AIによる意味的類似度を用いた日本語検索システムの開発

全　慶樹，近藤　正一，堀　武司

Development of Japanese Search Systems using AI-based Semantic Similarity

Keiki ZEN，Shouichi KONDOU，Takeshi HORI

抄　録

ウェブ検索エンジンなどの情報検索（Information Retrieval）システムでは，テキストの意味的な類似性（Semantic Textual Similarity, STS）を正しく認識することが技術的課題の一つとされている。近年，このようなテキストの意味的な類似度を高品質に計算する技術が開発され，検索エンジンなどへの活用が進められている。しかし，英語を中心として研究が進められているため日本語を対象とした活用事例は少なく，道内中小企業や自治体における活用もこれからの状況にある。そこで本研究では，日本語で記述されたテキストの意味的類似度にもとづいた検索技術の活用に向けて，公開されている日本語データセットを使用した高品質な意味的類似度モデルを開発した。あわせて本モデルを使用したシンプルな検索システムを開発し，検索性能と処理速度を評価した。その結果，公開されている他の意味的類似度モデルより高い検索性能を有しているとともに，標準的なPCにおいて十分な処理速度を有していることを確認した。

キーワード：AI，自然言語処理，意味的類似度，日本語，情報検索システム

Abstract

In information retrieval systems such as web search engines, one of the technical challenges is correctly recognizing semantic textual similarity (STS). In recent years, technology has been developed to calculate high-quality semantic textual similarity, and its use in search engines and other applications is progressing. However, as research is focused on English, there are few examples of its use in Japanese, and its use by small and medium-sized enterprises and local governments in Hokkaido is still in the future. Therefore, in this study, we developed a high-quality semantic similarity model using publicly available Japanese datasets, aiming to utilize search technology based on Japanese semantic textual similarity. We also developed a simple search system using this model and evaluated its search performance and processing speed. As a result, we confirmed that it has higher search performance than other publicly available semantic similarity models, and has sufficient processing speed on a standard PC.

KEY-WORDS : AI, NLP, semantic similarity, Japanese, information retrieval systems

1. はじめに

ウェブ検索エンジンなどの情報検索（Information Retrieval）システムでは，テキストの意味的な類似性（Semantic Textual Similarity, STS）を正しく認識することが技術的課題の一つとされている。人間が行うのと同様の言語理解にもとづき２つのテキストがどの程度類似しているのかを定量化できれば，ユーザーの意図を反映した柔軟な検索が可能になると考えられている。たとえば「ジャガイモ」と検索したとき，ユーザーは「ジャガイモ」に関する情報に加えて「馬鈴薯」などの同義語や類義語に関する情報も必要としている場合が多く，意味的な類似性を考慮した検索技術が重要となる。また，「病気に強いジャガイモ」と検索したとき，「病気のジャガイモ」に関する情報は不要な場合が多く，単語レベルではなくテキスト全体で意味を認識できる技術が必要とされている。近年，このようなテキストの意味的な類似度を高品質に計算

事業名：経常研究
課題名：AIを用いた自然言語処理による文書データからの情報抽出技術の研究

する技術が開発され，検索エンジンなどへの活用が進められている。これにより一種の情報検索とみなせる業務，たとえばFAQ（Frequently Asked Questions，よくある質問回答）データベースを参照してユーザーの問い合わせに回答するヘルプデスク業務などを自動化できると考えられている。しかし，英語を中心として研究が進められているため日本語を対象とした活用事例は少なく，道内中小企業や自治体における活用もこれからの状況にある。

そこで本研究では，日本語で記述されたテキストの意味的類似度にもとづいた検索技術の活用に向けて，公開されている日本語データセットを使用した高品質な意味的類似度モデルを開発し，検索性能を評価した。あわせて本モデルを使用したシンプルな検索システムを開発し，標準的なPCにおける処理速度を評価した。

2. 日本語を対象とした意味的類似度モデルの開発

本研究では，日本語Wikipediaデータセット等で事前学習済みの言語モデルにSentence-BERT[1]アルゴリズムを適用し，日本語の意味的類似度モデルを開発した。日本語の意味的類似度モデルの学習と検証には，公開されているJSNLI[2]，JSICK[3]，JGLUE[4] から構築した日本語データセットを使用した。性能評価には，自治体業務で使用されている非公開FAQデータベースから構築したデータセットを使用し，他の研究機関から公開されている意味的類似度モデルと検索性能を比較した。

2.1 アルゴリズムの選定

テキストの意味的な類似度を計算するアルゴリズムの多くは，文章や単語などの自然言語の構成要素をベクトル空間上のベクトルとしてマッピングする「埋め込み」と呼ばれる変換を行う。埋め込みにより各テキストはベクトル化されるため，ベクトル間のコサイン類似度（式１）やユークリッド距離，マンハッタン距離などの類似性尺度が適用可能となり，これをテキスト間の意味的類似度として利用する。たとえばコサイン類似度の場合，類似する２つのテキストに対応するベクトルの「なす角」が小さくなるため，類似性が高いほど値は１に近くなる（図１）。

$$\text{cosine similarity} = S(\boldsymbol{a}, \boldsymbol{b}) := \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\| \|\boldsymbol{b}\|} = \cos(\theta) \qquad (1)$$

$\boldsymbol{a}, \boldsymbol{b}$: ベクトル，$\theta$: $\boldsymbol{a}, \boldsymbol{b}$ のなす角

自然言語処理分野では，これまでも様々な埋め込みアルゴリズムが提案されており，２層のニューラルネットワークを使用したWord2vec[5] は単語の埋め込みアルゴリズムとして広く知られている。しかし，複数の単語をそのまま埋め込むことができないため，テキスト全体の意味をどのように扱うかが課題とされていた。

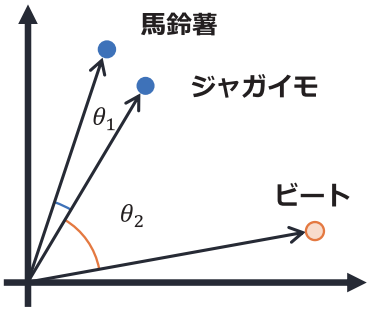

図１　Word2vec による単語埋め込みのイメージ
（類似単語のベクトルはなす角が小さい）

近年，ニューラルネットワークの一種であるTransformerを使用した言語モデルが自然言語処理の幅広いタスクに適用できることが明らかとなり，BERT[6] などの言語モデルを再学習させることで文章などの複数の単語からなるテキストの意味的類似度を高品質に計算できるようになった。これは埋め込みを介さずに意味的類似度を計算するアルゴリズムであり，２つのテキストをモデルに入力して類似度を直接計算する（図２）。このアルゴリズムは大量のパラメータを持つ大規模な言語モデルを使用するため，コサイン類似度と比べて１回の類似度計算に多くの時間を要する。したがって，ユーザーの入力に応じてリアルタイムに類似度計算を行う情報検索システムでは，検索対象の増加に伴い処理時間が増大するため利用に適さなかった。

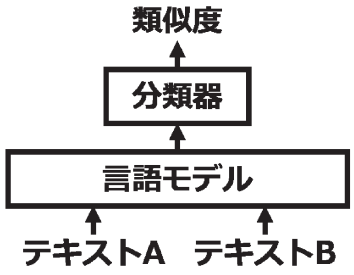

図２　言語モデルによる類似度の計算

一方，言語モデルの再学習を行わずに言語モデルの出力をそのまま埋め込みベクトルとみなして意味的類似度を計算するアルゴリズムが知られている。言語モデルはトークンと呼ばれる単位でデータの入出力を行っており，入出力はトークンの系列となる。このアルゴリズムでは，出力の先頭にあるCLSトークンや出力全体の平均を入力テキストの埋め込みベクトルとして利用する。埋め込みによる類似度計算は高速に実行できるため，情報検索システムでの利用も可能だが，再学習を行っていないため埋め込みの品質が低く，検索性能が十分ではなかった。

以上のアルゴリズムを踏まえて言語モデルを高品質な埋め込み表現となるように再学習させるSentence-BERTアルゴリズムが提案されており，情報検索システムにおいて高品質な意味的類似度が利用可能となった。本研究では，日本語の意味的類似度モデルの開発にSentence-BERTアルゴリズムを使用した。

2.2 Sentence-BERTアルゴリズムの詳細

Sentence-BERTは，高品質な埋め込み表現を獲得するために言語モデルの再学習を行うアルゴリズムである。学習時は，パラメータの共有された2つの言語モデルを出力部分で接続したネットワーク構造を持つ（図3）。学習済みモデルの利用時は，各テキストを学習済みモデルでベクトル化し，コサイン類似度などで意味的類似度を計算する。ベクトル化の計算方法として，CLSトークンを使用するCLSプーリング，最大のトークンを使用するMAXプーリング，すべてのトークンの平均を使用するMEANプーリングがあり，通常はMEANプーリングを使用する。

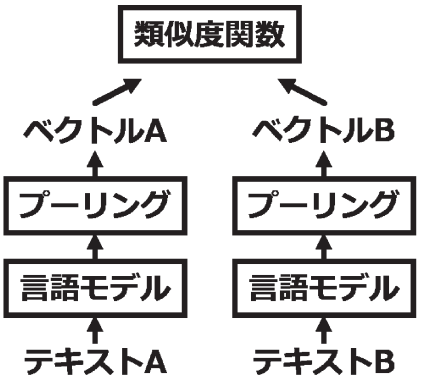

図3　Sentence-BERT のネットワーク構造

学習時の損失関数としてReimersらの論文では，Softmax Lossが使用されている。これは，ベクトルA，ベクトルB，及びその差分のベクトルを連結させ，ソフトマックス関数を計算する方法である。その他の損失関数としてMultiple Negatives Ranking Lossが知られている。これは，意味の類似しているテキストペアを複数同時に利用する計算方法であり，Softmax Lossより高品質な埋め込み表現を学習できるとされている。具体的には，類似テキストペアのデータセット（a_n, b_n）が与えられたとき，類似ペア（a_i, b_i）と同時に（a_i, $b_{j\neq i}$）を非類似ペアとして学習に利用する。空間上の類似ペア（a_i, b_i）の距離を最小化するのと同時に非類似ペア（a_i, $b_{j\neq i}$）の距離を最大化するように学習を進めることで埋め込み表現を獲得する（図4）。また，この損失関数ではテキストペアのかわりにテキストのトリプレット（三つ組）を使用することで性能を改善できる。その場合は，類似ペア（a_n, b_n）に語彙的に似ているが意味は異なるテキストc_nを追加し，（a_n, b_n, c_n）をトリプレットとして使用する（c_nはhard negativeと呼ばれる）。

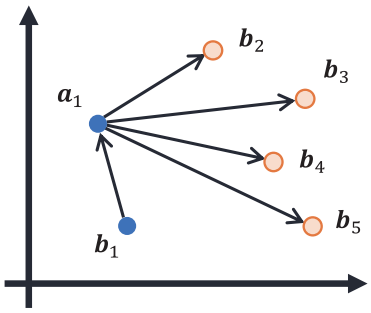

図4　複数の類似テキストペアを利用した学習

本研究では，ベクトル化の計算方法にMEANプーリングを，損失関数にMultiple Negatives Ranking Lossを使用し，学習には日本語データセットの類似テキストペアから作成したトリプレットを使用した。開発には，Reimersらが提供するSentence Transformers[7]フレームワークを使用した。

2.3 開発に使用したデータセット

Sentence-BERTによる意味的類似度モデルの開発には，意味の類似しているテキストペアのデータセットが必要となる。英語モデルの開発では，SNLIやMultiNLIといった自然言語推論（Natural Language Inference, NLI）の大規模なデータセットが利用される。本研究では，日本語の自然言語推論データセットとして利用可能なJSNLI，JSICK，JGLUEをデータセットの構築に使用した。各データセットの特徴は次のとおりである。

JSNLIは，英語のSNLIを日本語に翻訳したデータセットである。機械翻訳が使用されているが，機械翻訳の評価指標や人間による確認作業により，質の悪いデータをフィルタリングし，品質を高めている。フィルタリング後のデータ数は，学習用データが533,005ペア，評価用データが3,916ペアである。自然言語推論のデータでは，前提のテキストに対して仮説のテキストが正しければentailment，矛盾していれば

表1　JSNLI データセット

ラベル	前提	仮説
entailment	自転車で２人の男性がレースで競います。	人々は自転車に乗っています。
contradiction	５人の男の子が野球をしています。	バスケットボールをしている男の子のグループ。
neutral	白い犬が黒いロープを引く	犬は綱引きをしています。

contradiction，どちらとも判断できなければneutralのラベルが付与されている（表１）。entailmentラベルのテキストペアは意味が類似していることからSentence-BERTによる意味的類似度モデルの開発に利用できる。

次にJSICKは，英語のデータセットであるSICKを人手で日本語に翻訳し，ラベルを付与したデータセットである。データ数は，学習用データが5,000ペア，評価用データが4,927ペアであり，JSNLIと比べて少ないが人間による翻訳のためより高品質なデータセットであると考えられる。

最後にJGLUEは，日本語の言語理解能力を総合的に評価するためのベンチマーク（データセット群）である。英語の言語理解ベンチマークとしてはGLUEが広く使用されており，これに相当する日本語のベンチマークとして構築されている。翻訳を使用せず一から日本語のデータセットを構築しているため，自然な日本語テキストのデータセットであると考えられる。JGLUEは複数のデータセットから構成されており，文章分類タスクのMARC-ja，JCoLA，QAタスクのJSQuAD，JCommonsenseQA，文ペア分類タスクのJSTS，JNLIが含まれている。本研究では，JGLUEの中から自然言語推論のデータセットであるJNLIを使用した。データ数は，学習用データが20,073ペア，評価用データが2,434ペアである。

学習には，entailmentラベルのテキストペアをそのまま使用するのではなく，2.2節で述べたトリプレットを使用する。トリプレットの作成には，同じ前提テキストを持つentailmentラベルのテキストペアとcontradictionラベルのテキストペアを使用した。具体的には，entailmentペア（s, e）とcontradictionペア（s, c）から，（s, e, c）と（e, s, c）を作成した。JSNLIには同じ前提テキストを持つテキストペアが複数存在しており，この方法によりトリプレットを大量に作成できることから，学習用データセットの構築にはJSNLI（学習用）を使用した。同じ前提テキストを持つentailmentペアの選択方法には，ランダムに選択する方法とすべて選択する方法の２通りを検討した（表2）。

表２　学習用データセット

選択方法	データ数
ランダム	294,580
すべて	401,077

学習中のモデル選択で使用する検証用データセットの構築には，JSNLI（評価用），JSICK（評価用），JGLUE-JNLI（評価用）を使用した。検証は，テキストの集合から意味の類似したテキストを見つけるタスクで行うため，トリプレットの作成は不要である。データセットは対象のすべてのテキストとentailmentラベルから抽出された類似テキストペアのリストから構成される（表３）。

表３　検証用データセット

テキスト数	類似ペア数
13,202	2,873

学習済みモデルの性能評価には，モデル選択で使用した検証用データセットとは異なるものを使用する必要がある。本研究では，自治体業務で使用されている非公開FAQデータベースから性能評価で使用するテスト用データセットを構築した。FAQデータベースは，質問文と回答文のペアから構成されており，性能評価用に作業者１名に各質問文を文章（20 ～40字程度）とキーワード（２～５個程度）に要約してもらいデータセットを構築した（表4）。これらの要約された文章またはキーワードから質問文を検索するタスクにより検索性能を評価した。

表４　テスト用データセット

データ数	115
質問文平均文字数	65.6
要約文章平均文字数	36.3
要約キーワード平均個数	4.3

2.4　意味的類似度モデルの学習

Sentence-BERTによる意味的類似度モデルの学習には事前学習させたBERTなどの言語モデルが必要であり，本研究では公開されている日本語の言語モデルを使用した。表５に使用した言語モデルを示す。東北大学が公開している言語モデルの中から学習データセット，テキストのトークン化方法，モデルサイズが異なる複数のBERTモデルを使用した。また，rinna株式会社が公開している言語モデルの中から東北大学の言語モデルより大規模なデータセットで学習されている

表５　使用した日本語言語モデル

開発元	モデル名	学習データセット
東北大学	cl-tohoku/bert-base-japanese cl-tohoku/bert-base-japanese-whole-word-masking cl-tohoku/bert-base-japanese-char cl-tohoku/bert-base-japanese-char-whole-word-masking	日本語 Wikipedia（2019 年 9 月 1 日）
	cl-tohoku/bert-base-japanese-v2 cl-tohoku/bert-base-japanese-char-v2 cl-tohoku/bert-large-japanese cl-tohoku/bert-large-japanese-char	日本語 Wikipedia（2020 年 8 月 31 日）
rinna 株式会社	rinna/japanese-roberta-base	日本語 Wikipedia ＆ 日本語 CC-100

RoBERTa[8] モデルを使用した。

すべての学習において，最大入力トークン数を256と設定したほか，学習プロセスを高速化するための自動混合精度（Automatic Mixed Precision, AMP）を使用した。その他に意味的類似度モデルの性能に影響を与える設定項目が複数あり，検証用データセットでの性能を確認しながら手動で調整した（表 6 ）。

表 6　学習設定の調整

設定項目	調整範囲	調整後
言語モデル	表 5 参照	cl-tohoku/ bert-large-japanese
バッチサイズ	100～500	500
最大エポック数	1～20	15
データセット	表 2 参照	ランダム

上記設定で学習した際の検証用データセットに対する性能の推移を図 5 に示す。2 エポック目の性能が最も高く，その後低下する。これは学習用データセットに対する過学習を検証用データセットにより検出できているためと考えられる。

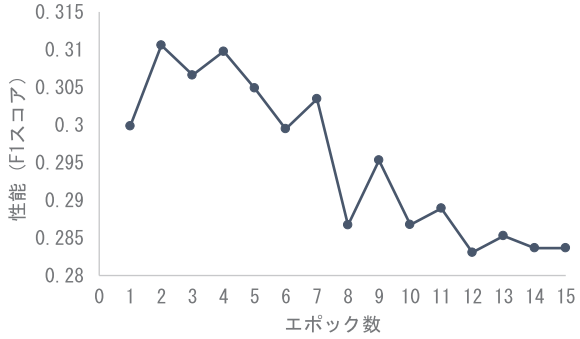

図 5　検証用データセットに対する性能の推移

2.5　学習済みモデルの性能評価

学習済みモデルの性能をテスト用データセットにより評価した。性能指標として正解テキストの検索順位にもとづくMRR（Mean Reciprocal Rank）を使用し，文章とキーワードによる検索をそれぞれ評価した。また，公開されている意味的類似度モデルには日本語に対応しているものがあり，それらのモデルと性能を比較した。具体的には，Sentence Transformersの多言語モデルと日鉄ソリューションズのSonobe Isao氏の日本語モデルを比較対象とした。結果を表 7 に示す。本研究で開発した意味的類似度モデルが，どちらの検索タスクにおいても高い性能を示すことを確認した。

3. 意味的類似度モデルを用いた検索システムの開発

開発した意味的類似度モデルはキーワード検索における性能が高いことから，検索エンジン等での活用が期待できる。そこで意味的類似度モデルを使用したシンプルな検索システムを開発し，標準的なPCにおける処理速度を評価した。

3.1　検索システムの開発

検索システムは，ウェブブラウザから利用できるウェブアプリケーションとして開発した。検索システムの概要を図 6 に示す。具体的には，ウェブサーバーであるnginxとPythonのHTTPサーバーであるGunicornがウェブアプリケーションを提供する。ウェブアプリケーションフレームワークにはPythonのFlaskを使用し，Sentence Transformersフレームワークを使用した意味的類似度モデルによる検索を実行する。

検索対象のテキストは，あらかじめ意味的類似度モデルでベクトル化されデータベースに保存されているため，検索時のベクトル化は不要である。検索時は，ユーザーが入力した検索キーワード（テキスト）をベクトル化し，データベースの中からコサイン類似度の高いテキストを検索結果として提示する。

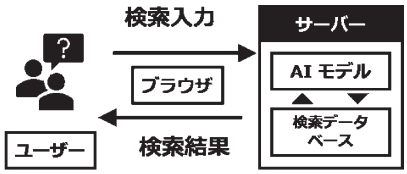

図 6　検索システムの概要

実際の検索画面を図 7 に示す。この例では，子育てに関連するFAQのデータセット（CC-BY 4.0 子育てオープンデータ協議会）を検索対象としている。ユーザーが「子供に必要なワクチン」と検索したとき「子どもの予防接種について教えてください。」というFAQが検索結果として提示されており，「子供」と「子ども」や「ワクチン」と「予防接種」の意味的な類似性を認識できていることが確認できる。

表 7　モデルの性能評価

開発元	モデル名	文章検索性能	キーワード検索性能
本研究	-	**0.989**	**0.955**
Sonobe Isao	sonoisa/sentence-bert-base-ja-mean-tokens-v2	0.968	0.877
Sentence Transformers	sentence-transformers/paraphrase-multilingual-mpnet-base-v2	0.971	0.899
	sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2	0.960	0.852
	sentence-transformers/distiluse-base-multilingual-cased-v2	0.950	0.865

図 7　システムの検索画面

3.2　検索システムの処理速度評価

本システムを標準的なPC内で構築し，処理速度を評価した（表 8 ）。検索対象として子育てFAQデータセットと和英辞書EDICT2の日本語見出しを使用した。計算には，CPU（Intel Core i5-10210U 1.6GHz）を使用し，ニューラルネットワークの計算を高速化するためのGPUアクセラレーションは使用していないが，PyTorchフレームワークによる動的量子化を検討した。毎秒処理できる検索の問い合わせ数は検索対象のデータ数の増加に伴い低下するが，小規模なシステムの運用であれば十分な処理速度であると考える。

表 8　システムの処理速度評価

検索対象	データ数	処理性能［query/s］	
		量子化なし	量子化あり
子育て FAQ	661	0.59	14.34
EDICT2	202,253	0.51	1.84

4. おわりに

本研究では，公開されているデータセットを使用して日本語を対象とした意味的類似度モデルを開発した。自治体業務で使用されているFAQを対象に検索性能を評価し，公開されている他の意味的類似度モデルより高い検索性能を有していることを確認した。あわせて検索エンジン等での活用を検討するため，ウェブブラウザから利用できる検索システムを開発し，標準的なPCにおいて十分な処理速度を有していることを確認した。開発した意味的類似度モデルは幅広い場面で活用可能であり，特にインターネット上のクラウドサーバーなどで扱うにはリスクが高い機密情報を対象とした検索システムを，オンプレミスシステム（社内で構築し運用するシステム）やスタンドアロンPCで構築する場合に有用と考える。また，OpenSearchなどの検索エンジンソフトウェアが埋め込みによる検索機能の開発を進めており，内部のモデルを本研究で開発した意味的類似度モデルに置き換えることで日本語に対する検索性能の改善が期待できる。

謝辞

本研究では，データ提供やシステムテストにおいて北海道総合政策部次世代社会戦略局情報政策課，北海道出納局総務課にご協力いただきました。ここに記して感謝いたします。

参考文献

1 ） Reimers, Nils, and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982-3992. (2019)
2 ） 日本語SNLI (JSNLI) データセット 1.1 (2020年10月 5 日) https://nlp.ist.i.kyoto-u.ac.jp/?日本語SNLI（JSNLI）データセット
3 ） 谷中　瞳，峯島宏次：JSICK: 日本語構成的推論・類似度データセットの構築，人工知能学会全国大会論文集，第35回, (2021)
4 ） Kurihara, Kentaro, Daisuke Kawahara, et al. “JGLUE: Japanese General Language Understanding Evaluation.” Proceedings of the 13th Language Resources and Evaluation Conference (LREC). pp. 2957-2966. (2022)
5 ） Mikolov, Tomas, et al. “Efficient Estimation of Word Representations in Vector Space.” International Conference on Learning Representations (ICLR) 2013. (2013)
6 ） Devlin, Jacob, et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 4171-4186. (2019)
7 ） Sentence Transformers: Multilingual Sentence, Paragraph, and Image Embeddings using BERT & Co. https://github.com/UKPLab/sentence-transformers
8 ） Liu, Yinhan, et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” arXiv preprint arXiv: 1907.11692. (2019)