Few-shot News Recommendation via Cross-lingual Transfer

ABSTRACT

The cold-start problem has been commonly recognized in recommendation systems and studied by following a general idea to leverage the abundant interaction records of warm users to infer the preference of cold users. However, the performance of these solutions is limited by the amount of records available from warm users to use. Thus, building a recommendation system based on few interaction records from a few users still remains a challenging problem for unpopular or early-stage recommendation platforms. This paper focuses on solving the *few-shot recommendation* problem for news recommendation based on two observations. First, news at different platforms (even in different languages) may share similar topics. Second, the user preference over these topics is transferable across different platforms. Therefore, we propose to solve the *few-shot news recommendation* problem by transferring the user-news preference from a many-shot source domain to a few-shot target domain. To bridge two domains that are even in different languages and without any overlapping users and news, we propose a novel unsupervised cross-lingual transfer model as the news encoder that aligns semantically similar news in two domains. A user encoder is constructed on top of the aligned news encoding and transfers the user preference from the source to target domain. Experimental results on two real-world news recommendation datasets show the superior performance of our proposed method on addressing few-shot news recommendation, comparing to the baselines.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

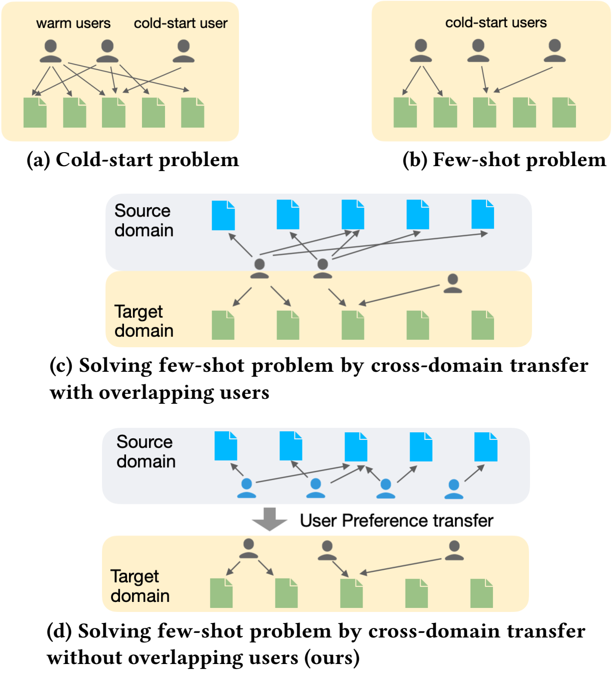
KEYWORDS

News recommendation; Cross domain recommendation; Transfer learning

ACM Reference Format:
. 2018. Few-shot News Recommendation via Cross-lingual Transfer. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

News recommendation aims to help users find news that they are interested in over the massive options [8, 21, 30]. Such personalized recommender systems boost users' reading experience and business revenue of news platforms [31, 33]. It has thus garnered increasing attention and has been tackled by a number of methods [28–30, 34]. The general idea of these methods is to characterize the preference of a user based on the news that the user has previously read and then infer the interest of the user on other news. The quality of recommendation results thus highly depend on the amount of historical user-news interactions to use. Like all other recommendation problems, news recommendation also suffers from the cold-start problem when the targeted user only has interactions with few news before, as illustrated in Fig. 1(a).



(a) Cold-start problem

(b) Few-shot problem

(c) Solving few-shot problem by cross-domain transfer with overlapping users

(d) Solving few-shot problem by cross-domain transfer without overlapping users (ours)

Figure 1: Illustration of problem formulations and solutions. (a) Cold-start problem: predict the preference of cold-start users while having many warm users in the system. (b) Few-shot problem: predict the preference of cold-start users while no warm users are available. (c)/(d) Solving few-shot problem by cross-domain transfer with/without overlapping users between two domains.

A number of works have tried to infer the preference of cold-start users based on the behavior of warm users who have browsed a large amount of news [10, 14, 18, 37]. Though promising results have been obtained by making use of warm users to help cold users in recommendation, these solutions become incapable when no warm users are available to use. This is the so-called *few-shot user recommendation* scenario, where all users have few user-item interactions, as shown in Fig. 1(b). This scenario can often be found at unpopular or early-stage recommendation platforms. Addressing the few-shot recommendation at these platforms is actually a chicken-and-egg problem. The performance of the recommender system at these platforms relies on the observable user-item interactions and in turn, the user-item interactions rely on the performance of the recommender system as users may not find their interested items and leave the platform [4]. To address this issue in

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

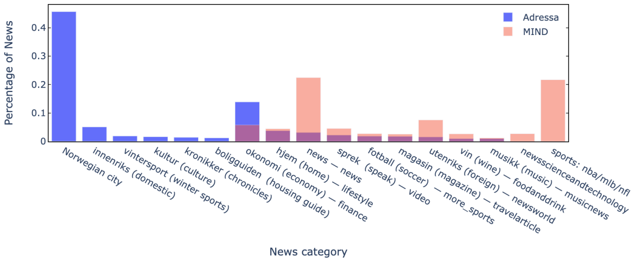


Figure 2: Overlapping news category distribution between Adressa (in Norwegian) and MIND (in English) dataset.

news recommendation, an approach in [27] is proposed to use the browse information of these cold-start users themselves in other domains. This is a feasible idea to transfer the knowledge from a rich source domain to the few-shot target domain. However, these methods require the existence of overlapping users between two domains [7, 20, 27] (as shown in Fig. 1(c)), and are thus limited to only transfer knowledge via the overlapping users.

We focus on a more challenging but widely existing problem that users in a target domain are all cold-start users, and they are not available in other domains for playing the role of knowledge transfer. As shown in Fig. 1(d), these users have only a few interaction records in one single domain. In this case, the previous meta-learning based [13, 15, 26] and transfer learning based solutions [7, 20, 27] become incapable, due to the limited amount of user-news interactions. Even a strong domain-specific news encoder cannot always help in this case, because the inference of users' interest is limited more by the lack of user-news interactions, rather than the understanding of news content. Take zero-shot cold users as special examples, a news encoder doesn't help on guessing the users' interest. Recommendation decision is then likely made by blind guess, unless external information from another domain can help on profiling the common patterns of users' interest.

We are therefore motivated to build a few-shot cross-domain news recommendation system, which works without the requirement of overlapping users to bridge the two domains. Picturing our application on two widely used public news recommendation datasets: MIND (English news published around 2020) and Adressa (Norwegian news published around 2017), we have two observations that support our investigation of the cross-domain (even cross-lingual) transfer. First, news in different domains may share similar categories and have somewhat similar topics. Figure 2 shows that the news categories in MIND and Adressa dataset have some overlapping topics, such as general *news*, *financial news*, *home/lifestyle* related news and so on. Second, although users in two domains are different, some general reading preference over news topics is transferable from the domain of one language to anther.

Based on these observations, we propose to build a *few-shot cross-lingual news recommendation system*. Since there are no overlapping users or news to bridge the two domains, transferring the user preference patterns from one domain to another has to address the *domain shift issue*, which can be attributed to the difference of news category distribution between two domains. As shown in Figure 2, the news category distributions of two datasets are different, although there are overlapping topics. The news about *Norwegian city* is the majority in Adressa, while *sport news* and general *news* are the two major topics in MIND. Since the preference of users should be inferred over the news distribution in their own domain, it is therefore essential to minimize the gap of user characterization between the source and target domain. When the target domain is cold for all users, it is a new challenging problem for addressing the domain shift issue by only a few user-news interactions. The previous study of language shift in NLP tasks like name entity recognition [9, 36] developed methods that align different language text with similar topic distribution and are not suitable for solving the gap of user characterization between domains in recommendation. Multilingual pretrained language models can be employed to represent news of different languages in the same space. However, the representation only takes into account the news content, and cannot catch how they are liked by users. In news recommendation models, news should be represented to reflect not only its content but also its attractiveness to users.

We design a cross-domain recommendation system consisting of news and user encoder shared between the source and target domain and propose three strategies for alleviating the domain shift, *Cross-domain Extension*, *Random Masking*, and *News-Alignment*. The *Cross-domain Extension* strategy is to construct augmented news which are semantically similar across two domains. The *Random Masking* strategy works by randomly taking the original news or one of the augmented news during training. In this way, the training of the shared news encoder is exposed to both the source and target domain. The inference of user preference later based on the news encoder thus covers the topics in target domain as well. The *News-Alignment* strategy is designed to drive similar news from two domains to be close in the same representation space. It presents all augmented news in pairs, and sends them to train the shared news encoder, such that the paired news have similar representations. In this common news representation space, we then get the user representations by the shared user encoder based on users' browsing history. The recommendation problem in target domain is thus solved with the transferred knowledge from the source domain.

We summarize the contributions of this work as follow: (1) To the best of our knowledge, we are the first one to use the cross lingual knowledge to solve the few-shot news recommendation problem in a challenging setting of no common users or news existing between the source and target domain. (2) We propose to use a shared recommender model across two domains to transfer the user-news preference patterns. For combating with the domain shift issue, we design three strategies to align the source and target domain in the same representation space. (3) We demonstrate the effectiveness of our proposed method with thorough experiments on two real-world datasets. The results show that our method has a consistent improvement above the baselines.

2 RELATED WORK

Neural News Recommendation: On the recent benchmark news recommendation dataset MIND [34], various deep learning-based news recommendation architectures have been developped, such as NPA [28], NAML [29], and NRMS [30]. Pre-trained language models [33] and even multilingual pre-trained models [32] have been used as advanced news encoder for improving the recommendation performance. Besides, cold-start issues have been also investigated in news recommendation by inferring the preference of cold users based on their activities in other domains [7, 27] or based on activities of other warm users in the same domain [10, 14, 18, 37].

Cold-start and Few-shot User Recommendation: In the general recommendation field, including news recommendation, cold-start problems for user refer to the situation where little is known about the preferences of the new user [10]. Most solutions [7, 10, 14, 18, 27, 37] aim at leveraging the behaviors of warm users to infer the preference of cold-start users. However, these methods become incapable in the scenario where no warm users are available and all users have few interactions with items. For this *few-shot user recommendation* problem, a popular solution is to apply meta-learning on making recommender models adaptable to users who have few interactions [13, 15, 26]. However, for many unpopular or early-stage recommendation platforms, these meta-learning based solutions become incapable due to the lack of training data for constructing meta-learning tasks. In [4], zero-shot recommenders are proposed to leverage the knowledge of a source dataset to improve the recommendation performance of a target domain without using any target domain data during training. Our proposed solution is flexible for either zero-shot or few-shot setting, depending on if any training data are available or not in the target domain.

Cross Domain Recommendation: Previous cross domain recommendation models [7, 20] are mainly designed for the cold-start problems, i.e., the user or item is new. For bridging the two different domains, they assume the existence of overlapping users or news between two domains. Our study is in a realist setting that has no requirement on the existence of overlapping users or news. The study of domain adaptation recommendation [35] also doesn't require overlapping users or items. However, the adaptation methods require extensive training data in both domains and thus are not applicable for the few-shot scenario where the training samples in the target domain are scarce.

Cross-lingual Transfer and Multilingual Pretrained Language Models: Cross-lingual techniques have promoted various NLP tasks. One popular idea is to translate the source language text to the target language. Based on the translated source text, the translation-then-align method [12, 16, 36], Bilingual method [16, 36] and many other task-specific cross-lingual methods [9, 36] have been designed to reduce the language shift between two domains. These aforementioned works are designed to transfer the text-level knowledge between languages which are in the same topics for text-level tasks, e.g., the user reviews sentiment analysis about similar topics in different languages in SemEval-2016 task [22]. For the few-shot news recommendation task, each user is represented by a sequence of reading history news and the user preference is inferred based on the sequence of news. Because of the news topics deviation, the target domain reading history has a different topic distribution from the source domain reading history. The inference of user preference in a cross-lingual model thus should address the domain shift issue. Multilingual pretrained language models can be an effective technique to find news with similar content in different languages. They can be employed as a news encoder for representing news from different domains in the same space. However, the news will be encoded solely based on the content, without considering the relevance of news in users' reading history. Two pieces of news with different content and topics can have similar representation if users who read one also prefer to read the other. Hence, the news encoder in a recommendation model should represent news by taking into account both its content and its history of being read by users.

3 THE PROPOSED FEW-SHOT NEWS RECOMMENDATION MODEL

3.1 Notation and Problem Definition

We let the source domain be noted as s and the target domain be noted as t. The U^s, U^t and D^s, D^t represent the user and news set of source domain and target domain, respectively. The user-news interactions in source and target domain available for training are noted as τ^s and τ^t, respectively. Each user $u \in U$ has a reading history news sequence $[d_1, d_2, ..., d_{len(u)}]$, where len measures the length of a sequence or the size of a set. Each news $d \in D$ is represented by a token sequence $[w_1, w_2, ..., w_{len(d)}]$. For a given user $u \in U$ and a candidate news set $C = \{d_i\}, i = 1...|C|$, the recommendation task is to predict the preference score of this user on each candidate news: r_i, $i = 1...|C|$. In the few-shot recommendation problem, we have only the ground-truth label y_i for a few news that the user u likes (y_i=1) or not (y_i=0). We build a cross-lingual news recommendation system, which predicts the preference of user $u \in U^t$ in the target domain by leveraging the abundant user-news interactions τ^s in the source domain and the few-shot or zero-shot interactions τ^t in the target domain.

3.2 Overall Framework

The framework of our proposed method is shown in Figure 3. The shared news encoder and user encoder are jointly trained by the source domain training set τ^s and the target domain training set τ^t. Note that τ^t is much smaller than τ^s, and there is no overlapping between τ^s and τ^t. After training, the news and user encoder are used in the target domain for predicting r^t_{ud}, the preference of user u on news d in target domain. There is a module of Cross-domain Extension (CDE) in training. For a news in the source domain $d \in D^s$, it goes through the CDE module and gets d' (the d's most similar news from the target domain) and d'' (a translation of d in the target language). Such extension is designed to alleviate the language shift and content shift for training the shared news encoder after Random Masking (RM). The RM operation randomly masks two news among (d, d', d'') and sends the remaining one to news encoder for participating the training. In the iterative training process, this RM operation blends the augmented news in the original news and makes the shared news encoder exposed to both the source and target domain. the source and target domain be better aligned. The news encoding vectors are then sent to user encoder to produce user encoding vectors. The preference prediction r^s_{ud} and r^t_{ud} are calculated by running dot product on the encoding vectors of candidate news and user. One additional News-Alignment module is applied to the source domain news and the augmented news for driving the source and target domain into the same representation space, so the domain shift can be further alleviated.

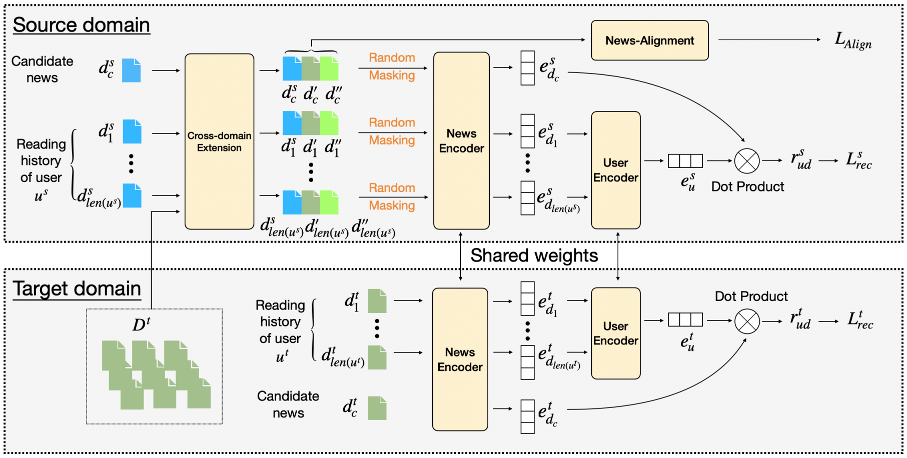


Figure 3: Framework of our proposed method.

3.3 Base Recommendation Network

We first describe the base neural recommendation model for obtaining users representations, news representations and constructing the recommendation loss to train the neural network model. We choose the state-of-the-art model NRMS [30] as our base network. The source domain and target domain share the same base network to encode news and users, and then conduct recommendations.

News Encoder ϕ: Our news encoder is shared by the source and target domain. A news d from either source or target domain is first tokenized by bpe [6] and then sent to a multilingual shared dictionary to obtain the representations of each token $[e_{w_1}, e_{w_2}, \ldots, e_{w_{len(d)}}]$. Then the token representations are given to a multi-head self-attention and attention network to produce the news representation e_d,

$$e_d = \phi(d) = Attention(MultiHeadAttn([e_{w_1}, e_{w_2}, \ldots, e_{w_{len(d)}}])). \quad (1)$$

User Encoder φ: The user representation is built from the reading history $[d_1, d_2, ..., d_{len(u)}]$ of this user. Thus, the news representation $\phi(d_i)$ in the history is sent to an attention network to produce,

$$e_u = \varphi(u) = Attention([\phi(d_1), \phi(d_2), \ldots, \phi(d_{len(u)})]). \quad (2)$$

Training and Inference: Given the representation of a user and a candidate news, the preference value can be computed by dot product, $r_{ud} = e_u^T e_d$. In the training stage, we have a positive news d liked by a user u, and sample four random negative news that the user may not like. We thus have $[d^+, d_1^-, d_2^-, d_3^-, d_4^-]$ and its corresponding prediction score $[r_{ud}^+, r_{u1}^-, r_{u2}^-, r_{u3}^-, r_{u4}^-]$. The recommendation loss can be defined by soft-max on the predicted scores. The overall loss function defined on training data is

$$\mathcal{L}_{rec} = -\sum_{(u,C)\in\tau}\ \sum_{i\in C_{positive}} \log \frac{\exp\left(r_{ui}^+\right)}{\exp\left(r_{ui}^+\right) + \sum_{j=1}^4 \exp\left(r_{uj}^-\right)}, \quad (3)$$

where $positive(C)$ is the positive news among the candidate news.

3.4 Cross-domain Extension

The key of designing a cross-lingual news recommendation system is to address the *domain shift issue*, which can be attributed to the difference of news distribution between two domains. Although multilingual pre-trained language models can represent text of different languages in the same embedding space and alleviate the language shift to certain extend, they are mostly trained only with token-level objectives to align tokens in different languages [1, 17]. If employing them as news encoder, news will be encoded solely based on the content, without considering the relevance of news in terms of users' preference. In the task of news recommendation, news representation should be learned based on not only its content but also its attractiveness to users.

To address this domain shift issue, we introduce the news of the target domain into the source domain during training and simulate the interactions of source users in the target domain and also the interaction of target users in the source domain. This so-called *Cross-domain Extension* (CDE) module is our first strategy to alleviate the domain shift via augmenting news across two domains. Specifically, there are two ways of augmentation. First, we use a multilingual pre-trained language model to find the most similar news $d' \in D^t$ in the target domain for each news $d \in D^s$ in the source domain:

$$d' = FindMostSimilar(d, D^t).$$

Concretely, we first use Sentence Transformers [24] and Multilingual DistilBertModel [25] to obtain the embeddings for each news in the source domain and target domain. Based on the embeddings of news, we use Hnswlib [19] which is a fast approximate nearest neighbor search tool to find the most similar news $d' \in D^t$ for each $d \in D^s$. To exclude unrelated news in nearest neighbor search, we set a similarity threshold ranging from 0 to 1 to control the selection process. For each $d \in D^s$, if its similarity score to the nearest neighbor is still lower than the threshold, it fails to find a similar news in the target domain and will have no d' in augmentation. We explore the influence of the similarity threshold in Section 4.

To weaken the language shift between two domains, we translate each news in the source domain to the language in target domain. Machine translation has been regarded as an effective technique to rewrite input text to a semantic similar translated text whose grammar and syntax are more similar to the target domain. Formally, for each news in the source domain $d \in D^s$, we use the Google translate[1] to translate it into the target language

$$d'' = Translate(d).$$

News d and d'' will be the two forms of expression about the same news content in different languages.

The *Cross-domain Extension* is operated before training and can be regarded as a pre-processing method and run offline. The augmented news set of d, denoted as $A_d = \{d, d', d''\}$, will be used in the next *Random Masking* module for simulating the interactions of users with news across different domains.

3.5 Random Masking, News and User Encoding

After introducing the news of the target domain into training data, we design *Random Masking*, the second strategy for reducing the domain shift. The goal is to blend the augmented news in the original news for training the shared news encoder by making it exposed to both the source and target domain. Concretely, the random masking operation is applied to $A_d = \{d, d', d''\}$ by randomly masking two of them and sending the remaining one to the news encoder (for news d that doesn't have a similar target news, randomly mask one of $\{d, d''\}$ and send the remaining one to the news encoder). The idea is analogous to code-switch used in cross-lingual NLP tasks [23, 36], which randomly replaces words in a sentence with the corresponding translation words in a different language to align words in different languages by mixing their context information. We propose a similar operation in the recommendation task for simulating the interactions of users with news across different domains.

Imagine that one training sample in the source domain includes a user $u \in U^s$ with reading history $[d^s_1, d^s_2, ..., d^s_{len(u)}]$, a candidate news set C and the label for each candidate news which indicates the user likes or dislike the news. For each news in the training sample (including both the history and candidate set), we have its augmented set $A_{d_i} = \{d_i, d'_i, d''_i\}$. In the iterative training process, the random masking operation will dynamically create enriched reading history and candidate news set for user u, which is a mixture of news in source and target domain. Let Γ be the random masking function. We then get the news representation for d_i:

$$e_{d_i} = \phi(\Gamma(A_{d_i})),$$

and the user representation for u:

$$e_u = \varphi(u) = Attention([\phi(\Gamma(A_{d_1})), \phi(\Gamma(A_{d_2})), \ldots, \phi(\Gamma(A_{d_{len(u)}}))]).$$

In this way, users in the source domain are virtually interacting with news in both source and target domain. This random masking operation is also beneficial for users in the target domain. If a user $u^t \in U^t$ interacted with news $d' \in D^t$ in history, the virtual interaction between $u^s \in U^s$ with this $d' \in D^t$ will pull u^t and u^s closer. Meanwhile, the news $d \in D^s$ and its translated d'' will be pulled close to u^t as well. The target user u^t is thus virtually interacting with news in the two domains. The shared news encoder ϕ and user encoder φ trained by these augmented user-news interactions reduce the deviation of news distribution between the source and target domain, and therefore have promoted generalization capacity when recommending news to cold users in the target domain. Note that there is no distortion of user preference in the random masking, because there is no deletion of the original user-news interactions, and d'_i and d''_i are derived from d_i to have similar content.

3.6 News-Alignment

To ensure that the news with similar content in one augmented news set $A_d = \{d, d', d''\}$ are embedded as representation vectors with high similarities, we design a news-alignment module, which explicitly pulls $\{d, d', d''\}$ close in the representation space. For each news $d \in D^s$, we firstly get its augmented news set $A_d = \{d, d', d''\}$, and then we use the news encoder to get representation of the source domain news $e_d = \phi(d)$, the translated news $e'_d = \phi(d')$, and the target domain news $e''_d = \phi(d'')$. The mean squared error loss (MSE) below is minimized to align $\{d, d', d''\}$ closer in the representation space,

$$L_{Align} = \sum_{d \in D^s} MSE(e_d, e'_d) + MSE(e_d, e''_d). \tag{4}$$

In this way, the news commonly interacted by users are further pulled closer, no matter the interactions are original or augmented, are in source or target domain. Even when the translated news d'' has an issue of translation quality, the alignment module helps to adjust the news encoder ϕ so the embedding of d'' does not deviate from that of d.

The overall loss function for the cross-domain new recommendation model consists of the recommendation loss in the source and target domain, as well as the alignment loss:

$$L = \alpha L_{Align} + \beta L^s_{rec} + L^t_{rec}, \tag{5}$$

where α and β are the trade off parameters. Our model is flexibly applicable to few-shot and zero-shot setting. In a zero-shot setting, there will be no target domain recommendation loss.

4 EXPERIMENT EVALUATION

4.1 Experiment settings

Datasets. There are five public news recommendation datasets including Plista (German) [11], Adressa (Norwegian)[2] [5], Globo (Portuguese) [2], Yahoo!(English)[3] and MIND (English)[4] [34]. The Globo and Yahoo! datasets only have word embeddings or word IDs for the news content. Since the original text of news is not available, we cannot get the multilingual embedding of news in these datasets, and thus cannot use them in the evaluation. The Plista dataset is not yet public available until the submission of this work. Hence, there are only two public news recommendation datasets available for our evaluation. We use MIND (English) and Adressa (Norwegian) to conduct our experiments. MIND is a news recommendation dataset released by Microsoft in 2020. It includes massive user-news interactions in MSN news website. Adressa is a news recommendation dataset released by NTNU in 2017. It includes Norwegian news and user-news interactions in Norwegian on the platform Adresseavisen which is a news website in Trondheim, Norway.

[1] https://translate.google.com.

[2] https://reclab.idi.ntnu.no/dataset with CC BY-NC-SA 4.0 license.

[3] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l

[4] https://msnews.github.io with Microsoft Research License Terms.

Few-shot (zero-shot) transfer setting. We construct two few-shot transfer settings: Adressa -> MIND and MIND -> Adressa. In both transfer settings, the training samples in the target domain are prepared in the few-shot scenario by randomly selecting 200 users, each of which only has two news interactions (2-shot), or four news interactions (4-shot). The zero-shot scenario has no training samples in the target domain. To imitate the real-world few-shot scenario, the news in selected trainig samples were published before the date of test set. More details about the training and test dataset can be found in Table 1. In MIND -> Adressa, we randomly selected 10,000 users and their news interactions from the training set of MINDlarge as the source domain training set. We randomly selected 40% of all users and their interactions in Adressa on Jan. 7, 2017 as the validation set and the other 60% as the test set in the target domain. In Adressa -> MIND, we randomly selected 10,000 Adressa users who click news on January 6, 2017 to construct the training set in source domain, and include also the reading history of these users collected from January 1, 2017 to January 5, 2017. For the target domain dataset, we randomly selected 40% of users and their interactions in dev set of MINDlarge as the validation set and the other 60% as the test set.

Implementation details and evaluation metrics. We implement our model by Pytorch and conduct all the experiments on a Linux server with GPUs (Nvidia RTX 3090). We use the Adam optimizer for training. The training epochs are set to 20, and the embedding dimension is set to 300, and the batch size is set to 80. We use AUC as the main evaluation metric. The performance measured by MRR, NDCG@5, NDCG@10 is also reported in the Appendix. In the MIND -> Adressa experiment, the learning rate is set to 3e-4. In the Adressa -> MIND experiment, the learning rate is set to 1e-4. For both settings, the hyper-parameter α is set to 1 and the β is set to 0.2. Each experiment is repeated for 5 times and the mean and the standard deviation are reported. We will release all codes after paper publication.

4.2 Baselines

There is no method specifically designed for cross-lingual few-shot news recommendation. We adopt common cross-domain and domain-adaption recommendation methods as baselines:

NRMS (Only Target data) [30]: This is our base model. In zero-shot setting, we evaluate it on the target domain test set directly. In few-shot setting, we train the model on the target domain training set and then evaluate it on the target domain test set.

ZERO-SHOT (Source + Target data) [36]: We adapt the framework of cross-lingual models designed for sentiment analysis in [36] as baselines for our recommendation tasks. In zero-shot setting, a recommendation model is trained on the source data and tested on the target domain test data. In the few-shot setting, the target domain training set is included in training as well. Testing is always on the target domain test set.

Translate-then-align (Translation + Target data) [36]: Following the Translate-then-align method in [36], news samples in the source language are translated into the target domain language by machine translation. Then the translated dataset are used for training in the zero-shot setting. In the few-shot setting, the target domain training set is also included for training.

Bilingual (Source + Translation + Target data) [36]: As a variant of **Translation + Target data**, this baseline uses both the translated dataset and the original dataset of source domain for training in the zero-shot setting. In the few-shot setting, the target domain training set is additionally included for training.

TDAR [35]: This is a text-enhanced domain adaptation recommendation model. It requires training data in both source and target domain. We thus only have it for comparison in the few-shot setting.

4.3 Overall recommendation performance in zero-shot and few-shot setting

The overall recommendation performance measured by AUC is reported in Table 2. The results of other metrics are presented in Table 4 in Appendix A.1. From the results shown in Table 2 and 4, we have the following observations:

1) Our method outperforms all baselines in all settings. In the results of zero-shot setting in Table 2, our method achieves 4.69% AUC improvement over the best baseline in MIND -> Adressa and 3.63% in Adressa -> MIND. In the 2-shot and 4-shot setting, the AUC improvements over the best baseline are 3.9% and 7.69% in MIND -> Adressa, 0.75% and 2.31% in Adressa -> MIND, respectively.

2) The inferior performance of ZERO-SHOT (Source + Target) indicates that fine-tuning model on the source language and then inferring on the target language is not an effective method for knowledge transfer because the transfer is restricted to the implicit domain alignment from the pre-training process as we discussed in the Related Work section. The Translate-then-align and Bilingual methods are worse than our methods, due to the language shift and content shift issue we mentioned in the Introduction section.

3) TDAR has low performance in the low-resource cross lingual news recommendation scenario. This is mainly because TDAR works better when topics overlap more between two domains. The evaluation data in the two domains were published in different periods and thus have content shift. Therefore, it is difficult for TDAR to align the content distribution between two domains.

To demonstrate the capability of our model on alleviating the domain shift when characterizing user preference, we visualize the users embedding of two domains in Figure 4. Comparing to the baseline ZERO-SHOT (Source + Target), our method pulls close the user embedding vectors of two domains. There also exist users in one domain but embedded close to users in the other domain, i.e., the red (blue) points in the region of blue (red) points. Our method is thus verified to be able to transfer user preference patterns, even without requiring bridging users/news commonly exist in two different domains.

It is worth to discuss interesting findings from the experiment results in two different transfer settings (MIND -> Adressa vs Adressa -> MIND): the improvements in the MIND -> Adressa setting are all better than the improvements in the Adressa -> MIND setting. This can be explained by the property of source/target domain. MIND contains a large amount of English news covering a wide range of topics, while Adressa only contains a small number of Norweigian news and limited topics in Norway. Thus, in the MIND -> Adressa setting, the source domain contains more user preference knowledge of the target domain, comparing to the case in the Adressa -> MIND transfer setting. So the improvements are more significant.

Table 1: Dataset statistics in the two few-shot (and zero-shot) transfer settings (Note that the test set in target domain is larger than the training set in the few-shot news recommendation scenario)

Setting			MIND -> Adressa			Adressa -> MIND		
			0-shot	2-shot	4-shot	0-shot	2-shot	4-shot
Training Set	Target Domain	#users	0	200	200	0	200	200
		#interactions	0	400	800	0	400	800
	Source Domain	#users	10000			10000		
		#news	31099			2097		
		#interactions	1083619			22700		
Test set		#users	2194	816	837	1000	864	843
		#interactions	4674	1734	1777	1627	1380	1278

Table 2: Zero-shot and few-shot recommendation performance comparison in terms of AUC. The best result is in bold font, and the best baseline is underlined. The unsolvable zero-shot case is indicated by –.

Cross-domain Setting		MIND -> Adressa			Adressa -> MIND		
training data available in target domain		0-shot	2-shot	4-shot	0-shot	2-shot	4-shot
NRMS (Only Target) [30]		0.482±0.012	0.564±0.006	0.542±0.004	0.506±0.012	0.507±0.014	0.490±0.016
ZERO-SHOT (Source + Target) [36]		0.512±0.014	0.549±0.018	0.529±0.010	0.522±0.014	0.528±0.021	0.520±0.014
Trans.-Align (Trans. + Target) [36]		0.504±0.016	0.551±0.009	0.538±0.015	0.523±0.004	0.533±0.012	0.509±0.015
Bilingual (Source+Trans.+Target) [36]		0.505±0.014	0.539±0.017	0.546±0.014	0.518±0.014	0.528±0.005	0.513±0.012
TDAR [35]		–	0.509±0.010	0.507±0.013	–	0.498±0.012	0.515±0.014
Ours	Random Masking	0.521±0.011	0.570±0.016	0.574±0.009	**0.542±0.005**	0.535±0.004	**0.532±0.011**
	Random Masking+News-Align	**0.536±0.014**	**0.586±0.017**	**0.588±0.020**	0.540±0.005	**0.537±0.008**	0.531±0.006

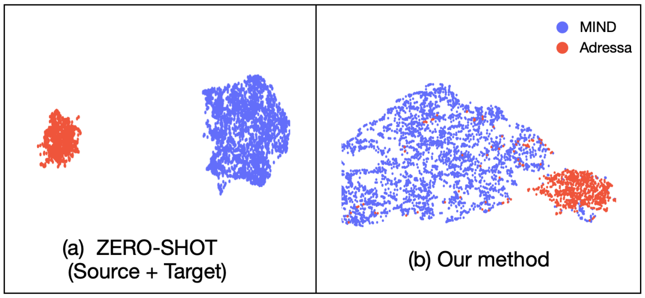


Figure 4: User embeddings visualization of ZERO-SHOT (Source + Target) method and our method in the MIND -> Adressa 4-shot setting.

In real-world scenarios, the MIND -> Adressa transfer setting is also more realistic than the Adressa -> MIND setting, because most news websites are English. The abundant user-news interaction records from English websites can be used to solve the few-shot problem in other low-resource news recommendation platforms.

4.4 Ablation studies

To further investigate the effect of three strategies designed for reducing domain shift, we conduct ablation studies in the MIND -> Adressa 4-shot setting. The results are shown in Table 3.

The use of target domain news $d' \in D^t$ in Cross-domain Extension. From Table 3, we can observe that all models with target domain news d' in Cross-domain Extension have better performance than those without it, indicating the selected target domain news in Cross-domain Extension can be regarded as an effective bridge to help transfer useful knowledge from the source domain to the target domain.

Table 3: The impact of Cross-domain Extension, Random Masking and News-Alignment, evaluated in the MIND -> Adressa 4-shot setting.

Method		AUC
Trans.-Align (Trans. + Target) [36] (Best Baseline)		0.546±0.014
without target news d'	Random Masking	0.558±0.011
	News-Alignment	0.559±0.009
	Random Masking + News-Alignment	0.569±0.007
with target news d'	Random Masking	0.574±0.009
	News-Alignment	0.559±0.009
	Random Masking + News-Alignment	**0.588±0.020**

Random Masking and News-Alignment. In both "with target news" and "without target news" setting, the performance of Random Masking and News-Alignment are better than the best baseline. The performance of Random Masking + News-Alignment are the best. This demonstrates that each module is beneficial to the model and they can compensate with each other and help model achieve better performance.

4.5 Experiments with different amounts of users from the target domain

We also evaluate the impact on transfer performance when varying the amounts of few-shot users from the target domain in training. In the MIND -> Adressa setting, we randomly sample 500, 1500, 2500, 3500, 4500 users who have less than 10 interactions in the target domain as the few-shot training samples and sample 8000 users as the test set. We select two strong baselines for comparison and the experiment results are shown in Figure 5 (more results are presented in Appendix A.2). We find that our method consistently achieves better performance, especially when the number of users

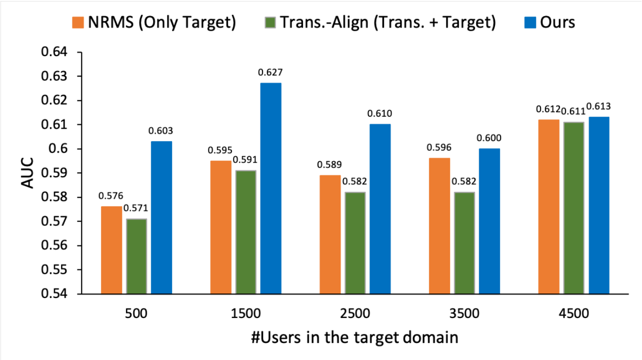


Figure 5: Performance when varying the amount of users involved in training from the target domain

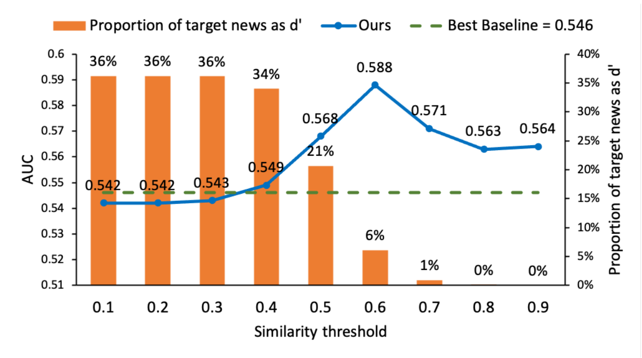


Figure 6: Influence of the similarity threshold for selecting the most similar target domain news

in training from the target domain is 500, 1500, and 2500. With the increase of training users, the improvements over the baselines decrease. This is an expected trend because domain transfer is less useful when more training data are available in target domain.

4.6 Influence of the similarity threshold

In the Cross-domain Extension module, we use a similarity threshold to control the selection of the most similar target domain news. The above reported results are from the setting of this threshold = 0.6, based on a grid search process. Here, we explore the influence of this similarity threshold when varying it from 0.1 to 0.9 in the MIND -> Adressa 4-shot setting. In Figure 6, the x-axis denotes the similarity threshold. The primary y-axis on the left denotes the AUC and the secondary y-axis on the right denotes the proportion of the target domain news selected as the most similar news d'. We find that if the similarity threshold is small (0.1∼0.4), although over 30% target news are selected as d' to participate in training, the recommendation model does not have a good performance. This is because the selected target news d' are not actually similar to source news d. The low threshold makes the unrelated news pass through training. With the increase of the similarity threshold, less but truly similar target domain news are used as d' for training. The threshold = 0.6 is the best setting, which introduces 6% of the target domain news in training. Note that if the threshold is large (0.7∼0.9), there are only few similar news found between the source domain and target domain. Although our method then has downgraded performance, it can still outperform the best baseline (Bilingual (Source+Trans.+Target)), demonstrating the effectiveness of our method on taking advantage of the introduced similar target news.

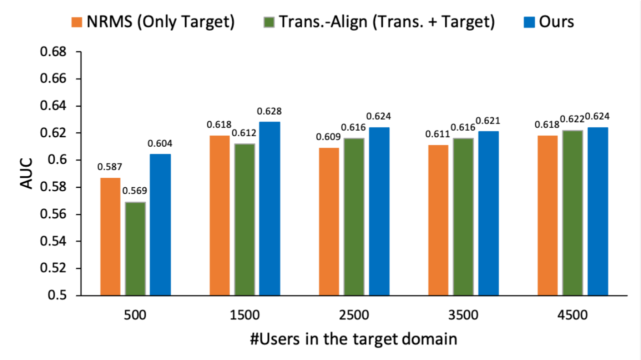


Figure 7: Performance when using the Multilingual Pretrained Language Model as news encoder

4.7 Experiments with the multilingual pretrained language model

The above reported experiments are from our method using NRMS [30] as the base recommendation model. In fact, our method is flexible for replacing the news encoder by any other language models. To further explore whether our strategies of reducing domain shift work well with other news encoders, we employ the multilingual pretrained language model - MultilingualBert [3] as news encoder. The training of news encoder will be based on the same strategies but just fine-tuning the language model parameters. We use MultilingualBert also for the NRMS and Trans.-Align(Trans.+Target) baseline. The performance comparison in terms of AUC is shown in Figure 7. More results are presented in Appendix A.3. We can see that for all methods in most settings, using the pretrained language model as the news encoder can relatively achieve better performance than the previous news encoder in NRMS. Comparing Figure 5 and 7, we can see that two baseline models benefit more from the multilingual news encoder than our method. However, our method can still outperform the baselines, at all different amounts of target domain users in training.

5 CONCLUSION

Few-shot news recommendation is a challenging and practical task for many unpopular or early-stage platforms. In this paper, we firstly solve this problem by cross-lingual transfer. To address the key challenge of domain shift in building a cross-lingual news recommendation model, we employ a shared news and user encoder between two domains to help transfer user-news preference patterns. We design three components which align two domains to reduce the domain shift. Based on existing news datasets, we construct two few-shot transfer settings and conduct extensive experiments. Experiment results show our model can outperform the baselines. A broader impact of our work is to help build recommender systems for countries or regions with low-resource languages. However, since our model is trained offline and may not be able to work well for breaking news that's never been seen before in neither the source and target domain.

REFERENCES

[1] Alexis Conneau and Guillaume Lample. 2019. Cross-Lingual Language Model Pretraining. In *NeurIPS*. Article 634.

[2] Gabriel de Souza Pereira Moreira, Felipe Ferreira, and Adilson Marques da Cunha. 2018. News Session-Based Recommendations using Deep Neural Networks. In *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.

[4] Hao Ding, Yifei Ma, Anoop Deoras, Yuyang Wang, and Hao Wang. 2021. Zero-Shot Recommender Systems. *arXiv preprint arXiv:2105.08318* (2021). arXiv:2105.08318 [cs.LG]

[5] Jon Atle Gulla, Lemei Zhang, Peng Liu, Özlem Özgöbek, and Xiaomeng Su. 2017. The Adressa Dataset for News Recommendation. In *Int. Conf. Web Intell.* 1042–1048.

[6] Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *LREC*.

[7] Guangneng Hu and Qiang Yang. 2021. TrNews: Heterogeneous User-Interest Transfer Learning for News Recommendation. In *EACL*. 734–744.

[8] Wouter IJntema, Frank Goossen, Flavius Frasincar, and Frederik Hogenboom. 2010. Ontology-Based News Recommendation. In *EDBT/ICDT Workshops*.

[9] Andrew Johnson, Penny Karanasou, Judith Gaspers, and Dietrich Klakow. 2019. Cross-lingual Transfer Learning for Japanese Named Entity Recognition. In *NAACL*. 182–189.

[10] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems – Survey and roads ahead. *Inf.Process.Manage.* 54, 6 (2018), 1203–1227. https://doi.org/10.1016/j.ipm.2018.04.008

[11] Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. 2013. The Plista Dataset. In *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge*. 16–23. https://doi.org/10.1145/2516641.2516643

[12] Roman Klinger and Philipp Cimiano. 2015. Instance Selection Improves Cross-Lingual Model Training for Fine-Grained Sentiment Analysis. In *CoNLL*. 153–163.

[13] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. 2019. MeLU: Meta-Learned User Preference Estimator for Cold-Start Recommendation. In *KDD*. 1073–1082. https://doi.org/10.1145/3292500.3330859

[14] Jingjing Li, Mengmeng Jing, Ke Lu, Lei Zhu, Yang Yang, and Zi Huang. 2019. From Zero-Shot Learning to Cold-Start Recommendation. In *AAAI*. Article 514, 8 pages. https://doi.org/10.1609/aaai.v33i01.33014189

[15] Ruirui Li, Xian Wu, Xian Wu, and Wei Wang. 2020. Few-Shot Learning for New User Recommendation in Location-Based Social Networks. In *WWW*. 2472–2478.

[16] Xin Li, Lidong Bing, Wenxuan Zhang, Zheng Li, and Wai Lam. 2021. Unsupervised Cross-lingual Adaptation for Sequence Tagging and Beyond. *arXiv preprint arXiv:2010.12405* (2021). arXiv:2010.12405 [cs.CL]

[17] Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common Sense Beyond English: Evaluating and Improving Multilingual Language Models for Commonsense Reasoning. In *ACL*. 1274–1287.

[18] Chen Lin, Runquan Xie, Xinjun Guan, Lei Li, and Tao Li. 2014. Personalized news recommendation via implicit social experts. *Inf. Sci.* 254 (2014), 1–18.

[19] Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. In *TPAMI*. 824–836.

[20] Tong Man, Huawei Shen, Xiaolong Jin, and Xueqi Cheng. 2017. Cross-Domain Recommendation: An Embedding and Mapping Approach. In *IJCAI*. 2464–2470.

[21] Owen Phelan, Kevin Mccarthy, Mike Bennett, and Barry Smyth. 2011. Terms of a feather: Content-based news recommendation and discovery using twitter. In *In ECIR*. 448–459.

[22] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 19–30.

[23] Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP. In *IJCAI*. 3853–3860.

[24] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP*. https://arxiv.org/abs/1908.10084

[25] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *EMNLP*. https://arxiv.org/abs/2004.09813

[26] Huimin Sun, Jiajie Xu, Kai Zheng, Pengpeng Zhao, Pingfu Chao, and Xiaofang Zhou. 2021. MFNP: A Meta-optimized Model for Few-shot Next POI Recommendation. In *IJCAI*. 3017–3023. https://doi.org/10.24963/ijcai.2021/415

[27] Michele Trevisiol, Luca Maria Aiello, Rossano Schifanella, and Alejandro Jaimes. 2014. Cold-Start News Recommendation with Domain-Dependent Browse Graph. In *RecSys*. 81–88.

[28] Chuhan Wu, Jianqiang Huang, Fangzhao Wu, Yongfeng Huang, Mingxiao An, and Xing Xie. 2019. NPA: Neural news recommendation with personalized attention. In *KDD*. 2576–2584.

[29] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Attentive Multi-View Learning. In *IJCAI*. 3863–3869.

[30] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Multi-Head Self-Attention. In *EMNLP*. 6389–6394.

[31] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2022. Personalized News Recommendation: Methods and Challenges. *ACM Trans. Inf. Syst.* (2022). https://doi.org/10.1145/3530257

[32] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering News Recommendation with Pre-Trained Language Models. In *SIGIR*. 1652–1656.

[33] Chuhan Wu, Fangzhao Wu, Yang Yu, Tao Qi, Yongfeng Huang, and Qi Liu. 2021. NewsBERT: Distilling Pre-trained Language Model for Intelligent News Application. In *EMNLP Findings*. 3285–3295.

[34] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In *ACL*. 3597–3606.

[35] Wenhui Yu, Xiao Lin, Junfeng Ge, Wenwu Ou, and Zheng Qin. 2020. Semi-supervised Collaborative Filtering by Text-enhanced Domain Adaptation. In *KDD*. 2136–2144.

[36] Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021. Cross-lingual Aspect-based Sentiment Analysis with Aspect Term Code-Switching. In *EMNLP*. 9220–9230.

[37] Li Zheng, Lei Li, Wenxing Hong, and Tao Li. 2013. PENETRATE: Personalized news recommendation using ensemble hierarchical clustering. *Expert Syst. Appl.* 40, 6 (2013), 2127–2136. https://doi.org/10.1016/j.eswa.2012.10.029

A ADDITIONAL EXPERIMENTAL RESULTS

A.1 NDCG and MRR for zero-shot and few-shot transfer

Here we report the NDCG@5, NDCG@10 and MRR performance of our methods and all baselines in both MIND -> Adressa and Adressa -> MIND zero-shot and few-shot setting. As shown in Table 4, our method achieves better performance in all metrics and all settings.

A.2 NDCG and MRR for experiments with different amounts of users from the target domain

As shown in Figure 8, the NDCG and MRR performance of our method are better than other baselines in the different amounts of users from the target domain setting.

A.3 NDCG and MRR for experiments with the Multilingual Pretrained Language Model

As shown in Figure 9, the NDCG and MRR performance of our method are better than other baselines in the experiments with the Multilingual Pretrained Language Model.

Table 4: Zero-shot and few-shot recommendation performance comparison. The best result is in bold font, and the best baseline is underlined. The unsolvable zero-shot case is indicated by –. Each experiment is repeated for 5 times and the mean and the standard deviation are reported.

(a) Metric: NDCG@5

Cross-domain Setting		MIND -> Adressa			Adressa -> MIND		
training data available in target domain		0-shot	2-shot	4-shot	0-shot	2-shot	4-shot
NRMS (Only Target) [30]		0.552±0.009	0.616±0.005	0.601±0.004	0.247±0.008	0.274±0.004	0.263±0.011
ZERO-SHOT (Source + Target) [36]		0.574±0.009	0.602±0.013	0.587±0.008	0.271±0.012	0.298±0.020	0.279±0.010
Trans.-Align (Trans. + Target) [36]		0.579±0.012	0.615±0.006	0.600±0.010	0.289±0.006	0.321±0.008	0.298±0.004
Bilingual (Source+Trans.+Target) [36]		0.577±0.010	0.607±0.014	0.608±0.014	0.288±0.009	0.314±0.008	0.303±0.005
TDAR [35]		–	0.581±0.005	0.576±0.010	–	0.277±0.007	0.280±0.010
Ours	Random Masking	0.588±0.005	0.625±0.009	0.625±0.005	**0.308±0.007**	0.323±0.014	**0.312±0.009**
	Random Masking+News-Align	**0.598±0.010**	**0.633±0.013**	**0.635±0.015**	0.306±0.005	**0.334±0.009**	**0.312±0.006**

(b) Metric: NDCG@10

Cross-domain Setting		MIND -> Adressa			Adressa -> MIND		
training data available in target domain		0-shot	2-shot	4-shot	0-shot	2-shot	4-shot
NRMS (Only Target) [30]		0.642±0.008	0.685±0.005	0.682±0.004	0.310±0.006	0.344±0.005	0.324±0.008
ZERO-SHOT (Source + Target) [36]		0.656±0.006	0.675±0.010	0.664±0.006	0.327±0.011	0.357±0.021	0.345±0.005
Trans.-Align (Trans. + Target) [36]		0.665±0.011	0.687±0.005	0.681±0.007	0.352±0.004	0.377±0.012	0.359±0.004
Bilingual (Source+Trans.+Target) [36]		0.663±0.008	0.681±0.013	0.688±0.014	0.346±0.006	0.375±0.006	0.361±0.007
TDAR [35]		–	0.660±0.005	0.662±0.008	–	0.342±0.008	0.343±0.008
Ours	Random Masking	0.672±0.004	0.695±0.007	0.702±0.003	0.364±0.004	0.383±0.011	**0.371±0.006**
	Random Masking+News-Align	**0.679±0.009**	**0.698±0.010**	**0.708±0.013**	**0.365±0.004**	**0.389±0.006**	0.370±0.005

(c) Metric: MRR

Cross-domain Setting		MIND -> Adressa			Adressa -> MIND		
training data available in target domain		0-shot	2-shot	4-shot	0-shot	2-shot	4-shot
NRMS (Only Target) [30]		0.381±0.009	0.436±0.005	0.426±0.006	0.244±0.006	0.271±0.002	0.257±0.007
ZERO-SHOT (Source + Target) [36]		0.397±0.007	0.423±0.012	0.404±0.006	0.260±0.011	0.281±0.019	0.270±0.007
Trans.-Align (Trans. + Target) [36]		0.408±0.013	0.437±0.006	0.425±0.007	0.280±0.005	0.305±0.011	0.284±0.005
Bilingual (Source+Trans.+Target) [36]		0.404±0.008	0.429±0.015	0.433±0.016	0.273±0.007	0.303±0.007	0.281±0.006
TDAR [35]		–	0.407±0.007	0.403±0.008	–	0.264±0.008	0.270±0.007
Ours	Random Masking	0.415±0.005	0.445±0.008	0.448±0.004	**0.295±0.004**	0.309±0.013	0.303±0.005
	Random Masking+News-Align	**0.422±0.009**	**0.449±0.012**	**0.455±0.014**	0.294±0.003	**0.317±0.005**	**0.304±0.004**

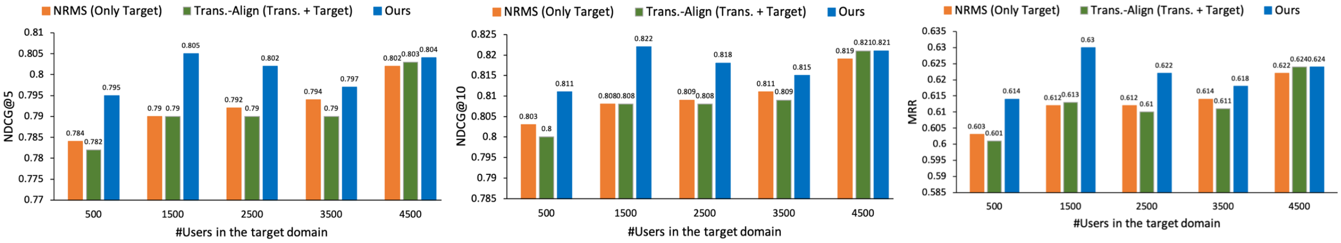


Figure 8: NDCG@5, NDCG@10 and MRR when varying the amount of users involved in training from the target domain.

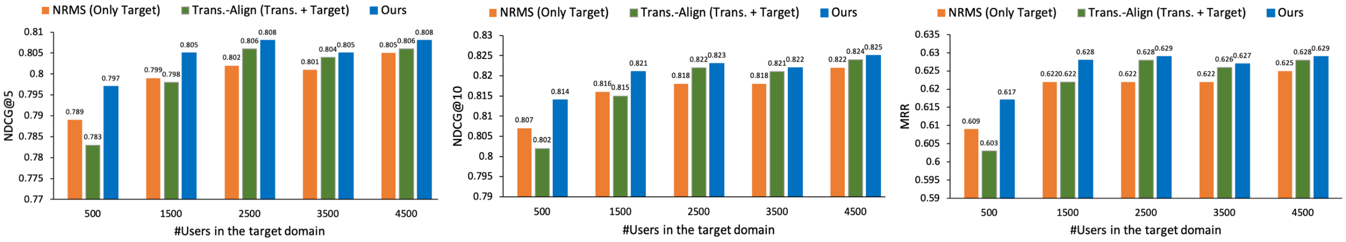


Figure 9: Experiments with the Multilingual Pretrained Language Model in terms of NDCG@5, NDCG@10 and MRR.