*INTERSPEECH 2021*
30 August – 3 September, 2021, Brno, Czechia

# Semantic Transportation Prototypical Network for Few-shot Intent Detection

*Weiyuan Xu*[1], *Peilin Zhou*[1], *Chenyu You*[2], *YueXian Zou*[1,3,*]

[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]Department of Electrical Engineering, Yale University, CT, USA
[3]Peng Cheng Laboratory, Shenzhen, China

xuwy@stu.pku.edu.cn, {zhoupl,zouyx}@pku.edu.cn, chenyu.you@yale.edu

## Abstract

Few-shot intent detection is a problem that only a few annotated examples are available for unseen intents, and deep models could suffer from the overfitting problem because of scarce data. Existing state-of-the-art few-shot model, Prototypical Network (PN), mainly focus on computing the similarity between examples in a metric space by leveraging sentence-level instance representations. However, sentence-level representations may incorporate highly noisy signals from unrelated words which leads to performance degradation. In this paper, we propose Semantic Transportation Prototypical Network (STPN) to alleviate this issue. Different from the original PN, our approach takes word-level representation as input and uses a new distance metric to obtain better sample matching result. And we reformulate the few-shot classification task into an instance of optimal matching, in which the key word semantic information between examples are expected to be matched and the matching cost is treated as similarity. Specifically, we design Mutual-Semantic mechanism to generate word semantic information, which could reduce the unrelated word noise and enrich key word information. Then, Earth Mover's Distance (EMD) is applied to find an optimal matching solution. Comprehensive experiments on two benchmark datasets are conducted to validate the effectiveness and generalization of our proposed model.

**Index Terms**: few-shot learning, intent detection, metric learning, spoken language system

## 1. Introduction

Intent detection (ID) is a challenging task in building task-oriented spoken dialogue systems, which aims to capture underlying intents from given utterances. During past years, deep learning methods such as convolutional neural network (CNN) [1, 2, 3], recurrent neural network (RNN) [4, 5, 6] and graph neural network (GNN) [7] have been applied to this task and achieved excellent performance. Moreover, pre-trained language models [8, 9, 10] have also been explored to better detect the intent implied in the given utterance. Despite of the promising results of these models, they often require large amounts of labelled data, which is impractical in real world applications because annotation is both time-consuming and costly.

In this paper, we focus on intent detection with only a few training samples available, which is formally called few-shot intent detection problem. Figure 1 gives an example of 3-way-1-shot intent detection, where only one annotated sample is available for each three new intents. Essentially, few-shot intent detection is a few-shot text classification (TC) task which has become a hot topic recent years. Based on which aspect is the key point, data or model, we divide existing few-shot TC methods into two categories: *data augmentation* (DA) and *model improvement*. DA aims to enrich the annotated data and reuse deep models. [12, 13] proposed pseudo-labeling methods for unlabelled data, using semi-supervised learning. But the acquisition of the unlabelled data is also difficult. In order to tackle the hindrance of collecting unlabelled data, several generative DA approaches are given. [14] proposed back-translation which utilizes language translation; [15] put forward Easy Data Augmentation that consists of five word operations, such as synonym replacement. Though DA can address the data scarcity to some extent, it still suffers from the incorrectness of augmented data and error can be accumulated to downstream tasks. Thus, *Model improvement* is introduced as another solution, which refers to designing dedicated models. Some end-to-end few-shot classification models are discovered in computer vision (CV) field, including Matching Networks [16], Prototypical Networks (PN) [17] and Relation Networks [18]. Generally, these models compute the similarity between query and support set and then, classify the query into the class with highest rank. [19] investigated the model performance on few-shot TC task in order to check if these models are suitable for nature language processing (NLP). They first used an encoder to extract sentence-level vector, which can be regarded as global feature vector, from word embedding. Then, they fed the vector into aforementioned few-shot models and conducted experiments on four text classification dataset. Their experiment results demonstrates the superior performance of PN in few-shot TC tasks.

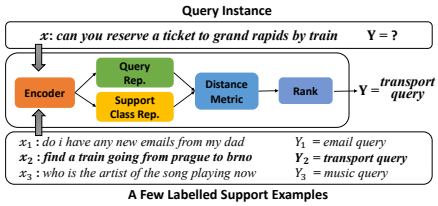



Figure 1: *3-way-1-shot intent detection on NLUE [11]*

Although existing methods have achieved promising results, we observe that unrelated word may drive global feature representations from same class far apart in a metric space. Abundant data and large model size could alleviate the data noise to some extent, but the noise negatively impacts the few-shot model whose parameters and labelled training data are few [20]. Moreover, the global feature may lose the discriminative information maintained by words in sentence. Consider the sentence *"Book a train earlier in case of traffic jam"* with intent *Transport ticket*. We can easily figure out its intent rather than *query weather* when noticing the key words *"book"* and *"train"* while other words gives less information. Therefore, a new distance metric is needed, which is capable of fully uti-

* Corresponding author.



http://dx.doi.org/10.21437/Interspeech.2021-548

lizing discriminative information of words and minimizing the noise from irrelevant information.

To alleviate above-mentioned issues, we propose a novel metric-based method, Semantic Transportation Prototypical Network (STPN), to solve few-shot intent detection problem. Different from the original Prototypical Networks, our approach takes word-level local representation matrix as input and uses a new distance metric to obtain better sample matching performance. Specifically, we reformulate the sample matching problem into transportation optimization problem and adopt Earth Mover's Distance (EMD) [21] to find an optimal solution. Our model will be further detailed in section 2.4. Our contribution are three parts: 1)To the best of our knowledge, STPN is the first few-shot intent detection model focusing on word-level discriminative information, which is consistent to human cognition; 2) We adopt EMD as a distance metric to intent classification, providing innovation of similarity measurement between sentences, and we design Mutual-Semantic mechanism to obtain semantic weight of words; 3) Experimental results on a intent detection dataset and a text classification dataset demonstrate the effectiveness and generalization of our method.

# 2. The Proposed Approach

## 2.1. Task Definition

Suppose we have a large labelled training set $D_{train}$. Our goal is to develop a classifier that learns knowledge from $D_{train}$, so that it can make predictions over new classes, in which we only have a few annotated examples. In general, these few annotated data consist of support set $D_{support}$, and a query set $D_{query}$ is also included to test the accuracy of the classifier on new classes. Formally, if the support set contains $K$ labelled examples for each of the $N$ unique classes, we call this few-shot problem a $N$-way-$K$-shot problem. For this paper we consider $N = 5$ and $K = 5$ or 1. Obviously, $K$ is too small to train a good supervised classifier because such a data-scarcity setting will bring about an overfitting problem. Therefore, we follow the episode training strategy [16] to construct multiple training episodes for $N$-way-$K$-shot problem. For each training episode, we firstly sample a $N$-way-$K$-shot support set $S$ and a query set $Q$ from $D_{train}$, and then, both $S$ and $Q$ are fed to the model to minimize loss. In the testing phase, same episode mechanism is applied to report the model performance on new classes which have never appeared in the training phase.

## 2.2. Revisiting Earth Mover's Distance

The Earth Mover's Distance (EMD), which is also known as Wasserstein distance in mathematics, is often utilized to measure the distance between two sets of weighted objects or distributions. The EMD is initially regarded as a solution to the transportation problem: Suppose that a set of suppliers $\mathcal{S} = \{s_i\}_{i=1}^{m}$ are required to transport something to a set of demanders $\mathcal{D} = \{d_j\}_{j=1}^{k}$ and the transportation cost $C = \{c_{i,j}\}, i \in [1, m], j \in [1, k]$ is given. The goal of EMD is to find a least-expensive flow $\widetilde{X}$, in the transportation plan set $\mathcal{X} = \{x_{i,j}\}$ where $i \in [1, m], j \in [1, k]$. Formally, EMD can be formulated as a linear programming problem:

$$\begin{aligned} \underset{x_{ij}}{\text{minimize}} \quad & \textstyle\sum_{i=1}^{m}\sum_{j=1}^{k} c_{ij}x_{ij} \\ \text{subject to} \quad & x_{ij} \geqslant 0, i \in [1, m], j \in [1, k] \\ & \textstyle\sum_{j=1}^{k} x_{ij} = s_i, \quad i \in [1, m] \\ & \textstyle\sum_{i=1}^{m} x_{ij} = d_j, \quad j \in [1, k] \end{aligned} \tag{1}$$

where $s_i$ represents the supply units of supplier $i$, $d_j$ denotes the demand of $j$-th demander, and $x_{i,j}$ represents the number of units transported from supplier $i$ to demander $j$.

## 2.3. Prototypical Networks

Prototypical Networks (PN) were introduced by [17], and [19] has proved its effectiveness in both CV and NLP field. The key idea behind PN is very simple and straightforward. It learns a feature space and computes the distance between samples in that space to make classification.

Formally, given an instance $S_i^c = \{w_1, \ldots, w_T\}$, which is the $i$-th sample for class $c$ with $T$ words, it will be embedded to the semantic space as $X_i^c = [x_1, \ldots, x_T]$, where $x_i = W_e w_i \in \mathbb{R}^d$ and $W_e$ refers to the embedding matrix whose dimension is $d$. Several methods can be used to obtain the instance feature vector, like averaging and CNN. Once $V_i^c$, the vectors of instances belonging to class $c$ are obtained, the prototype vector for class $c$ can be calculated by:

$$P^c = \frac{1}{K}\sum_{i=1}^{K} \mathcal{F}_\theta(V_i^c) \tag{2}$$

where the $\mathcal{F}_\theta$ represents a learnable linear function that could transform feature spaces to better distinguish different classes. Afterwards, given a query vector $V^q$, the similarity between the query and the prototype of class $c$ can be computed as follows:

$$s^{q,c} = \frac{exp(\mathcal{D}(V^q, P^c))}{\sum_{c'=1}^{N} exp(\mathcal{D}(V^q, P^{c'}))} \tag{3}$$

where $\mathcal{D}$ is a distance metric function. Note that in the origin PN paper, the authors only used the euclidean distance as $\mathcal{D}$, but in our experiments the cosine distance will also be evaluated for a more comprehensive comparison.

## 2.4. Semantic Transportation Prototypical Networks

During past years, although a lot of few-shot learning methods have been proposed, Prototypical Networks still work as a strong baseline. This claim could be demonstrated in [19] because, with the same experimental setting, Prototypical Networks perform better than many few-shot TC methods discovered over the last few years. Besides, [19] also proved that there are two key factors that determine PN's performance: data representation and the distance metric. Therefore, inspired by these insights, we propose Semantic Transportation Prototypical Network (STPN) model for few-shot intent detection task, which is illustrated in Figure 2. STPN is the upgrade version of origin PN, taking both two influential factors into consideration. For one thing, instead of computing the similarity between sentence-level representations, our STPN tends to utilize the discriminative information of words, which are more diverse and expressive. For another thing, we replace the distance metric with EMD to transport semantic information and retrieve a better matching result.

Our approach first decomposes the input sentences into local representations and then, obtains the optimal matching cost from EMD to measure the distance between input sentences. Concretely, we use pre-trained language model as the sentence embedding matrix $\mathbf{Q} \in R^{T \times d}$, where $T$ and $d$ denote the sentence length and embedding dimension. Each row of sentence embedding matrix is the embedding vector of the corresponding word in sentence and can be seen as the local word feature.

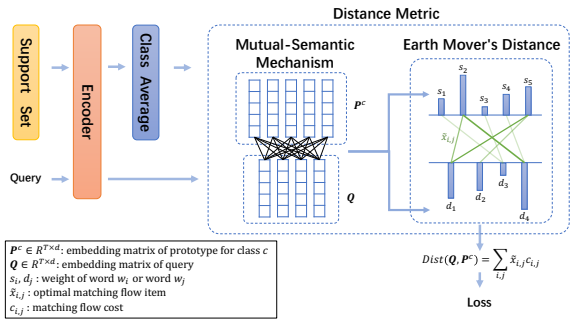



Figure 2: *Semantic Transportation Prototypical Networks.*

Thus, the similarity between two sentences can be measured by the semantic matching cost between local feature collections.

Following the origin EMD equation in (1), we obtain the cost unit by computing the pairwise cosine similarity between local word features $\mathbf{u}_i$ and $\mathbf{v}_j$ from two sentences:

$$c_{ij} = 1 - \frac{\mathbf{u}_i^T \mathbf{v}_j}{||\mathbf{u}_i|| ||\mathbf{v}_j||} \tag{4}$$

where fewer semantic matching cost are needed for similar words. Next, the semantic weight of word is computed, with larger weight word playing a more important role in the comparison, and smaller ones hardly influencing the matching result. This can be interpreted as the key word matching mechanism and is consistent with human intuitions: people can separate similar short sentences with their semantic overlap words. The semantic overlapping degree of the words is computed by our proposed Mutual-Semantic mechanism, shown as (5) and (6). We perform dot product between local word feature in sentences to generate the semantic relevance of words:

$$s_i = max(\frac{1}{T} \sum_{j=1}^{T} \mathbf{u}_i^T \mathbf{v}_j, 0) \tag{5}$$

where $max(\cdot)$ ensures the weight to be non-negative and $d_j$ can be obtained in the same manner. Then, the weights are normalized to make sure collections in both side share the same size:

$$\tilde{s}_i = \frac{T s_i}{\sum_{j=1}^{T} s_j} \tag{6}$$

Finally, the optimal matching flows $\widetilde{X}$ is acquired by solving EMD and the similarity score $s$ between query sentence and prototype sentence can be computed by (7).

$$s(Q, P^c) = \sum_{i=1}^{T} \sum_{j=1}^{T} c_{i,j} \tilde{x}_{i,j} \tag{7}$$

# 3. Experiments

## 3.1. Datasets

We evaluate our approach on **NLUE**, a dialogue intent detection dataset and **Huffpost** [22], a traditional text classification dataset. **NLUE** is released by [11] and we utilize a subset of utterances covering 64 intents. The vocab size is 2074 and its average sentence length is 6. We randomly choose 30 intents as training data and 8 intents as validation data, while the remaining 26 intents are considered as test data. **HuffPost** headlines consists of news headlines published on HuffPost between 2012 and 2018. [22] split these headlines into 41 classes. The vocab size is 8218 and its average sentence length is 11. The sentences of huffpost are less grammatical than formal sentences.

## 3.2. Baselines and Implementation Details

Based on the fundamental works from [19], where PN shows the best performance on TC task, we set up various combinations of word embedding techniques, sentence representations and distance metric for PN as baselines.

**Word Embedding Techniques**. We use three pre-trained language models, fastText [23], GloVe [24] and BERT [8], as the embedding matrix to help encode sentences.

**Sentence Representations.** We evaluate three sentence-level representations and a word-level representation. *AVG* represents an example as the mean of all its word embeddings. *IDF* represents an example as the weighted average of its word embeddings, in which weights are given by the inverse document frequency over all dataset. *CNN*, followed by [25], retrieves the representation by applying 1D convolution over the input words. Word-level representation(WORD) gives up the assembling operation and represents an example with a matrix, whose row is the embedding of the corresponding word.

**Distance metric**. Cosine similarity(*cos*) and Euclidean Distance(*l2*) are two commonly used distance metric.

**Implementation Details**. Baselines are the combination of aforementioned models. All parameters are optimized using Adam with a learning rate of 0.001. During meta-training, we sample 100 training episodes per epoch with an early stop strategy: when the validation loss fails to improve for over 10 epochs. Finally, we evaluate the model performance based on 1000 testing episode. The average accuracy is reported over 5 different random seeds and each seed is run over 5 times. Our code implementation is partly based on the work of [26].

## 3.3. Results and Analysis

We first evaluate our model in 5-way-1-shot and 5-way-5-shot intent detection tasks on **NLUE**. The experiment results are shown in Table 1. Compared to all baselines with different settings, our model achieves the best performance. For 1-shot setting, our model improves accuracy at most by 6.6% and 2.0% at most for 5-shot setting. The empirical results demonstrates that our model is indeed effective in few-shot intent detection task, and the word-level representation based distance metric is more consistent to human cognition. And the significant improvement in 1-shot experiment indicates our model is more competitive in extremely low resource settings.

Further more, we conduct the same experiments on **Huffpost** to explore the generalization of our approach. Our approach achieves at most 1.4% accuracy improvement in 5-way-5-shot and at most 4.2% in 5-way-1-shot. The results on **Huffpost** in Table 1 indicates our proposed method not only works for ID task, but also can be adapted to general TC tasks.

For both datasets, we evaluate different word embedding techniques, such as fastText, GloVe and BERT. According to Exp. *A* to *F*, the baselines shows a trend that models using fastText and BERT are usually more competitive to the ones using GloVe, while other settings are the same. However, it is opposite when checking in our model, which can be seen in Exp. *J*. Specifically, GloVe can bring our model at least 0.5% performance improvement in 5-shot setting on NLUE and at most 4.5% improvement in 1-shot on NLUE. We postulate that fastText and BERT are likely to containing more contextual infor-

Table 1: *Experiments on 5-way-1-shot and 5-way-5-shot classification. In column Rep., SENT means sentence-level representation, like [CLS] in BERT. A dash(-) means the setting is not available. Specially, the results with a dash in CNN row are limited to the data because NLUE contains one word sentences and CNN is not compatible.*

| Method | | | NLUE - 5 shot | | | NLUE - 1 shot | | | Huffpost - 5 shot | | | Huffpost - 1 shot | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Exp. ID | Rep. | Dist. | fastText | GloVe | BERT | fastText | GloVe | BERT | fastText | GloVe | BERT | fastText | GloVe | BERT |
| A | AVG | cos | 57.3 | 53.6 | 54.3 | 44.1 | 42.3 | 44.6 | 41.4 | 40.3 | 39.9 | 32.3 | 32.0 | 31.5 |
| B | AVG | l2 | 66.9 | 67.1 | 67.4 | 49.4 | 49.6 | 51.2 | 48.6 | 47.4 | 47.5 | 34.2 | 32.5 | 33.2 |
| C | CNN | cos | - | - | - | - | - | - | 38.3 | 38.3 | 37.3 | 30.7 | 30.4 | 30.1 |
| D | CNN | l2 | - | - | - | - | - | - | 41.4 | 42.0 | 46.0 | 32.0 | 32.1 | 33.3 |
| E | IDF | cos | 58.5 | 56.1 | 56.8 | 45.4 | 43.4 | 44.4 | 42.4 | 41.5 | 39.6 | 32.9 | 32.6 | 31.7 |
| F | IDF | l2 | 67.9 | 69.5 | 67.3 | 49.7 | 48.6 | 50.6 | 48.8 | 48.4 | 49.3 | 34.1 | 32.5 | 33.2 |
| G | SENT | cos | - | - | 50.1 | - | - | 41.4 | - | - | 40.4 | - | - | 31.6 |
| H | SENT | l2 | - | - | 66.4 | - | - | 48.9 | - | - | 49.6 | - | - | 33.6 |
| I | WORD | WRD | 49.9 | 52.0 | 50.6 | 35.8 | 38.4 | 38.6 | 35.3 | 34.8 | 39.1 | 26.4 | 27.0 | 30.3 |
| J | **WORD** | **EMD** | **69.9** | **71.4** | **67.8** | **55.2** | **56.2** | **51.7** | **49.3** | **49.8** | **50.0** | **35.9** | **36.8** | **35.8** |

mation than GloVe, and GloVe provides more word-level information. So, GloVe is more consistent to our approach which utilize word information to make sample comparison.

We also study two different distance metric, cosine similarity and Euclidean distance, which are commonly used. Uniformly, models using Euclidean distance performs better and at least 1.3% accuracy improvement is achieved in 5-way-1-shot on Huffpost. Mathematically, cosine similarity and Euclidean distance can be given by (8) and (9):

$$Cos(\mathbf{u}, \mathbf{v}) = 1 - cos(\mathbf{u}, \mathbf{v}) = 1 - \frac{\mathbf{u}^T \mathbf{v}}{\lambda_\mathbf{u} \lambda_\mathbf{v}} \tag{8}$$

$$\begin{aligned} Eul(\mathbf{u}, \mathbf{v}) &= ||\mathbf{u} - \mathbf{v}||_2 = ||\lambda_\mathbf{u} \widehat{\mathbf{u}} - \lambda_\mathbf{v} \widehat{\mathbf{v}}||_2 \\ &= \sqrt{\lambda_\mathbf{u} \lambda_\mathbf{v} (2Cos(\mathbf{u}, \mathbf{v}) + (\lambda_\mathbf{u} - \lambda_\mathbf{v})^2)} \end{aligned} \tag{9}$$

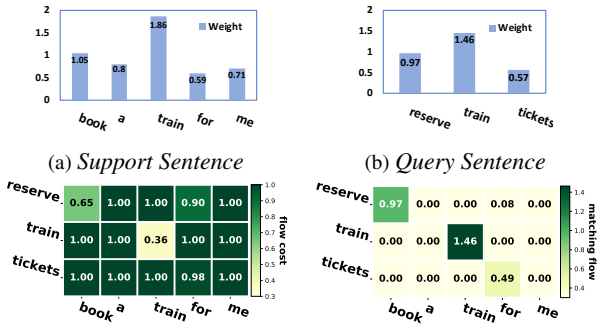
where $\lambda_\mathbf{u}$ and $\lambda_\mathbf{v}$ denote the norm of the vector and, $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$ is the direction vector. From (9), we can see that cosine similarity can be obtained in Euclidean distance in some way. So, we surmise that Euclidean distance is usually better than cosine similarity because it gives both the norm and angle information.

### 3.4. Ablation Study

We conduct ablation study on our proposed method. The results are shown with the Exp. *I* and *J* in Table 1 and our approach outperforms in all settings. *WRD* distance metric is given by [27] to improve the research in semantic textual similarity task. *WRD* also uses EMD algorithm, while the weighting factor and the transportation cost are given by the norm and angle of local word feature vector. We re-implement *WRD* following its paper. The difference between *WRD* and our distance metric is the input parameters for EMD, which is the weight of words. The experiment result shows that the semantic weight generated by our Mutual-Semantic mechanism is interpretative and consistent to the explanation in section 2.4.

### 3.5. Visualization

We visualize the weights and optimal flow to illustrate how our approach gives a good matching similarity. Figure 3a and Figure 3b show an example sentence pair with same label *transport ticket* and their word semantic weights generated by our approach. The semantic overlap between words is given by Figure 3c, and deeper color means less semantic overlap degree. Figure 3d represents the number of weights transported to corresponding position to construct the semantics of support sentence, with




(a) *Support Sentence* (b) *Query Sentence*

(c) *Matching Flow Cost* (d) *Matching Flow*

Figure 3: *Weights and EMD result generated by our approach.*

deeper color giving higher flow. As we can see, our approach gives more attention to the label related key words, e.g. book, train and reserve, and the semantic transportation between them given by EMD is consistent to our recognition.

## 4. Conclusion

We propose a new few-shot intent detection method, Semantic Transportation Prototypical Networks (STPN), which utilizes word-level representation of instance, and matches key word semantic of sentences. Also, we adopt EMD as a distance metric, providing an innovation of matching similarity measurement. Our proposed Mutual-Semantic mechanism, generating word semantic weights for EMD, enriches key word semantic information and reduces the noise signal from unrelated word. The comprehensive experiments indicate the effectiveness and generalization of our approach.

## 5. Acknowledgements

This paper was partially supported by Shenzhen Science & Technology Fundamental Research Programs (No: JSGG20191129105421211 & GXWD20201231165807007-20200814115301001).

# 6. References


[1] G. Tür, L. Deng, D. Hakkani-Tür, and X. He, "Towards deeper understanding: Deep convex networks for semantic utterance classification," in *ICASSP*. IEEE, 2012, pp. 5045–5048.

[2] P. Xu and R. Sarikaya, "Convolutional neural network based triangular CRF for joint intent detection and slot filling," in *ASRU*. IEEE, 2013, pp. 78–83.

[3] C. Zhang, W. Fan, N. Du, and P. S. Yu, "Mining user intentions from medical queries: A neural network based heterogeneous jointly modeling approach," in *WWW*. ACM, 2016, pp. 1373–1384.

[4] S. V. Ravuri and A. Stolcke, "Recurrent neural network and LSTM models for lexical utterance classification," in *INTERSPEECH*. ISCA, 2015, pp. 135–139.

[5] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *INTERSPEECH*. ISCA, 2016, pp. 685–689.

[6] Y. Wang, Y. Shen, and H. Jin, "A bi-model based RNN semantic frame parsing model for intent detection and slot filling," in *NAACL-HLT (2)*. Association for Computational Linguistics, 2018, pp. 309–314.

[7] J. Hu, G. Wang, F. H. Lochovsky, J. Sun, and Z. Chen, "Understanding user's query intent with wikipedia," in *WWW*. ACM, 2009, pp. 471–480.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT (1)*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *NeurIPS*, 2020.

[10] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *NAACL-HLT*. Association for Computational Linguistics, 2018, pp. 2227–2237.

[11] X. Liu, A. Eshghi, P. Swietojanski, and V. Rieser, "Benchmarking natural language understanding services for building conversational agents," *CoRR*, vol. abs/1903.05566, 2019.

[12] A. Blum and T. M. Mitchell, "Combining labeled and unlabeled data with co-training," in *COLT*. ACM, 1998, pp. 92–100.

[13] Z. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 11, pp. 1529–1541, 2005.

[14] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *EMNLP*. Association for Computational Linguistics, 2018, pp. 489–500.

[15] J. W. Wei and K. Zou, "EDA: easy data augmentation techniques for boosting performance on text classification tasks," in *EMNLP/IJCNLP (1)*. Association for Computational Linguistics, 2019, pp. 6381–6387.

[16] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *NIPS*, 2016, pp. 3630–3638.

[17] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *NIPS*, 2017, pp. 4077–4087.

[18] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*. IEEE Computer Society, 2018, pp. 1199–1208.

[19] T. Dopierre, C. Gravier, and W. Logerais, "A neural few-shot text classification reality check," *arXiv preprint arXiv:2101.12073*, 2021.

[20] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *CVPR*. IEEE, 2020, pp. 12 200–12 210.

[21] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[22] R. Misra and J. Grover, *Sculpting Data for ML: The first act of Machine Learning*, 01 2021.

[23] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext.zip: Compressing text classification models," *CoRR*, vol. abs/1612.03651, 2016.

[24] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*. ACL, 2014, pp. 1532–1543.

[25] Y. Kim, "Convolutional neural networks for sentence classification," in *EMNLP*. ACL, 2014, pp. 1746–1751.

[26] Y. Bao, M. Wu, S. Chang, and R. Barzilay, "Few-shot text classification with distributional signatures," in *ICLR*. OpenReview.net, 2020.

[27] S. Yokoi, R. Takahashi, R. Akama, J. Suzuki, and K. Inui, "Word rotator's distance," in *EMNLP (1)*. Association for Computational Linguistics, 2020, pp. 2944–2960.