技術紹介

土木特化埋め込みモデルの構築

Civil-Engineering Domain Adaptation on Japanese Embedding Model

山田　雄太 *1
YAMADA Yuta

西島　隆人 *1
NISHIJIMA Takato

四月朔日　勉 *2
WATANUKI Tsutomu

1. はじめに

当社では土木業界向けに AI を用いた検索システムの開発を進めており，その基盤として，土木知識を学習した埋め込みモデルの開発を行いました。埋め込みモデルとは単語や文章，画像の持つ意味を埋め込み表現と呼ばれる多次元ベクトルの形に変換する機械学習モデルです。ベクトル同士はコサイン類似度といった定量的指標によって比較できるため，埋め込み表現に変換することで意味的関連を考慮した検索が可能となります。しかし，汎用的な埋め込みモデルでは土木用語等の固有知識を正確に理解していません。そこで，土木学会の論文や土木工事共通仕様書などから土木に関連する文章を抽出し，それらを基に学習させることで土木知識を持つテキスト埋め込みモデルの構築をしました。

2. 土木特化埋め込みモデルの構築

(1) 土木特化埋め込みモデル構築時の課題

既存の大規模言語モデル（LLM: Large Language Model）に特定の専門分野に特化させるドメイン適応を行うには専門データセットによる学習が挙げられます。しかし，土木分野に特化したオープンなデータセットは存在しておらず，土木性能に対する評価も困難であったため，学習用・評価用データセットを自前で構築する必要があります。

また埋め込みモデルの学習では対照学習という手法が用いられ，この手法では意味的に類似・乖離した文章ペアが必要とされます。つまり，土木に関係した文書や Web 上の文章を集めるだけではなく，集めた土木文章から意味的類似・乖離を持った文章ペアを作成する必要があります。

(2) 学習用土木言語データセットの構築

土木に関する意味的類似・乖離を持った文章ペアを作成するために，収集した土木文章から QA（Question Answering），自然言語推論（NLI: Natural Language Inference）データセットの構築を進めました。 なお，QA データは質問と対応する回答に加え回答根拠となりうる文章の 3 つで構成されたデータ，NLI データはある前提文に対して同意である含意文と矛盾がある矛盾文の 3 つの文で構成されたデータです。

埋め込みモデルをはじめとした LLM の学習には膨大なデータが必要となります。近年，テキスト埋め込みの学習に LLM を用いて合成したデータを活用する研究[1)]が数多く報告されています。また合成データセットでチューニングしたモデルが人手で整備されたデータセットでチューニングしたモデルに匹敵する性能を示した報告[2)]もなされています。

以上を踏まえ，全文検索アルゴリズムによって人手に近い形で構築された NLI/QA データに加え，LLM を用いて合成した文章による学習用 NLI/QA データも学習用土木データセットに加え，合計約 540 万の土木 NLI/QA データから成るデータセットを構築しました（**表 1**）。元となる土木に関する文章は土木学会の論文誌や Web 上の土木用語集，Wikipedia の土木関連の記事から取得しました。

表 1　構築した学習用土木データセット

データ種別	データ数
NLI	3.04 M
QA	2.44 M

(3) 埋め込みモデルの学習時の課題

既存の埋め込み特化モデルに対して対照学習によるファインチューニングを行った結果を**表 2** に示します。**表 2** によると土木性能は向上した一方で，一般性能は劣化しています。

表 2　埋め込み特化モデルに対する学習による性能

	一般性能	土木性能
ファインチューニング前	79.89	87.82
ファインチューニング後	77.05	89.69

（※ 表 2 で示す性能は IR，STS タスクの平均性能）

既存の埋め込み特化モデルに対するファインチューニングでは一般性能と土木性能の両立が困難であったため，既存の言語モデルに対し埋め込み特化モデルとするための大規模ファインチューニングを行うことで一般性能と土木性能を両立する埋め込み

*1 川田テクノシステム㈱開発本部クラウド開発部 主任
*2 川田テクノシステム㈱開発本部クラウド開発部 部長

モデルの構築を目指しました。

(4) 埋め込みモデルとしての学習

学習データには構築した土木 NLI / QA データセットと一般の日本語データセットを用い，モデルの学習は 2 段階に分けて行いました。1 段階目は学習データに LLM で作成した弱教師データも含め，大量のテキストをモデルに学習を施しました。一方，2 段階目では日本語データセットは LLM を用いず作成されたもの，土木データセットは人手に近い形で作られたものと土木用語と説明文のペアを用いて，更にモデルの性能が改善するよう学習を施しました[1)]。

3. 構築した土木特化埋め込みモデルの評価

今回，埋め込みモデルの評価にあたっては，意味的類似度（STS: Semantic Textual Similarity）と情報検索（IR: Information Retrieval）タスクの性能を重要視します。日本語性能に関しては日本語埋め込みモデルの性能評価ベンチマークである JMTEB (Japanese Massive Text Embedding Benchmark）で評価しました。一方，土木性能に関しては学習データ同様にオープンなものは存在しないため構築しました。

土木性能評価用 STS データは類似度を判定する 2 文と文ペアの類似度から構成され，類似度は 6 段階で設定されます[3)]。Web 上の土木用語集から取得した文章をデータソースとしてカテゴリと全文検索アルゴリズムの評価値を基に文章ペアを構築し，全てのデータに対して人手で類似度評価を行い構築しました。一方，土木性能評価用 IR データセットは検索クエリ群とクエリに対する回答根拠が含まれる文章群の 2 つから構成されます。Web 上の土木用語集から取得した文章に対し，意味的分割と分割結果からクエリとなる質問文の生成を LLM を用いて行い，キーフレーズ抽出結果から候補を絞り込み，人手での確認を経てデータセットを構築しました。

構築した土木埋め込みモデルとオープンな日本語埋め込みモデルとの間で一般性能と土木性能を比較した結果をそれぞれ**表 3, 4** に示します（太字が当社で開発したモデル）。モデル形状が完全に同一である 70M パラメータの Ruri-v3 と比較すると，開発したモデルは上手く土木知識を学習できていると考えられます。また，ドメイン適応にあたって，一般性能の劣化も発生していないことが確認できます。

表 3　構築したモデルの一般性能

モデル名	#Param.	全体性能	STS	IR
kts-civil	**70M**	**75.77**	**83.83**	**79.38**
Ruri-v3	315M	77.24	81.22	81.89
PLaMo	1.05B	76.10	79.94	83.14
Ruri-v3	70M	75.48	79.82	79.96
Ruri-small-v2	68 M	73.30	82.91	73.94

表 4　構築したモデルの土木性能

モデル名	#Param.	土木性能	STS	IR
kts-civil	**70M**	**89.38**	**81.16**	**93.50**
Ruri-v3	315M	90.38	83.79	93.68
GLuCoSE	133M	90.18	78.79	95.88
Ruri-v3	70M	87.82	81.10	91.19
Ruri-small-v2	68 M	88.12	80.72	91.83

4. 考察

既存の埋め込みモデルに対するファインチューニングでなく，埋め込みモデルにするためのファインチューニングの方が一般性能の劣化が生じなかった理由としては破壊的忘却が発生しなかったことが考えられます。既存の埋め込みモデルに対するファインチューニングでは土木データを後から学習するため，一般的な日本語でも用いられる言葉の意味が土木分野特有の意味として上書きされる可能性があります。一方，埋め込みモデルにするためのファインチューニングでは一般の日本語データと土木データを同時に学習することができます。この方法では意味の上書きが起こらず，一般性能の劣化が起こらなかったと考えられます。

また当社で開発したモデルよりパラメータ数が多い Ruri-v3（315M）は， Ruri-v3（70M）と同一のデータセットで学習がなされています。つまり，Ruri-v3（315 M）の学習前モデルに対し今回学習で用いたデータセットで学習を行うことで，Ruri-v3（315M）より土木性能が高いモデルを構築することが可能と考えられます。

5. おわりに

当社が開発した土木特化埋め込みモデルは一般的な日本語性能（**表 3**），土木性能（**表 4**）ともに，既存の同程度のパラメータ数を持つ埋め込みモデルと比較して優れた性能を得られていることを確認しました。

今後はよりパラメータ数の大きなモデルに対して埋め込みモデルとしての同様の学習を施すことで，更に土木性能の高い埋め込みモデルの構築を進めていきます。また，開発した土木特化埋め込みモデルは basepage や保管管理システムにおける検索機能に組み込むことで，検索精度やユーザビリティの向上が期待されます。

参考文献

1) H. Tsukagoshi, R. Sasano : Ruri: Japanese General Text Embeddings, arXiv: 2409.07737, 2024.

2) 佐藤，塚越，笹野，武田：自動生成した NLI データを用いた教師なし文埋め込みの改良，言語処理学会第 30 回年次大会発表論文集， pp.1670-1675, 2024.

3) E. Agirre, *et al.* : SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation, SemEval-2016, pp.497-511, 2016.