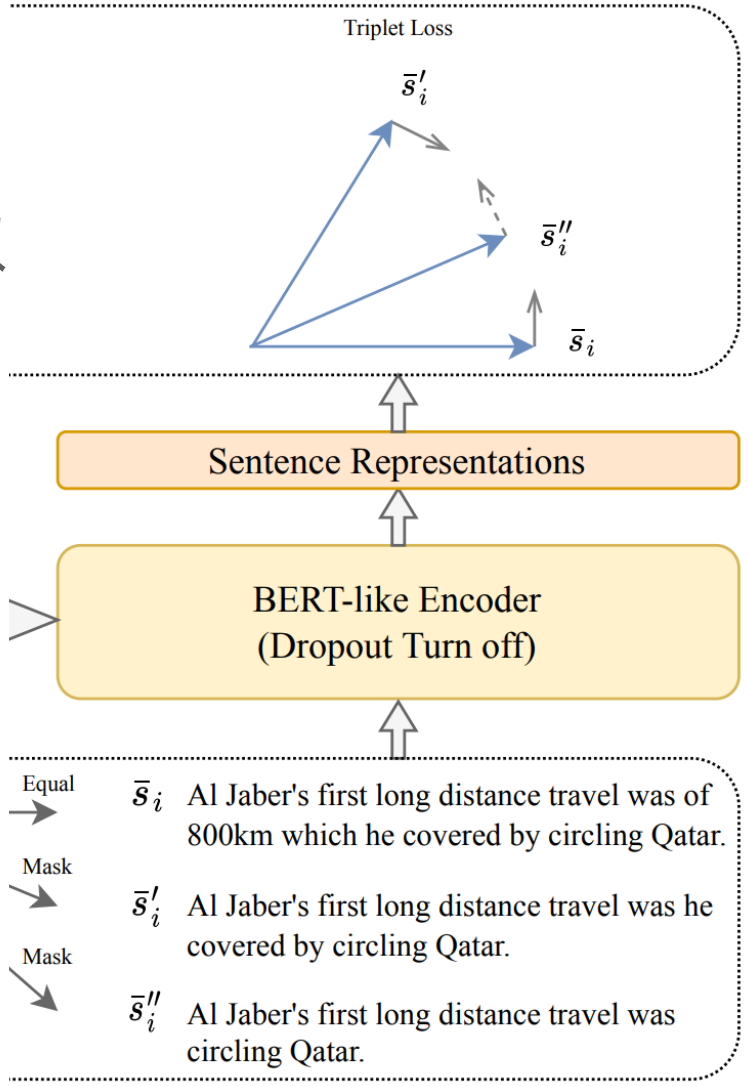
手法: 含意関係を捉える自己教師あり学習法

1. 角度ベースで決定境界にマージンを導入する損失関数を提案
2. 文の3つ組に対し，含意関係を捉える自己教師あり学習法を提案

- アンカー文$\bar{s}_i$に対し，20%, 40%のマスクした$\bar{s}_i'$と$\bar{s}_i''$を作成
- $\bar{s}_i$と$\bar{s}_i'$の類似度を$\bar{s}_i$と$\bar{s}_i''$の類似度に対して高くなるように学習

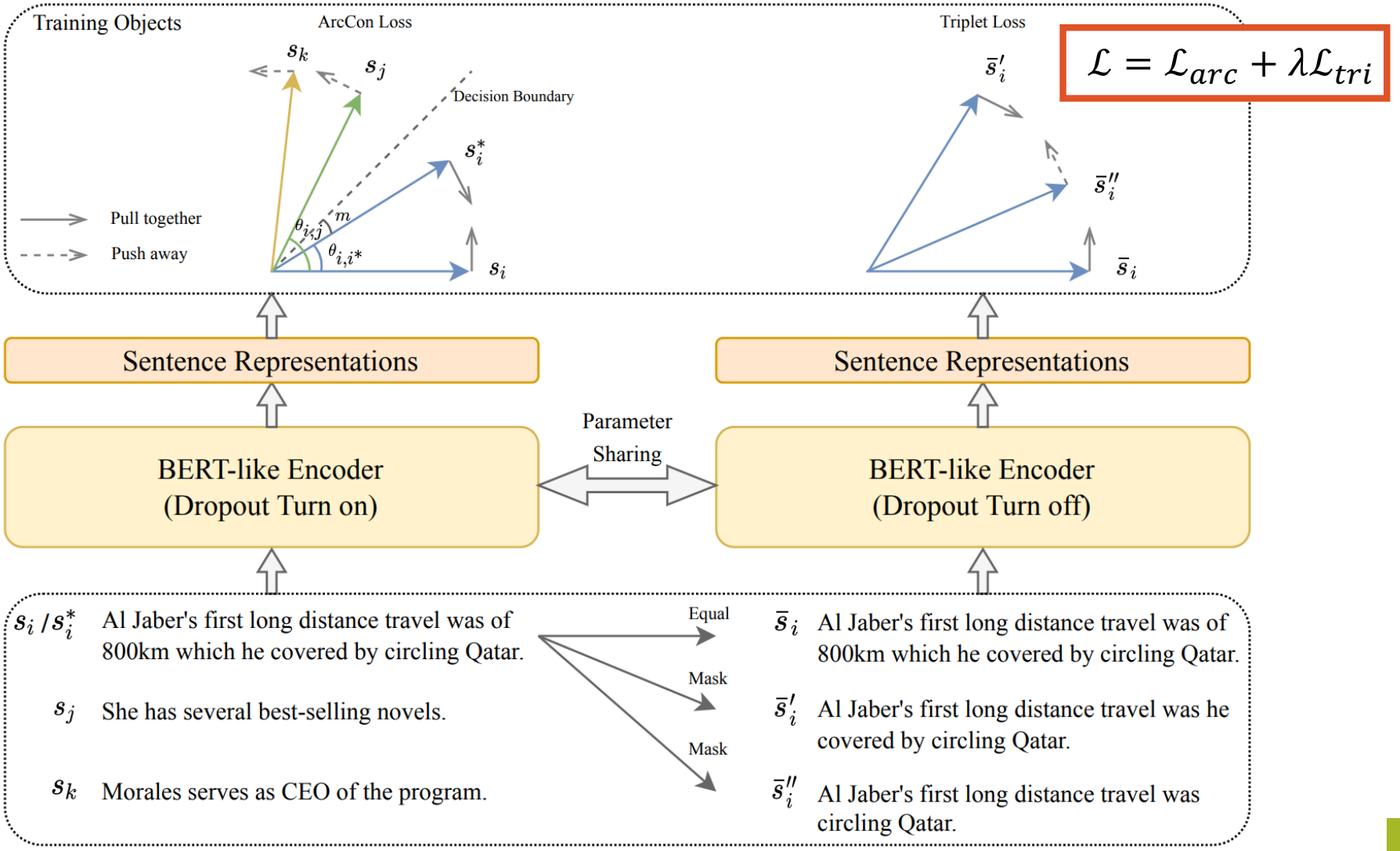
$$\mathcal{L}_{\text{tri}} = max\left(0, sim(\bar{h}_i, \bar{h}_i'') - sim(\bar{h}_i, \bar{h}_i') + m\right)$$

- 含意を満たさない事例も出現しうるが，無視できる程度
- ドロップアウトのノイズは，含意関係が不明瞭になる可能性があるため，適用しない



手法: ArcCSEの概要

1. 角度ベースで決定境界にマージンを導入する損失関数を提案
2. 文の3つ組に対し，含意関係を捉える自己教師あり学習法を提案



実験設定: 評価タスクと評価指標

- Semantic Textual Similarity (STS): 文ペア意味的類似度
  - 7タスク: STS12-16, STSb, SICK-R
  - 評価指標はスピアマン相関係数

- SentEval: ロジスティック回帰による分類
  - 7タスク (分類クラス数)
    - MR: 映画レビューの感情極性分類 (2)
    - CR: 商品レビューの感情極性分類 (2)
    - SUBj: 映画関連テキストの主観客観分類 (2)
    - MPQA: ニュース記事の意見極性分類 (2)
    - SST-2: 映画レビューの感情極性分類 (2)
    - TREC: 質問タイプの分類 (6)
    - MRPC: 2文の同義判定 (2)
  - 評価指標は正解率

実験結果: STSタスク

- ArcCSEはSimCSEより高いパフォーマンス
- ArcCon lossとTriplet lossもそれぞれ有効
  - w/o ArcCon lossは，NT-Xent lossとTriplet lossの組み合わせ

Method	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
GloVe (avg.)	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
$\text{BERT}_{\text{base}}$ (last avg.)	30.87	59.89	47.73	60.29	63.73	47.29	58.22	52.57
$\text{BERT}_{\text{base}}$-flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
$\text{BERT}_{\text{base}}$-whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-$\text{BERT}_{\text{base}}$	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-$\text{BERT}_{\text{base}}$	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
$\text{ConSERT}_{\text{base}}$	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
SimCSE-$\text{BERT}_{\text{base}}$	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
ArcCSE-$\text{BERT}_{\text{base}}$	**72.08**	**84.27**	**76.25**	82.32	79.54	**79.92**	**72.39**	**78.11**
w/o ArcCon loss	69.94	82.34	75.08	**83.08**	78.97	78.59	71.13	77.02
w/o Triplet loss	69.66	81.92	75.33	82.79	**79.55**	79.56	71.94	77.25
$\text{ConSERT}_{\text{large}}$	70.69	82.96	74.13	82.78	76.66	77.53	70.37	76.45
SimCSE-$\text{BERT}_{\text{large}}$	70.88	84.16	76.43	84.50	**79.76**	79.26	**73.88**	78.41
ArcCSE-$\text{BERT}_{\text{large}}$	**73.17**	**86.19**	**77.90**	**84.97**	79.43	**80.45**	73.50	**79.37**

実験結果: SentEval

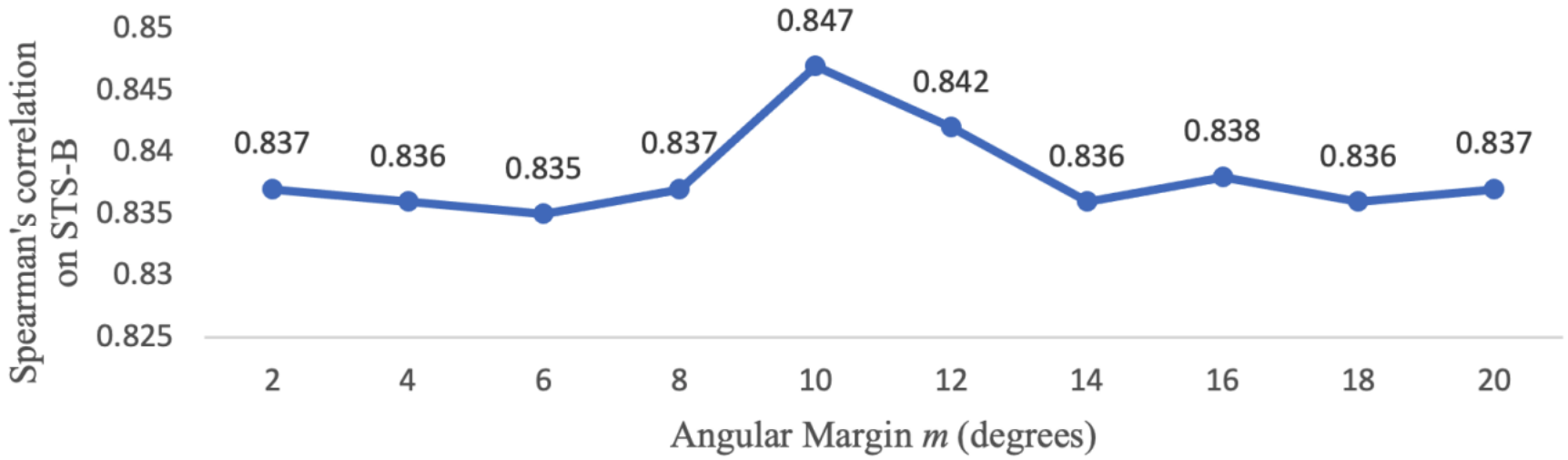
- ArcCSEは，$\text{BERT}_{\text{base}}$と$\text{BERT}_{\text{large}}$の両モデルでベースラインと同等以上の性能
  - ドメインに特化した文埋め込みとしての有効性を確認

Method	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
GloVe (avg.)	77.25	78.30	91.17	87.85	80.18	83.00	72.87	81.52
Skip-thought	76.50	80.10	93.60	87.10	82.00	92.20	73.00	83.50
$\text{BERT}_{\text{base}}$ (last avg.)	78.66	86.25	94.37	88.66	84.40	**92.80**	69.54	84.94
IS-$\text{BERT}_{\text{base}}$	81.09	**87.18**	94.96	88.75	**85.96**	88.64	74.24	85.83
SimCSE-$\text{BERT}_{\text{base}}$	**81.18**	86.46	94.45	88.88	85.50	89.80	74.43	85.81
ArcCSE-$\text{BERT}_{\text{base}}$	79.91	85.25	**99.58**	**89.21**	84.90	89.20	**74.78**	**86.12**
$\text{BERT}_{\text{large}}$ (last avg.)	84.30	89.22	95.60	86.93	89.29	91.40	71.65	86.91
SimCSE-$\text{BERT}_{\text{large}}$	**85.36**	**89.38**	95.39	89.63	90.44	91.80	**76.41**	88.34
ArcCSE-$\text{BERT}_{\text{large}}$	84.34	88.82	**99.58**	**89.79**	**90.50**	**92.00**	74.78	**88.54**

実験結果: 角度マージンの影響

- ArcCon lossで必要な角度マージン (m) のハイパーパラメータチューニングの結果
  - 2から20まで2ずつ変化
  - STS-Bの開発セットのスコアで探索

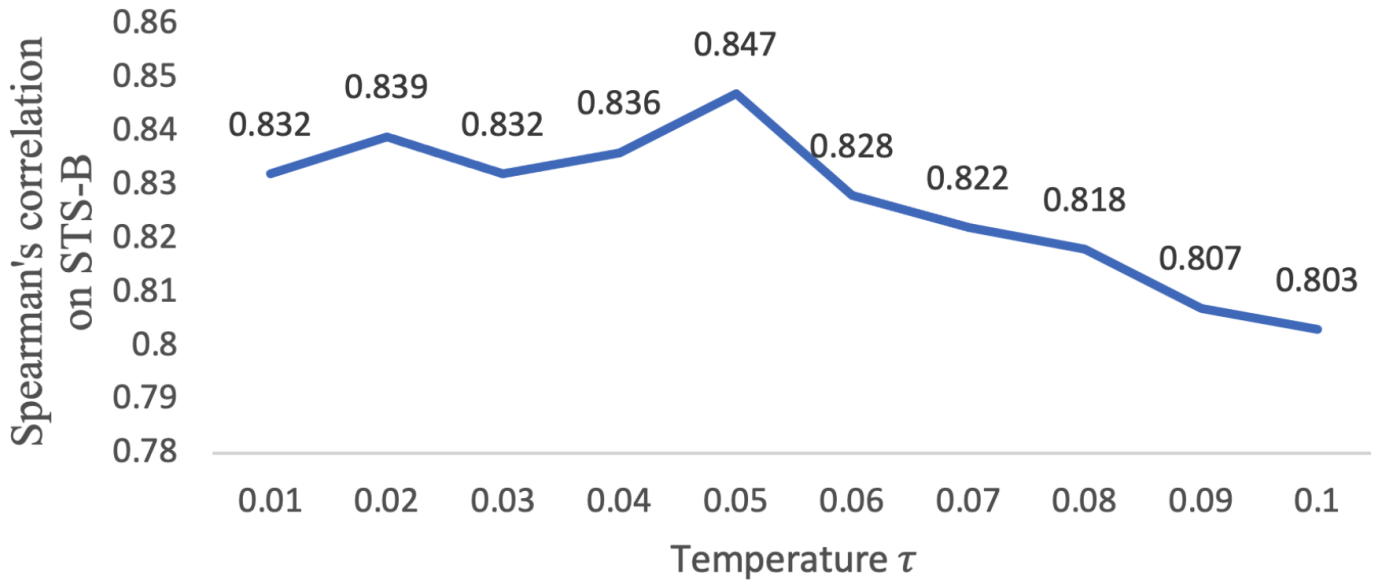
- mが10の時に最高スコア
  - マージン次第でスコアが1%前後変化



実験結果: 温度パラメータの影響

- ArcCon lossで必要な温度パラメータのハイパーパラメータチューニングの結果
  - 0.01から0.1まで0.01ずつ変化
  - STS-Bの開発セットのスコアで探索

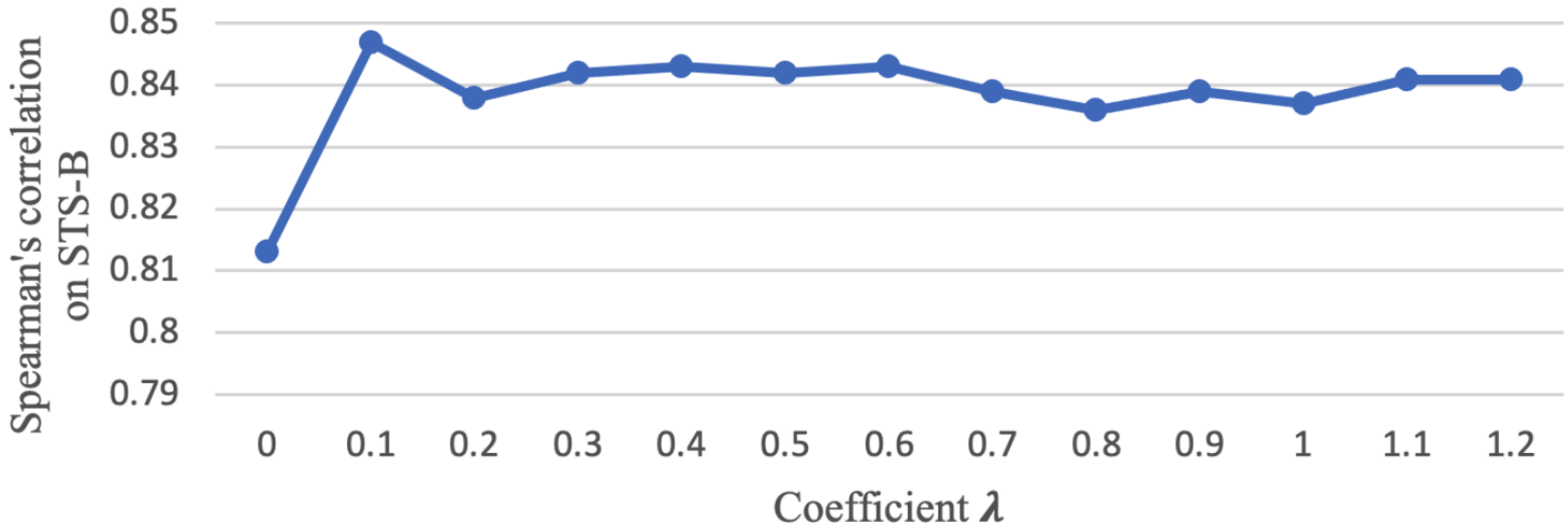
- τが0.05の時に最高スコア
  - 温度パラメータも探索する必要あり



実験結果: ArcCon lossとTriplet loss

- ArcCon lossとTriplet lossの間の相対的な重みλの調整結果
  - λを0から1.2まで0.1ずつ変化

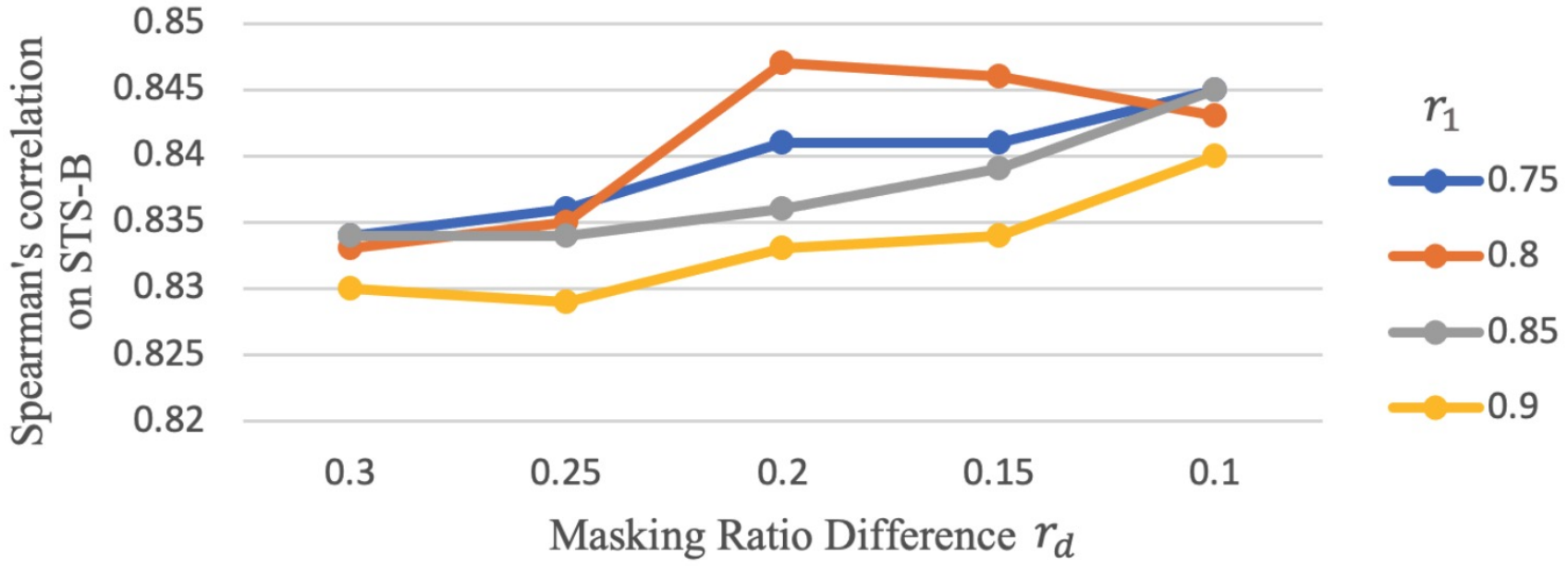
- λ=0.1のときに最大スコア
  - 適切なλを設定する必要あり



実験結果: マスク率の影響

- 文の3つ組 (s_1, s_2, s_3) に基づく学習時のマスク率の調査
  - 元の文s_1に対して，マスクしない率r_1を適用したものを文 s_2
  - 元の文s_1に対して，マスクしない率r_1+r_dを適用したものを文 s_3

- r_1=0.8と r_d=0.2のときに最大スコア
  - マスク率の差 r_d が大きい時にスコアが低下
  - タスク難易度のバランスが重要



実験結果: ドロップアウトの有無による影響

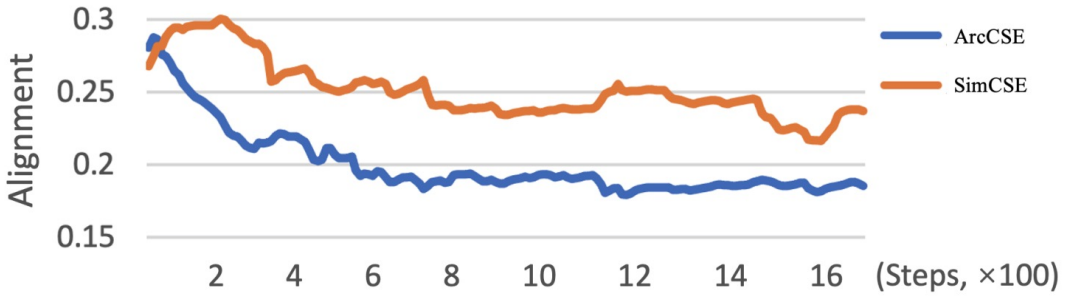
- ArcCon lossとTriplet lossにおけるエンコーダのドロップアウトの影響を調査
  - “on”と“off”はドロップアウトありとなしの設定
  - “mix”は正例ペアをそれぞれドロップアウトありとなしで構成した設定

- ArcCon lossでドロップアウトをオン，Triplet lossでドロップアウトをオフの設定が最良

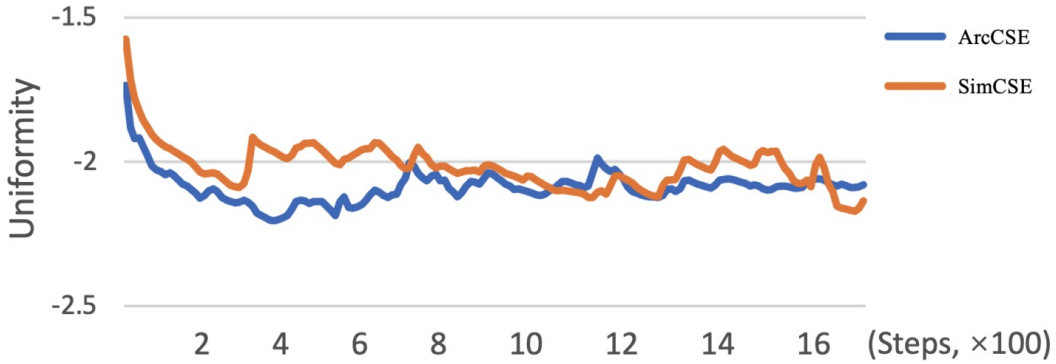
ArcCSE-BERT$_{base}$	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
w/ Dropout$_{on/off}$	**72.08**	**84.27**	**76.25**	**82.32**	**79.54**	**79.92**	**72.39**	**78.11**
w/ Dropout$_{mix/off}$	70.51	83.59	75.85	82.30	78.87	78.74	71.58	77.35
w/ Dropout$_{on/on}$	69.62	83.13	74.42	82.15	78.39	78.39	70.89	76.71

実験結果: アライメントと一様性

- 対照学習と密接に関連した表現の質を測定する指標 [7] であるアライメント (ℓ_{align}) と一様性 ($\ell_{uniform}$) を算出
  - アライメント: ArcCSEはSimCSEより良いパフォーマンス
  - 一様性: ArcCSEはSimCSEと同程度のフォーマンス



(a) ℓ_{align}



(b) ℓ_{uniform}

[7] Wang and Isola: Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere (ICML'20)

まとめ

- 目的
  - 高品質な文埋め込みの構成

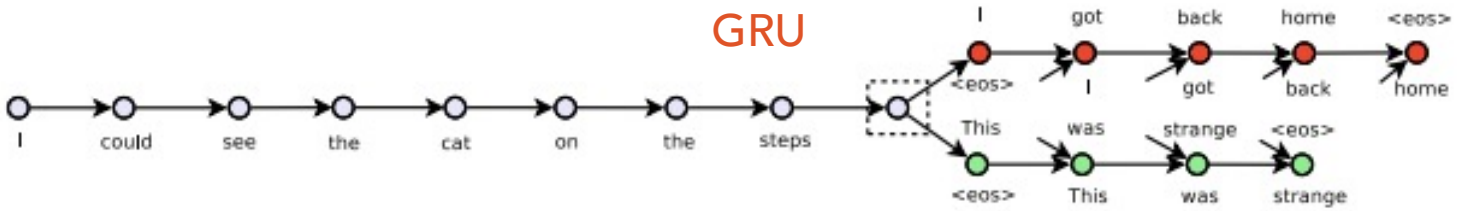
- 手法: ArcCSE
  1. 角度ベースで決定境界にマージンを導入する損失関数
  2. 文の3つ組に対し，含意関係を捉える自己教師あり学習法

- 実験結果
  - 従来の最先端モデルSimCSEを超えるパフォーマンス
  - それぞれの損失関数の有用性の確認

[付録] BERT以前の文埋め込み

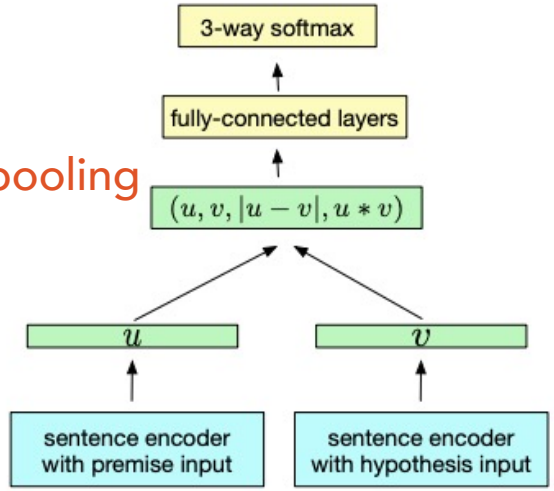
- Skip-Thought [12]
  - エンコードした文の前後の文をデコードするように学習

GRU



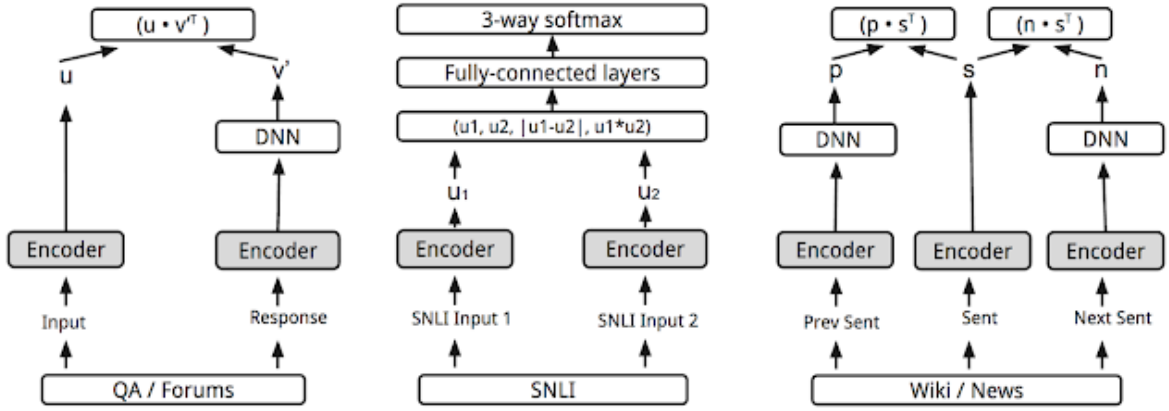
- InferSent [13]
  - NLIを学習
  - Siamese Network

Bi-LSTM
w/ max pooling



- Universal Sentence Encoder (USE) [14]
  - マルチタスク学習 (QA, NLI, modified Skip-Thought)

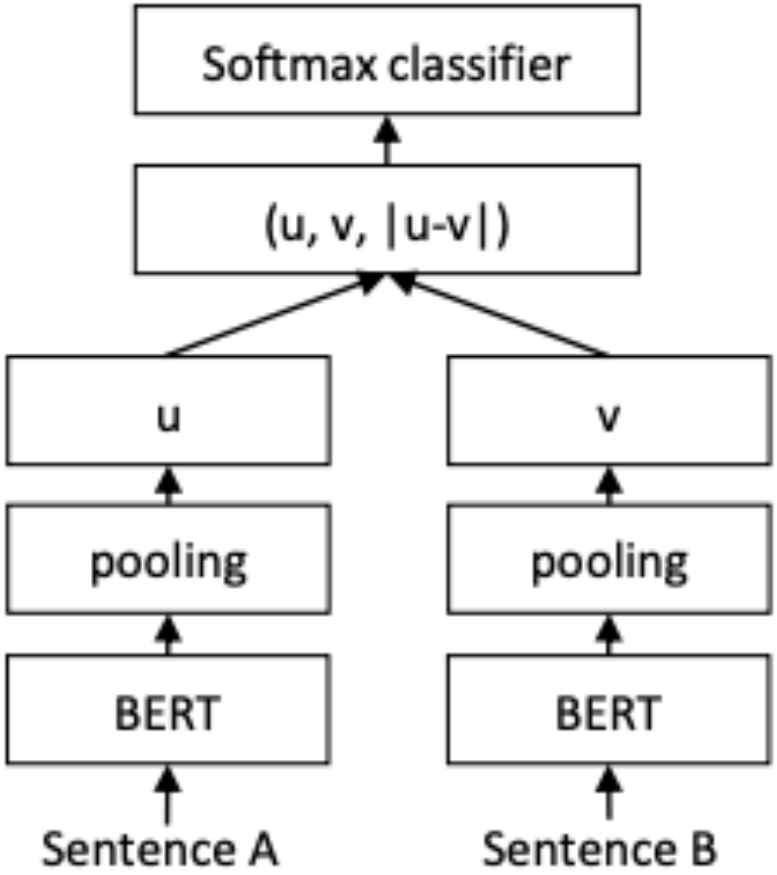
Transformer (Encoder)



[12] Kiros et al.: Skip-Thought Vectors (NIPS'15)
[13] Conneau et al.: Supervised Learning of Universal Sentence Representations from Natural Language Inference Data (EMNLP'17)
[14] Cer et al.: Universal Sentence Encoder for English (EMNLP-demo'18)

[付録] Sentence-BERT [15]

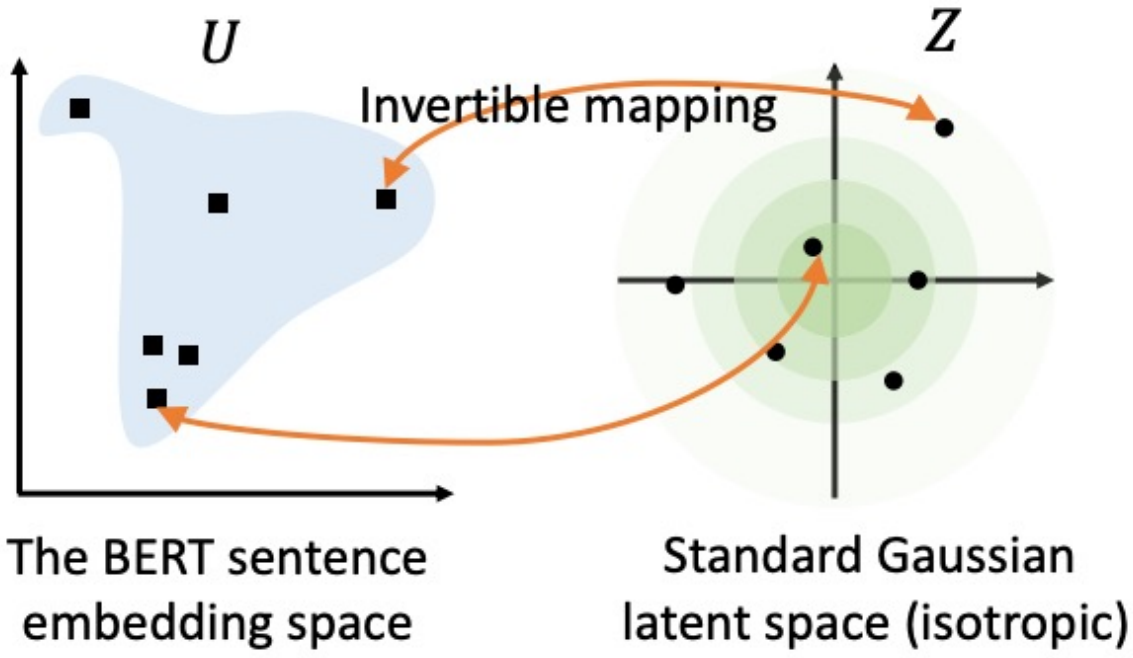
- NLIデータセット (SNLI, MultiNLI) を利用して，Siamese NetworkによるBERTのファインチューニング
  - 従来の文埋め込み手法infersentやUSEにインスパイア
  - 文ペアの含意/中立/矛盾の3値分類を学習
  - 平均プーリングが最良



[15] Reimers and Gurevych: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (EMNLP'19)

[付録] BERT-flow [16]

- BERTの埋め込み空間の異方性に関する分析
- 等方性を満たす標準ガウス潜在空間に逆変換可能な写像を学習
  - flowベースの生成モデルGlowを利用
  - ターゲットドメインのテキストで学習



[16] Li et al.: On the Sentence Embeddings from Pre-trained Language Models (EMNLP'20)

[付録] BERT-whitening [17]

- BERT-flowのような高度な手法でなくとも，白色化というシンプルかつ効率的な後処理技術で十分有効であることを実証
  - 白色化は，各特徴量を無相間化し，平均値を0，標準偏差を1へ変換
  - ターゲット/NLIデータセットから得られた白色化パラメータを利用

Algorithm 1 Whitening-k Workflow

Input: Existing embeddings $\{x_i\}_{i=1}^{N}$ and reserved dimensionality k

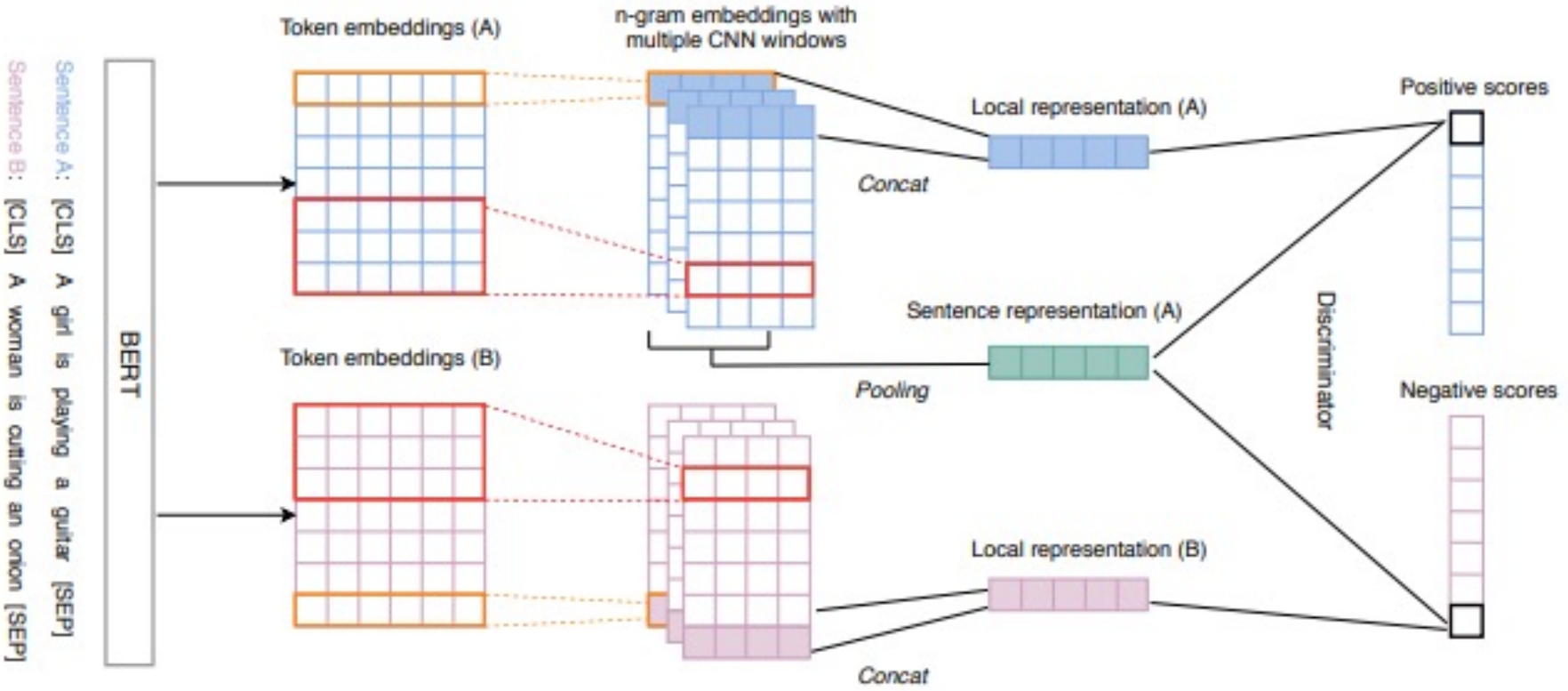
compute μ and Σ of $\{x_i\}_{i=1}^{N}$
compute $U, \Lambda, U^T = \text{SVD}(\Sigma)$
compute $W = (U\sqrt{\Lambda^{-1}})[:, :k]$
for $i = 1, 2, \cdots, N$ **do**
    $\tilde{x}_i = (x_i - \mu)W$
end for

Output: Transformed embeddings $\{\tilde{x}_i\}_{i=1}^{N}$

[17] Li et al.: On the Sentence Embeddings from Pre-trained Language Models (EMNLP'20)

[付録] IS-BERT (Info-Sentence BERT) [18]

- CNN層を通して文埋め込みを構成
  - 大域的な埋め込み (文埋め込み) と局所的な埋め込み (n-gram埋め込み) の相互情報量が最大となるように学習
  - ラベル情報なしのNLIで学習



[18] Zhang et al.: An Unsupervised Sentence Embedding Method by Mutual Information Maximization (EMNLP'20)