特許からの多言語上位下位関係の抽出

岩熊耕平, 福田悟志, 難波英嗣 （中央大学）

研究背景および目的

- テキストからの用語の上位下位関係の抽出は、自然言語処理タスクにおける重要なタスクであるが、従来のベンチマークでは、一般語の上位下位関係の抽出に焦点を当てたものが中心的であった。
- 専門用語の上位下位関係の抽出に焦点を当てたデータセットとして、Google Patent Phrase Similarity Datasetが存在するが、英語のみを対照としている。
- 近年、OpenAI EmbeddingsやE5など、多言語対応の埋め込み手法が提案されている。これらの手法は、言語が異なっていても、内容が類似していれば類似した埋め込み表現が得られる。

Google Patent Phrase Similarity Datasetと多言語埋め込み手法を用い、多言語に対応した専門用語の上位下位関係の抽出器を構築

提案手法

1. パターンに基づく上位下位関係候補ペアの抽出

「などの」「等の」といった定型表現に着目し、テキストから上位下位関係候補を抽出

2. OpenAI Embedding/ E5を用いた埋め込み表現の獲得

多言語埋め込みモデルであるOpenAI Embeddings 及び E5を用いて各用語の埋め込み表現を獲得

用語ペアの類似度判定

cos類似度を用いて用語ペア間の意味的類似度を測定

3. 上位下位関係識別機の構築

得られた用語ペアの埋め込み表現を用いて上位下位関係を予測する分類器を構築し、入力された用語ペアの関係性を判定する

学習データとは異なる言語の用語間の関係性を予測

使用データ

Google Patent Phrase Similarity Dataset

- 特許用語間の関係性を学習するためのデータセット。対象言語は英語。
- 用語ペアの関係性、類似度がまとめられている。類似度が1の場合はexact、0.75は同義語、0.5は上位下位関係、0.25は部分全体関係、0は無関係となっているため、類似度が正しく予測できれば、同時に用語間の関係が推測できることになる。
- 訓練用36473件(75%)、検証用2843件(5%)、評価用9232件(20%)に分割されている

anchor	target	context	rating	score
acid absorption	absorption of acid	B08	exact	1.00
acid absorption	acid immersion	B08	synonym	0.75
acid absorption	chemically soaked	B08	domain	0.25
acid absorption	acid reflux	B08	not rel.	0.00
gasoline blend	petrol blend	C10	synonym	0.75
gasoline blend	fuel blend	C10	hypernym	0.50
gasoline blend	fruit blend	C10	not rel.	0.00
faucet assembly	water tap	A22	hyponym	0.50
faucet assembly	water supply	A22	holonym	0.25
faucet assembly	school assembly	A22	not rel.	0.00

データ例

実験結果

- **埋め込み表現による類似度評価**

用語ペア間のcos類似度を測り、PearsonおよびSpearman順位相関係数で評価

Model	Pearson cor.	Spearman cor.
Word2Vec	0.437	0.483
BERT	0.418	0.409
Patent-BERT	0.528	0.535
Sentence-BERT	0.598	0.535
OpenAI Embed	0.581	0.564
E5	0.574	0.546

OpenAI EmbeddingsおよびE5は、従来手法と同等の精度が得られている。

結論・今後の課題

- OpenAI, E5による単語埋め込みは、BERTなどの従来モデルよりも意味的類似性を表現している
- 今後は、関係性分類器の構築や日本語用語との対応付けに取り組む