DEIM2021 H33-3

ゼロショット文書分類向けの情報源領域から学習データの選択手法

大畑　直輝† 　白井　匡人†† 　若林　啓††† 　劉　健全††††

† 島根大学 自然科学研究科　〒 690-8504 島根県松江市西川津町 1060
†† 島根大学 学術研究院理工学系　〒 690-8504 島根県松江市西川津町 1060
††† 筑波大学 図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
†††† 日本電気株式会社　バイオメトリクス研究所　〒 211–8666 神奈川県川崎市中原区下沼部 1753
E-mail: †n20m101@matsu.shimane-u.ac.jp, ††shirai@cis.shimane-u.ac.jp, †††kwakaba@slis.tsukuba.ac.jp,
††††jqliu@nec.com

あらまし　本研究では，学習データが存在しない対象領域の分類器を構築するための情報源領域の学習データの選択手法を提案する．一般的な文書分類では，クラスごとの学習データから得られる各クラスの特徴を基に未知の文書を既知のクラスへ分類する．ゼロショットテキスト分類では，対象のクラスの学習データが存在しないため，クラス名から得られる特徴を基に分類を行う．本研究では，対象領域での分類先となるクラス集合が与えられた際に，対象領域の各クラスを区別するための学習データを情報源領域から選択する．これにより対象領域のラベル付きデータを用いずに文書分類が行えることを示す．

キーワード　文書分類，ゼロショット学習

1. 前　書　き

分類問題では，クラスごとの特徴に基づきクラスが未知のデータを既知のクラスに分類する．一般的な分類問題では，クラスごとに学習データが存在するため学習データを基に各クラスを識別可能な特徴を獲得する．文書をクラスごとに分類する文書分類では，政治，スポーツ，IT といったクラスは文書の集合として表現され，単語の頻度や分布を基に未知文書のクラスを推定する．各クラスに所属する文書は文書で扱われる話題に関連して多様性を持つため，各クラスの特徴を得るには十分な数のラベル付き文書が必要となる．しかし，文書のラベル付けは人手によって行われるためコストが膨大となる．十分な学習データを用意できない場合，学習データが少ないクラスはそのクラスを表す特徴を獲得することが困難となる．

転移学習は，対象領域の解析に情報源領域から得られた情報を利用する枠組みである．このため，対象領域から十分な学習データが得られない場合でも解析が行える．対象領域に少量の学習データが存在する場合，情報源領域と対象領域のラベル付き文書を基にクラスの対応付けを行う．このような設定は帰納転移学習と呼ばれる．帰納転移学習を用いた文書分類では，対象領域に存在する学習データを基に情報源領域と対象領域のクラスの対応付けを行うことで，対象領域の学習データの不足を補うことができる．対象領域に学習データが存在しない場合，クラスの特徴を学習データから得られないため情報源領域のクラスと対応付けることが困難となる．

ゼロショット文書分類では，クラス名から得られる特徴を基にクラスが未知の文書を分類する [1] [2]．単純には対象領域の各クラスのクラス名に使われている単語を分散表現によりベクトル化し，文書のベクトルと比較することで分類先のクラスを決定することができる．しかし，多くても数単語のクラス名から得られる特徴は限定的であり，各クラスの文書は同じクラス内でも話題の違いなどに影響して様々な単語分布を持つため，本来の学習データから得られるクラスの特徴とは異なることが考えられる．本研究では対象領域で分類先となるクラス名の集合が与えられているという仮定の基で，対象領域のクラスを区別するための学習データを情報源領域から選択する手法を提案する．ここで情報源領域はラベル無し文書で構成される大規模な文書集合である．選択された学習データを基に分類器を構築することで，対象領域のラベル無し文書を各クラスに分類する．

第 2 章では関連研究について述べ，第 3 章では提案手法について述べる．第 4 章では実験により提案手法の有効性を示す．第 5 章で結論とする．

2. 関 連 研 究

2.1 ゼロショット学習

画像分類では，分類先となるクラス名自体が持つ特徴を用いて学習データに存在しないクラスの画像を分類可能な手法が提案されている [3] [4] [5] [2]．この学習手法はゼロショット学習と呼ばれている．多くのゼロショット学習手法では，クラス名と画像の特徴を共通の空間に埋め込む．これにより，クラス名同士の類似とクラス名が表す画像の特徴を捉えられるため，新規クラスの分類が可能となる．クラス名が持つ特徴を分類に活用するゼロショット学習の仕組みは文書分類にも適用できることが考えられるが，一般的な文書分類ではクラス名は単なる識別子としてしか扱われていない．

word2vec は分布仮説に基づき，該当の単語の周辺に出現する単語によって単語の意味を決定する [6]．クラス名が持つ特徴を分散表現で表すことによってクラスが持つ情報と文書の比較が可能となる．Veeranna らは，クラス名とテスト文書の全ての n-gram(n=1,2,3) に対してコサイン類似度を計算し，類

似度が閾値を超えるクラスに割り当てるマルチラベル文書分類手法を提案している [1]．Pushp らはクラス名のベクトルを単語のベクトルに連結し，該当するクラスである場合に 1，該当しない場合に 0 を出力する LSTM を学習する手法を提案している [7]．Ye らは，文章の最後尾にクラス名を付与した時にそのクラスに該当する場合は 1，該当しない場合は 0 を出力する BERT ベースの分類モデルを提案している [8]．さらに対象領域のラベル無し文書を用いてモデルを更新する自己学習を行うことで精度が上昇することを示している．

これらの手法では，分類先のクラス集合を仮定していないため，対象領域のクラスが一部しか定まっていない場合でも分類を行うことができる．しかし，クラス名と文章を入力とする 2 値分類器を学習するため，ラベル付き文書が必要となる．また，構築される分類器は対象領域のクラス集合を区別するための分類器ではないため，十分な精度が得られていない．

3. 提 案 手 法

3.1 問 題 設 定

本稿は，対象領域のクラス集合 $Y = \{y_1, y_2, ..., y_{|c|}\}$ が与えられた基で，対象領域の文書 $T = \{t_1, t_2, ..., t_i\}$ のクラス y^t を推定する．ここで各クラス名は単なる識別子ではなく,「経済」,「スポーツ」といったクラスに所属する文書の基準となる意味を持つ単語，または基準となる意味を説明する文章であることを仮定する．提案手法は，対象領域のクラスを区別する分類器を構築するためにクラス名が持つ特徴を基に情報源領域から各クラスの学習データとして使用する文書を選択する．情報源領域は大規模な文書の集合であり，その文書 $S = \{s_1, s_2, ..., s_j\}$ は，全てラベル無し文書である．提案手法では情報源領域，対象領域ともにラベル付けされた学習データは使用しない．

3.2 学習データの選択

本研究では，対象領域のクラスを区別するための学習データを情報源領域から選択する手法を提案する．対象領域にラベル付きの学習データが存在しないため，対象領域のクラス名と情報源領域の文書との類似度から情報源領域のラベル無し文書に擬似的なクラスラベルを付与する．学習データの選択では，対象領域のクラスを区別して特徴付けるために，類似度 1 位のクラスと 2 位のクラスの類似度の差を基に対象のクラスにのみ類似する文章を抽出する．学習データ選択の手順を以下に記述する．

1. 文章とクラス名の単語を分散表現によりベクトル化する．
2. 情報源領域のラベル無し文書に対して対象領域の各クラスとのコサイン類似度を計算する．
3. 類似度の上位のクラスから順にランキングする．
4. ランキング 1 位のクラスの類似度の値とランキング 2 位のクラスの類似度の値の差を計算し，1 位と 2 位の差が設定された閾値より大きい場合，1 位のクラスを疑似ラベルとして付与する．
5. 情報源領域の文書数分だけ繰り返し各クラスの疑似的な学習データを生成する．最終的に類似度の差が上位の k 個の文書を各クラスの学習データとして用いる．

手順 1 では大規模な学習データにより事前に学習された word2vec モデルを用いて情報源領域の文章と対象領域の各クラスの分散表現を獲得する．ここでは文章の分散表現は文章に出現する単語のベクトルの平均とする．

手順 2 において情報源領域の文章のベクトルを q，対象領域のクラスのベクトルを d とし，$|V|$ はベクトルの次元数とすると，コサイン類似度 $cos(q, d)$ は以下の式で求まる．

$$cos(q, d) = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}} \quad (1)$$

手順 3 では，対象領域のクラスの中で情報源領域の文書に関連している可能性が高いものを決定する．ランキングは以下の式で計算する．$SORT$ は括弧で囲まれた範囲を降順に並び替える処理を行う．

$$rank(|c|) = SORT\{cos(q, d_1), cos(q, d_2), ..., cos(q, d_{|c|})\} \quad (2)$$

手順 4 では，対象領域の該当するクラスにのみ類似する文書を選択するために，ランキング 1 位とランキング 2 位の差 $diff$ が閾値 $Threshold$ 以上となる文書を選択する．

$$diff = rank(1) - rank(2) \quad (3)$$

$$\begin{cases} \text{疑似ラベルを付与する} & (diff > Threshold) \\ \text{疑似ラベルを付与しない} & (diff \leqq Threshold) \end{cases} \quad (4)$$

手順 5 では，選択された文書数はクラスごとに異なるため，全てのクラスの文書数を揃える．文書数を揃える際は，ランキング 1 位と 2 位の差が大きい文書から順に選択する．

3.3 TextCNN を用いた文書分類

対象領域の文書を分類するために疑似ラベルを付与された学習データを基に TextCNN [9] を用いて分類を行う．TextCNN は CNN と同様に入力層，畳み込み層，プーリング層，全結合層で構成される文書分類モデルである．TextCNN の入力は分散表現で表された単語の系列であり，畳み込み層で異なる長さのフィルタを用いることで単語の系列の意味を考慮することができる．提案手法である情報源領域からの学習データの選択と TextCNN までの全体の流れを図 1 に示す．

4. 実　　験

4.1 データセット

情報源領域の文書集合には，20newsgroups，yahoo topic，REUTER データセットを使用する．20newsgroups データセットは 20 の異なるニュースグループからなる文書データである．各クラスに約 1,000 のデータがあり，合計は約 20,000 文書である．yahoo topic データセットは 10 のクラスが存在する大規模なデータセットである．このデータセットは，各 5 クラスずつの v0 と v1 に分けられている．各クラスに 130,000 の文書データがあり，合計は 1,300,000 文書である．Reuters Corpus は 1996 年 8 月 20 日から 1997 年 8 月 19 日までの 1 年分のニュース記事であり，1 つの記事に 128 種類からなるラベルが複数付

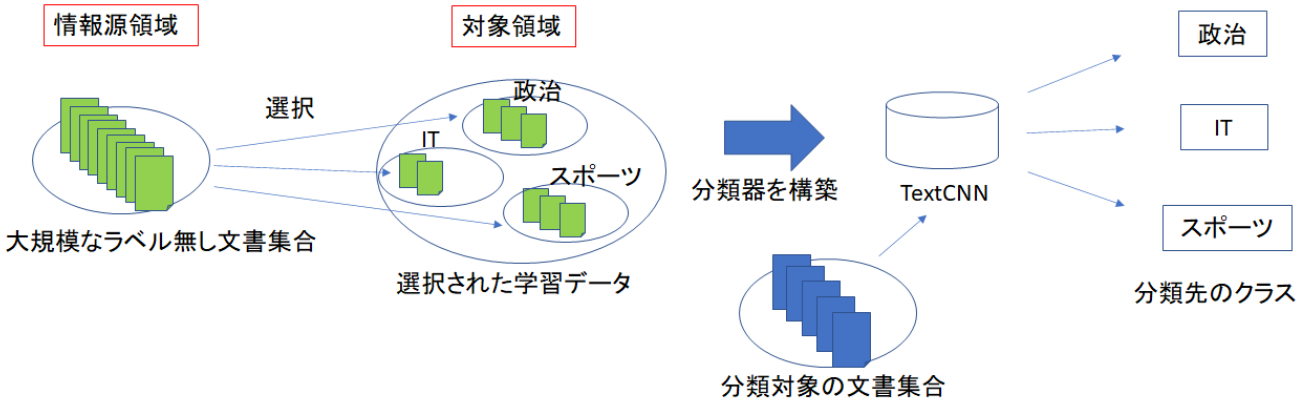

図 1　文書分類の流れ

いている．本研究では出現頻度が上位となる 60 クラスを使用する．合計の文書数は約 760,000 である．対象領域のテストデータには，DBpedia を使用する．DBpedia には，wikipedia から収集された 14 クラスの文書が含まれている．データセットは train と test に分かれており，各クラスには 40000 の学習データと 5000 のテストデータがある．提案手法では学習データのラベルは使用せず，ラベル無し文書として情報源領域に使用する．word2vec モデルは，Google news の記事から学習された事前学習済みモデルを使用する．情報源領域のデータセットを表 1 に示す．また，対象領域の評価に用いるデータセットを表 2 に示す．提案手法のパラメータとして，類似度の差の閾値は 0.05 とする．また，各クラスで類似度の差が上位となる 3000 文書を学習データとして使用する．

文書データ	クラス数	全文書数
20news	20	18,828
yahoo topic	10	1,300,000
REUTER	60	762,027
DBpedia$_{train}$	14	560,000

表 1　情報源領域の文書

文書データ	クラス数	各クラスの文書数	全文書数
DBpedia$_{test}$	14	5000	70000

表 2　対象領域の評価に用いる文書

4.2　評 価 方 法

評価指標として，再現率，適合率，F 値を使用する．これらは文書分類タスクの精度評価を行う一般的な評価指標である．再現率は，正解データのうちどれだけ正解と予測できたかの割合を表す．適合率はあるクラスであると予測したデータのうち実際に正解した割合を表す．F 値は再現率と適合率の調和平均から求める．これらは 0 から 1 の値を取り，1 に高いほど精度が高いことを示す．

4.3　比 較 手 法

比較手法には以下の 3 つの手法を用いる．

(1)word2vec:分散表現を用いてクラス名から取得したベクトルと文書のベクトルのコサイン類似度を計算し，類似度が最上位のクラスへ分類する．

(2)labelsimilarity:クラス名から取得したベクトルと文書に出現する全ての n-gram(n=1,2,3) に対してコサイン類似度を計算し，最上位の類似度が閾値を上回るとき，そのクラスに割り当てる．

(3)BERT:情報源領域の学習データを基に，文章の最後尾にクラス名を加えたものを入力として，該当するクラスである場合 1，該当しない場合 0 を出力する BERT モデルを構築する．情報源領域の学習データには yahoo topic データセットの 5 つのクラスからなる v0 を使用する．

4.4　実 験 結 果

実験結果を再現率，適合率，F 値の順で示す．word2vec を用いた場合は 0.595，0.518，0.511，labelsimilarity を用いた場合は 0.538，　0.531，　0.505，BERT を用いた場合は 0.547，0.464，0.445 となる．提案手法を用いた場合は 0.850，0.828，0.807 となる．F 値の平均は提案手法が最も高くなっている．

5

4.5　考　　察

表 3 より提案手法の F 値の平均は既存手法の word2vec から 0.296，labelsimilarity から 0.302，BERT から 0.362 改善していることがわかる．提案手法では，情報源領域のラベル無し文書を選択することによって対象領域のクラスを区別するための分類器を構築しているために，精度が上昇したと考えられる．しかし，図 2 より”Animal”や”Mean Of Transportation”クラスのように極端に低い精度となるクラスも発生している．表 4 より，この 2 つのクラスのベクトルは他のクラスと比較して似たクラスが存在しているわけではないが，クラスを特徴付けるための学習データが選択できていない．表 5 より，提案手法において情報源に DBpedia$_{train}$, $DBpedia_{test}$以外のデータセットも追加した方が精度が高いため情報源に他のデータセットを利用することが有効であることがわかる．表 6より, *Animal* クラスにおいて実際に選択されるデータ数が少ないことが原因で精

	word2vec			labelsimilarity			BERT			提案手法		
クラス名	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
Company	0.854	0.416	0.560	0.580	0.598	0.589	0.499	0.290	0.367	0.709	0.696	**0.703**
Educational Institution	0.501	0.755	0.602	0.426	0.604	0.500	0.821	0.809	0.815	0.841	0.972	**0.901**
Artist	0.861	0.331	0.479	0.597	0.405	0.483	0.220	0.572	0.318	0.894	0.926	**0.910**
Athlete	0.968	0.645	0.774	0.822	0.694	0.752	0.433	0.431	0.432	0.961	0.990	**0.975**
Office Holder	0.432	0.349	0.386	0.454	0.190	0.268	0.629	0.851	0.724	0.884	0.791	**0.835**
Mean Of Transportation	0.423	0.473	**0.447**	0.287	0.088	0.135	0.415	0.153	0.224	0.813	0.156	0.261
Building	0.611	0.474	0.534	0.527	0.365	0.431	0.326	0.199	0.247	0.625	0.955	**0.756**
Natural Place	0.201	0.496	0.286	0.171	0.169	0.170	0.313	0.947	0.471	0.549	0.909	**0.684**
Village	0.643	0.984	0.778	0.388	0.996	0.559	0.946	0.176	0.297	0.975	0.996	**0.986**
Animal	0.754	0.064	0.118	0.614	0.241	0.346	0.893	0.312	0.192	0.559	0.075	**0.462**
Plant	0.615	0.485	0.542	0.751	0.635	0.688	0.854	0.735	0.790	0.940	0.957	**0.949**
Album	0.785	0.931	0.852	0.711	0.920	0.802	0.613	0.186	0.286	0.964	0.988	**0.976**
Film	0.767	0.739	0.753	0.730	0.943	0.823	0.668	0.527	0.589	0.944	0.976	**0.960**
Written Work	0.509	0.624	0.561	0.471	0.581	0.520	0.451	0.496	0.473	0.900	0.973	**0.935**
macro avg	0.595	0.518	0.511	0.538	0.531	0.505	0.547	0.464	0.445	0.850	0.828	**0.807**

表 3　提案手法の結果

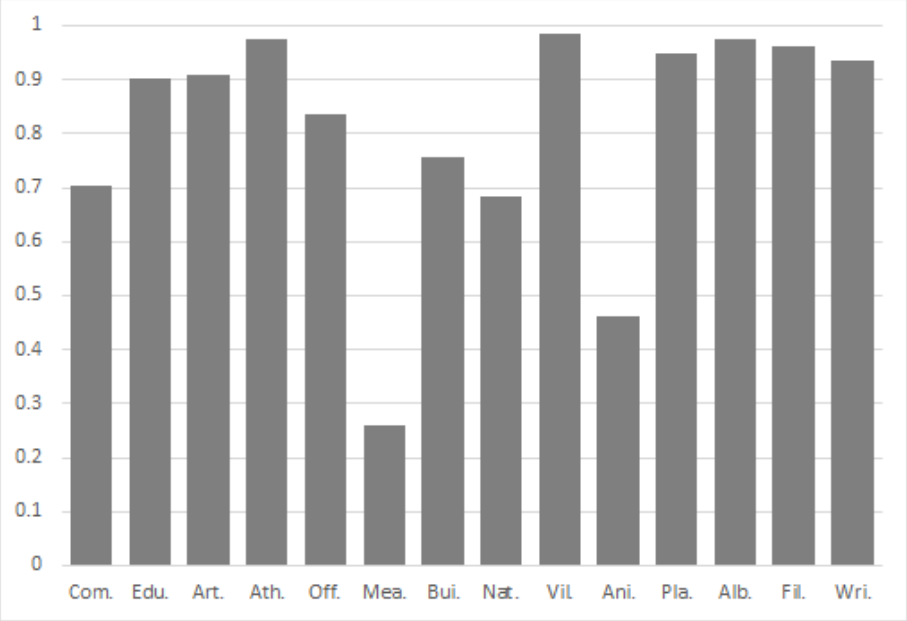

図 2　提案手法における各クラスの F 値

度が悪い可能性があることを示してる．

5. 結　　論

本研究では，学習データの無い対象領域の文書を分類するためにラベル無しの情報源領域から対象領域の各クラスの学習データを選択する手法を提案した．提案手法ではクラス名と文書の類似度を計算し，対象のクラスにのみ類似度が高い文書を選択することでラベル付き文書を用いずに対象領域のクラスを特徴付けることができる．実験により比較手法と比較して提案手法が有効であることを示した．今後の課題として，提案手法においても極端に分類精度の低いクラスが発生するため，このようなクラスの精度を改善する必要がある．

文　　献

[1] S.P. Veeranna, J. Nam, E.L. Mencıa, and J. Fürnkranz, “Using semantic similarity for multi-label zero-shot classification of text documents,” Proceeding of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Belgium: Elsevier, pp.423–428, 2016.

[2] L. Zhang, T. Xiang, and S. Gong, “Learning a deep embedding model for zero-shot learning,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.2021–2030, 2017.

[3] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, “Label-embedding for attribute-based classification,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.819–826, 2013.

[4] E. Gavves, T. Mensink, T. Tommasi, C.G. Snoek, and T.

	Com.	Edu.	Art.	Ath.	Off.	**Mea.**	Bui.	Nat.	Vil.	**Ani.**	Pla.	Alb.	Fil.	Wri.
Com.	1.000	0.186	0.082	0.070	0.155	0.130	0.040	0.165	0.184	0.034	0.193	0.042	0.007	0.132
Edu.	0.186	1.000	0.207	0.129	0.214	0.173	0.358	0.250	0.176	0.121	0.159	0.193	0.286	0.191
Art.	0.082	0.207	1.000	0.357	0.033	0.141	0.121	0.206	0.102	0.170	0.096	0.483	0.365	0.203
Ath.	0.070	0.129	0.357	1.000	0.054	0.137	0.093	0.169	0.156	0.160	0.100	0.198	0.140	0.163
Off.	0.155	0.214	0.033	0.054	1.000	0.158	0.198	0.083	0.086	0.056	0.130	0.094	0.090	0.140
Mea.	0.130	0.173	0.141	0.137	0.158	1.000	0.203	0.228	0.086	0.150	0.202	0.217	0.202	0.278
Bui.	0.040	0.358	0.121	0.093	0.198	0.203	1.000	0.301	0.200	0.207	0.230	0.085	0.200	0.225
Nat.	0.165	0.250	0.206	0.169	0.083	0.228	0.301	1.000	0.354	0.259	0.258	0.165	0.155	0.226
Vil.	0.184	0.176	0.102	0.156	0.086	0.086	0.200	0.354	1.000	0.118	0.105	0.069	0.085	0.086
Ani.	0.034	0.121	0.170	0.160	0.056	0.150	0.207	0.259	0.118	1.000	0.198	0.059	0.210	0.117
Pla.	0.193	0.159	0.096	0.100	0.130	0.202	0.230	0.258	0.105	0.198	1.000	0.162	0.087	0.085
Alb.	0.042	0.193	0.483	0.198	0.094	0.217	0.085	0.165	0.069	0.059	0.162	1.000	0.272	0.150
Fil.	0.007	0.286	0.365	0.140	0.090	0.202	0.200	0.155	0.085	0.210	0.087	0.272	1.000	0.233
Wri.	0.132	0.191	0.203	0.163	0.140	0.278	0.225	0.226	0.086	0.117	0.085	0.150	0.233	1.000
平均	0.173	0.260	0.255	0.209	0.178	0.236	0.247	0.273	0.200	0.204	0.215	0.228	0.238	0.231

表 4　クラス同士のコサイン類似度

情報源に利用するデータ	再現率	適合率	F 値
DBpedia_{train}	0.758	0.593	0.620
DBpedia_{test}	0.753	0.450	0.497
topic, reuter, 20news, DBpedia_{train}	0.850	0.828	0.807

表 5　情報源領域の使用文書における精度の比較

	DBpedia_{train}	topic, reuter, 20news, DBpedia_{train}
Com.	8927	68569
Edu.	30828	36738
Art.	7459	7930
Ath.	14665	19095
Off.	3804	87405
Mea.	7644	209047
Bui.	8642	9793
Nat.	22931	37266
Vil.	36284	37367
Ani.	420	4210
Pla.	10952	13149
Alb.	33741	40499
Fil.	20753	23988
Wri.	17699	21895

表 6　各クラスで選ばれた学習データの数

Tuytelaars, “Active transfer learning with zero-shot priors: Reusing past datasets for future tasks,” Proceedings of the IEEE International Conference on Computer Vision, pp.2731–2739, 2015.

[5] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, “Latent embeddings for zero-shot classification,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.69–77, 2016.

[6] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Advances in neural information processing systems, vol.26, pp.3111–3119, 2013.

[7] P.K. Pushp and M.M. Srivastava, “Train once, test anywhere: Zero-shot learning for text classification,” arXiv preprint arXiv:1712.05972, 2017.

[8] Z. Ye, Y. Geng, J. Chen, J. Chen, X. Xu, S. Zheng, F. Wang, J. Zhang, and H. Chen, “Zero-shot text classification via reinforced self-training,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp.3014–3024, 2020.

[9] Y. Kim, “Convolutional neural networks for sentence classification,” arXiv preprint arXiv:1408.5882, 2014.