

# 多言語文埋め込みの意味と言語の分離のための損失関数の分析

野々村 奏<sup>1</sup> 福島 啓太<sup>1</sup> 近藤 里咲<sup>1</sup> 梶原 智之<sup>1,2</sup>

<sup>1</sup> 愛媛大学 <sup>2</sup> 大阪大学

{nonomura@ai., fukushima@ai., kondo@ai., kajiwara@}cs.ehime-u.ac.jp

## 概要

本研究では、多言語文埋め込みが言語情報の影響を強く受ける問題に対し、意味情報と言語情報の2要素に分離するアプローチに着目する。分離のための訓練の損失関数は、意味情報または言語情報の各要素内において定義される要素内制約と、両者に跨る要素間制約に大別できる。既存手法ではこれらを併用しているが、制約の違いによる埋め込み分離への影響や、埋め込みモデルのアーキテクチャの違いの影響については明らかになっていない。実験の結果、エンコーダ由来モデルでは両制約の併用が有効であるのに対し、LLM由来モデルでは要素間制約のみが有効であることが示された。また、要素内制約は埋め込み空間の一様性を改善し、要素間制約は対訳文間の整合性を改善するという違いを確認した。

## 1 はじめに

多言語文埋め込み [1-4] は機械翻訳の品質推定 [5] や言語横断的な文類似度推定 [6] および情報検索 [7] など、多様な自然言語処理タスクで活用されている。しかし、埋め込みには言語に依存する情報が強く含まれる言語特異性の問題 [8] があり、対訳コーパスを用いた訓練により、言語情報とは独立な意味情報の獲得に向けた研究が進められてきた [8-10]。これらは複数の損失関数を用いたマルチタスク学習により、意味情報（以下、意味表現）と言語情報（以下、言語表現）を独立した要素として分離することで意味情報の抽出を図っている。

これらの一連の分離手法の中で、SEED [10] が最も高い性能を示している。図1のように、SEEDの訓練における損失関数は2種類に分類できる。1つ目は要素内制約である。これは意味表現または言語表現のいずれかの要素内で定義される損失関数である。具体的には、対訳文間では意味表現同士類似度を最大化する一方で、同一言語文間では言語表現同士の類似度を最大化する。この制約により対訳文

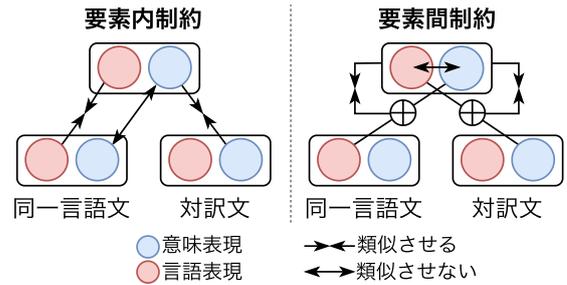


図1 要素内制約と要素間制約の概要図。

の意味表現同士および同一言語文の言語表現同士が類似するよう促される。2つ目は要素間制約である。これは意味表現と言語表現の両者に跨って定義される損失関数である。具体的には、文埋め込みが意味表現と言語表現の和で構成される前提のもと、意味表現を対訳文間で置換したり、言語表現を同一言語文間で置換したりする。この制約により置換関係にある対訳文の意味表現同士および同一言語文の言語表現同士が類似するよう促される。そのため、両制約は同様の役割を果たすと考えられるが、表現分離への影響の違いや、従来のエンコーダ由来および近年のLLM由来のモデルといったアーキテクチャの違いとの相性については不明である。

そこで本研究では、機械翻訳の品質推定タスク (Quality Estimation; QE [5]) と言語横断の文間類似度推定タスク (Semantic Textual Similarity; STS [6]) を対象に、両者の違いの傾向を探る。また、得られた意味表現の幾何学的特性 [11] の観点から、各制約が埋め込み空間に与える影響を分析する。

評価実験の結果、エンコーダ由来のモデル [1,2] には両制約の併用が有効であるのに対し、LLM由来のモデル [3,4] には要素間制約のみが有効であることが示された。また、幾何学的特性の観点では、要素内制約は埋め込み空間の一様性 (Uniformity) を改善する傾向があり、要素間制約は対訳文が空間上の近くに配置される性質 (Alignment) を改善する傾向があることを確認した。

## 2 損失関数の構成要素の分類

埋め込みの分離のための訓練における損失関数 [8–10] は、意味表現または言語表現の要素内で定義される要素内制約と、2要素に跨る要素間制約に分類できる。いずれにせよ、対訳文間の意味表現が類似し、同一言語間の言語表現が類似する設計であるため、分離を実現する上ではどちらか片方の制約だけでも成立すると考えられる。本節では、まずモデルの構成を述べ、その後、本研究の分析対象である要素内制約と要素間制約について説明する。

### 2.1 モデルの構成

言語  $X$  の文を  $x$ 、言語  $Y$  の文を  $y$  とし、それぞれの文埋め込みを  $e_x, e_y$  とする。文埋め込みには、言語非依存の意味情報と言語依存の言語情報の両方が含まれる。これらを独立した要素として分離するために、文埋め込みが意味表現  $e^{(m)}$  と言語表現  $e^{(l)}$  の和で構成されると仮定する。既存手法の SEED [10] に従い、意味表現抽出器 (Multi-Layer Perceptron; MLP) を用いて意味表現を抽出 ( $e^{(m)} = \text{MLP}(e)$ ) し、元の文埋め込みと意味表現の差分 ( $e^{(l)} = e - e^{(m)}$ ) を言語表現として定義する。その後、意味表現が言語非依存な文の意味情報を、言語表現が文の言語情報を獲得できるように損失関数を定義する。要素内制約と要素間制約の概要を図 1 に示す。

### 2.2 要素内制約

要素内制約  $L_{\text{intra}}$  は意味表現または言語表現のいずれかの要素内で定義され、 $L_{\text{intra}} = L_{\text{mean}} + L_{\text{lang}}$  で表される。以降、バッチ内の  $i$  番目の対訳文対を  $(x_i, y_i)$  とする。また、 $j$  は、 $i$  と同バッチ内の異なるインデックスを表すものとする。

**意味表現損失** 意味表現において対訳文間  $(x_i, y_i)$  は同じ意味情報を持つため類似度を最大化し、同一言語内の異なる文間  $(x_i, x_j)$  は異なる意味情報を持つため最小化する。よって以下の式 (1) を定義する。

$$L_{\text{mean}} = 2 \left( 1 - \cos \left( e_{x_i}^{(m)}, e_{y_i}^{(m)} \right) \right) + \max \left( 0, \cos \left( e_{x_i}^{(m)}, e_{x_j}^{(m)} \right) + \cos \left( e_{y_i}^{(m)}, e_{y_j}^{(m)} \right) \right). \quad (1)$$

**言語表現損失** 言語表現において同一言語間  $(x_i, x_j), (y_i, y_j)$  は同じ言語情報を持つため類似度を最大化する。よって以下の式 (2) を定義する。

$$L_{\text{lang}} = 2 - \cos \left( e_{x_i}^{(l)}, e_{x_j}^{(l)} \right) - \cos \left( e_{y_i}^{(l)}, e_{y_j}^{(l)} \right). \quad (2)$$

### 2.3 要素間制約

要素間制約  $L_{\text{inter}}$  は、意味表現と言語表現の両方に跨って定義され、 $L_{\text{inter}} = L_{\text{cross}} + L_{\text{sep}}$  で表される。

**交差復元損失** 対訳文間  $(x_i, y_i)$  は同一の意味情報を持つため、 $x_i$  の言語表現と  $y_i$  の意味表現の和は  $e_{x_i}$  を再構成でき、 $y_i$  の言語表現と  $x_i$  の意味表現の和は  $e_{y_i}$  を再構成できる。一方で、同一言語文間  $(x_i, x_j)$  および  $(y_i, y_j)$  は同じ言語情報を持つため、 $x_i$  の意味表現と  $x_j$  の言語表現の和は  $e_{x_i}$  を再構成でき、 $y_i$  の意味表現と  $y_j$  の言語表現の和は  $e_{y_i}$  を再構成できる。よって以下の式 (3) を定義する。

$$L_{\text{cross}} = 4 - \cos \left( e_{x_i}, e_{y_i}^{(m)} + e_{x_i}^{(l)} \right) - \cos \left( e_{y_i}, e_{x_i}^{(m)} + e_{y_i}^{(l)} \right) - \cos \left( e_{x_i}, e_{x_i}^{(m)} + e_{x_j}^{(l)} \right) - \cos \left( e_{y_i}, e_{y_i}^{(m)} + e_{y_j}^{(l)} \right). \quad (3)$$

**分離損失** 抽出された意味表現と言語表現が互いに重複した情報を持たないように、類似度を最小化する。よって以下の式 (4) を定義する。

$$L_{\text{sep}} = \max \left( 0, \cos \left( e_{x_i}^{(m)}, e_{x_i}^{(l)} \right) \right). \quad (4)$$

## 3 タスクの性能に関する分析

QE タスクおよび言語横断の STS タスクを対象に、分離方法の違いによる性能を評価する。QE は参照訳を用いずに翻訳品質を推定するタスクである。対訳コーパスのみを訓練に使用する本手法においては教師なし設定<sup>1)</sup>として取り組む。STS は与えられた 2 文間の意味的類似性を推定するタスクである。どちらのタスクも、所与の 2 文間の意味表現をもとに余弦類似度を算出し、人手評価値とのピアソン相関係数によりモデルの性能を評価する。

### 3.1 実験設定

**データセット** WMT20 QE タスクには 6 つの言語対<sup>2)</sup>が含まれる。各言語対について、原言語文と目的言語文の文対 1,000 件に対し、人手による品質評価スコアが付与されたデータセットが提供されている。なお、品質推定の対象となる機械翻訳モデルは fairseq ツールキット [12] で訓練された Transformer モデル [13] である。モデルの訓練には

1) 教師あり設定では「原言語文・目的言語文・人手評価値」で訓練するが、教師なし設定では人手評価値を使用しない。

2) 内訳は、多資源言語対の英語-ドイツ語 (en-de) および英語-中国語 (en-zh)、中資源言語対のルーマニア語-英語 (ro-en) およびエストニア語-英語 (et-en)、少資源言語対のネパール語-英語 (ne-en) およびシンハラ語-英語 (si-en) である。 <https://github.com/facebookresearch/mlqe>

表1 訓練用データの内訳（それぞれ10%は検証用）

WMT20 QE		SemEval-2017 STS	
言語対	文対数	言語対	文対数
en-de, en-zh	100万	en-it, en-tr	50万
ro-en, et-en	20万	en-de, en-es, en-fr	20万
ne-en, si-en	5万	en-ar, en-nl	3万

表2 機械翻訳の品質推定（ピアソン相関係数）

モデル名	手法	多資源	中資源	少資源	Avg.
		en-de	ro-en	ne-en	
LaBSE	訓練なし	0.084	0.705	0.547	0.396
	要素内制約	0.187	<b>0.726</b>	<b>0.634</b>	0.481
	要素間制約	0.143	0.698	0.517	0.401
	両方	<b>0.191</b>	0.725	0.633	<b>0.482</b>
mE5-large-instruct	訓練なし	0.143	<b>0.767</b>	0.549	0.446
	要素内制約	0.213	0.758	0.592	0.493
	要素間制約	0.196	0.766	<b>0.602</b>	0.485
	両方	<b>0.215</b>	0.761	0.596	<b>0.497</b>
Gemini Embedding	訓練なし	0.159	0.789	0.699	0.515
	要素内制約	<b>0.219</b>	0.788	0.681	0.529
	要素間制約	0.204	<b>0.798</b>	<b>0.701</b>	<b>0.536</b>
	両方	0.218	0.792	0.686	0.534
Qwen3-Embedding-8B	訓練なし	0.186	0.755	0.626	0.456
	要素内制約	<b>0.222</b>	0.754	0.616	0.473
	要素間制約	0.218	<b>0.763</b>	<b>0.632</b>	<b>0.481</b>
	両方	<b>0.222</b>	0.757	0.625	0.478

タスクで利用可能な対訳コーパスの一部<sup>3)</sup>を使用した。SemEval-2017 STS タスクには7つの言語対<sup>4)</sup>が含まれる。各言語対において250文対と人手評価値の組が評価用に提供されている。モデルの訓練には、先行研究 [8, 10] と同様に対訳コーパスの Tatoeba<sup>5)</sup> から訓練用データを抽出した。内訳を表1に示す。

**モデル** 本実験における訓練の対象は MLP のみであり、1層の線形層のみで構成される。文埋め込みの獲得には、エンコーダ由来の埋め込みモデルとして LaBSE<sup>6)</sup> [1] および mE5-large-instruct<sup>7)</sup> [2] を、

3) <http://www.statmt.org/wmt20/quality-estimation-task.html> から先行研究 [8-10] と同程度の量を無作為抽出した。

4) 内訳は、多資源言語対の英語-イタリア語 (en-it) および英語-トルコ語 (en-tr)、中資源言語対の英語-ドイツ語 (en-de)、英語-スペイン語 (en-es)、英語-フランス語 (en-fr)、少資源言語対の英語-アラビア語 (en-ar) および英語-オランダ語 (en-nl) である。 <https://public.ukp.informatik.tu-darmstadt.de/reimers/sentence-transformers/datasets/>

5) <https://tatoeba.org/ja/>

6) <https://huggingface.co/sentence-transformers/LaBSE>

7) <https://huggingface.co/intfloat/multilingual-e5-large-instruct>

8) <https://ai.google.dev/gemini-api/docs/embeddings>

表3 言語横断の文類似度推定（ピアソン相関係数）

モデル名	手法	多資源	中資源	少資源	Avg.
		en-it	en-fr	en-nl	
LaBSE	訓練なし	0.760	0.759	0.755	0.734
	要素内制約	0.778	0.782	<b>0.776</b>	0.751
	要素間制約	0.750	0.749	0.753	0.727
	両方	<b>0.780</b>	<b>0.783</b>	<b>0.776</b>	<b>0.753</b>
Gemini Embedding	訓練なし	0.908	0.901	0.899	0.886
	要素内制約	0.903	0.895	0.896	0.886
	要素間制約	<b>0.913</b>	<b>0.906</b>	<b>0.905</b>	<b>0.893</b>
	両方	0.908	0.901	0.900	0.890

LLM 由来のモデルとして Gemini Embedding<sup>8)</sup> [3] および Qwen3-Embedding-8B<sup>9)</sup> [4] を使用した。文埋め込みの獲得時には、エンコーダ由来のモデルでは単に入力文のみをモデルに与え、LaBSE では最終層の CLS トークンを、mE5-large-instruct では最終層の各トークンの平均プーリングを用いた。Gemini Embedding では API を使用し、埋め込み取得時の Task type には STS を指定した。Qwen3-Embedding-8B は評価リポジトリ<sup>10)</sup>を参考に、STS のタスク時の指示 “Retrieve semantically similar text” を入力文に付与し、最終層の EOS トークンを文埋め込みとした。LLM 由来の2つのモデルは複数の埋め込み次元数に対応した Matryoshka Representation Learning [14] のモデルであるが、本実験ではそれぞれ最大次元数の 3,072 次元および 4,096 次元を使用した。

**比較手法** 本研究の分析対象である要素内制約と要素間制約、そして両制約の併用 (SEED [10] と同じ構成) により訓練した場合を比較する。

**ハイパーパラメータ** 訓練には HuggingFace Transformers [15] を使用し、バッチサイズを 512、最適化手法を Adam [16]、学習率を  $1e-4$  とした。検証用データとして訓練データの 10% を無作為抽出した。検証用データにおける損失が3エポック連続で改善しない場合に訓練を終了した。

### 3.2 実験結果

**WMT20 QE タスク** 表2には各資源区分から1言語対の結果と全6言語対の平均性能 (Avg.) を記載した。いずれの単一の制約においても、訓練なしと比較して平均性能が向上しており、表現分離が適切に実行された。エンコーダ由来のモデル (LaBSE と mE5-large-instruct) では全6言語対の中で、どちら

8) <https://ai.google.dev/gemini-api/docs/embeddings>

9) <https://huggingface.co/Qwen/Qwen3-Embedding-8B>

10) <https://github.com/QwenLM/Qwen3-Embedding>

も3言語対において両方の制約を課した場合が最も高く、次いで要素内制約が高かった。平均性能も同様の順位関係である。一方で、LLM由来のモデル(Gemini EmbeddingとQwen3-Embedding-8B)では、平均性能を含めどちらも3言語対で要素間制約に基づくモデルが最も高かった。

**SemEval-2017 STS タスク** 表3には、QEタスクと同様に各資源区分から1言語対の結果と全7言語対の平均性能(Avg.)を記載した。エンコーダ由来のLaBSEでは全7言語対のうちのすべての言語対において両方の制約を課した場合が最も高く、LLM由来のGemini Embeddingについては6つの言語対において要素間制約が高い性能となった。平均性能も同様の傾向だった。以上より、WMT20 QEとSemEval-2017 STSの実験結果には共通して、エンコーダ由来のモデルには両制約を用いる手法が有効で、LLM由来のモデルには要素間制約のみを用いる手法が有効であることが明らかとなった。

## 4 幾何学的特性に関する分析

各制約が埋め込み空間に与える影響を明らかにするために、幾何学的特性を評価するAlignmentおよびUniformity [11]の指標を用いて分析する。これらの指標は、対照学習をはじめとする表現学習手法を、埋め込み空間の観点から評価するために用いられる[17–19]。ここでは、入力文 $x$ の意味表現に対してL2正規化した埋め込み $\tilde{h}(x)$ を用いる。

**Alignment** Alignmentは対訳文間の意味表現が、正規化された埋め込み空間上でどの程度近接しているかを表す指標である。本研究では、言語 $X$ の文を $x$ 、言語 $Y$ の文を $y$ とし、対訳ペア $(x, y)$ の分布を $p_{\text{pos}}$ としたとき以下の式(5)で定義される。

$$L_{\text{align}} = \mathbb{E}_{(x, y) \sim p_{\text{pos}}} \|\tilde{h}(x) - \tilde{h}(y)\|^2. \quad (5)$$

この値が小さいほど、対訳文対が埋め込み空間上で互いに近接し、言語間の整合性が高いといえる。

**Uniformity** Uniformityは正規化された埋め込みが、空間全体にどの程度均等に分散しているかを測る指標である。本研究では、全言語対のデータの周辺分布を $p_{\text{data}}$ としたとき以下の式(6)で定義する。

$$L_{\text{uniform}} = \log \mathbb{E}_{x, z \stackrel{\text{i.i.d.}}{\sim} p_{\text{data}}} e^{-2\|\tilde{h}(x) - \tilde{h}(z)\|^2}. \quad (6)$$

この値が小さいほど、埋め込み同士が超球面上に広く分散していることを示す。これは、埋め込み空間

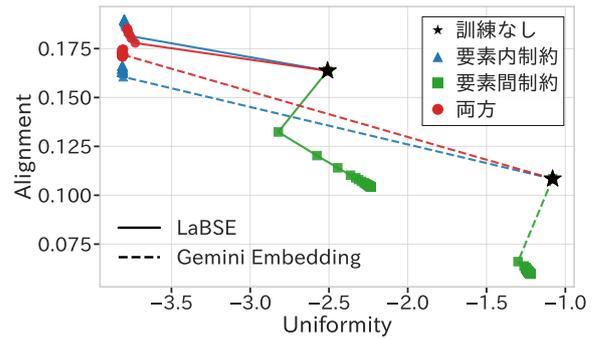


図2 SemEval-2017の訓練におけるエポックごとにAlignmentとUniformityの数値をプロットした図。

の限られた領域への集中を抑制し、空間全体をより有効に活用できていることを意味する。

**分析結果** 各言語対について検証データから無作為に抽出した1,000件を使用し、訓練時のエポック毎にAlignmentとUniformityを測定した。図2に、LaBSEとGemini Embeddingにおける結果を示す。いずれも、要素内制約あるいは両方の場合にUniformityの値が大幅に低下し、埋め込み空間を広く有効に活用できているが、Alignmentは悪化した。これは、要素内制約に含まれる意味表現損失(1)が、異なる文対に対する強い斥力として機能し、埋め込み空間の一様化を促したと考えられる。一方で、要素間制約の場合はUniformityの値はあまり変化しないものの、Alignmentの値が大きく低下し、言語間の整合性を改善した。これは、文埋め込みを再構成するという枠組みがUniformityを維持しつつ、表現分離が適切に機能しAlignmentが改善したと考えられる。以上より、エンコーダ由来のモデルではUniformityの改善が、LLM由来のモデルではAlignmentの改善がそれぞれ性能向上に寄与したと示唆される。ただし、モデルのアーキテクチャとこれらの幾何学的特性との間にある関係性の解明は今後の課題である。

## 5 おわりに

本研究では、多言語文埋め込みを意味表現と言語表現に分離する2つの訓練手法に着目し、両者の違いを分析した。その結果、エンコーダ由来モデルでは要素内制約と要素間制約の併用が有効である一方、LLM由来モデルでは要素間制約のみの場合に高い性能を示すことを明らかにした。さらに、幾何学的特性の分析から、要素内制約は埋め込み空間の一様性を改善し、要素間制約は対訳文間の整合性を改善するという違いを確認した。

## 謝辞

本研究はJST BOOST(課題番号:JPMJBY24036821)の支援を受けたものです。

## 参考文献

- [1] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavzhagan, and Wei Wang. Language-agnostic BERT Sentence Embedding. In **Proc. of ACL**, pp. 878–891, 2022.
- [2] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 Text Embeddings: A Technical Report. **arXiv:2402.05672**, 2024.
- [3] Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, Feng Han, Andreas Dourmoglou, Nithi Gupta, Fedor Moiseev, Cathy Yip, Aashi Jain, Simon Baumgartner, Shahrokh Shahi, Frank Palma Gomez, Sandeep Mariserla, Min Choi, Parashar Shah, Sonam Goenka, Ke Chen, Ye Xia, Koert Chen, Sai Meher Karthik Duddu, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghiya, Zach Gleicher, Karan Gill, Zhe Dong, Mojtaba Seyedhosseini, Yunhsuan Sung, Raphael Hoffmann, and Tom Duerig. Gemini Embedding: Generalizable Embeddings from Gemini. **arXiv:2503.07891**, 2025.
- [4] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. **arXiv:2506.05176**, 2025.
- [5] Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. Findings of the WMT 2020 Shared Task on Quality Estimation. In **Proc. of WMT**, pp. 743–764, 2020.
- [6] 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 **Proc. of SemEval**, pp. 1–14, 2017.
- [7] Jian-Yun Nie. **Cross-Language Information Retrieval**. Morgan & Claypool Publishers, 2010.
- [8] Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation. In **Proc. of EMNLP**, pp. 7764–7774, 2021.
- [9] Yuto Kuroda, Tomoyuki Kajiwara, Yuki Arase, and Takashi Ninomiya. Adversarial Training on Disentangling Meaning and Language Representations for Unsupervised Quality Estimation. In **Proc. of COLING**, pp. 5240–5245, 2022.
- [10] Keita Fukushima, Tomoyuki Kajiwara, and Takashi Ninomiya. Reversible Disentanglement of Meaning and Language Representations from Multilingual Sentence Encoders. In **Proc. of MRL**, pp. 265–270, 2025.
- [11] Tongzhou Wang and Phillip Isola. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In **Proc. of ICLR**, pp. 9929–9939, 2020.
- [12] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In **Proc. of NAACL**, pp. 48–53, 2019.
- [13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Proc. of NeurIPS**, 2017.
- [14] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka Representation Learning. In **Proc. of NeurIPS**, pp. 30233–30249, 2022.
- [15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In **Proc. of EMNLP**, pp. 38–45, 2020.
- [16] Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In **Proc. of ICLR**, 2015.
- [17] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In **Proc. of EMNLP**, pp. 6894–6910, 2021.
- [18] Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings. In **Proc. of NAACL**, pp. 4207–4218, 2022.
- [19] Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Improving In-context Learning of Multilingual Generative Language Models with Cross-lingual Alignment. In **Proc. of NAACL**, pp. 8058–8076, 2024.