# Genre Separation Network with Adversarial Training for Cross-genre Relation Extraction

**Ge Shi**[1], **Chong Feng**[1]*, **Lifu Huang**[2], **Boliang Zhang**[2],
**Heng Ji**[2], **Lejian Liao**[1], **Heyan Huang**[1]

[1] Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, School of Computer, Beijing Institute of Technology, Beijing, 100081, China
{shige, fengchong, liaolj, hhy63}@bit.edu

[2] Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
{huangl7, zhangb8, jih}@rpi.edu

## Abstract

Relation Extraction suffers from dramatical performance decrease when training a model on one genre and directly applying it to a new genre, due to the distinct feature distributions. Previous studies address this problem by discovering a shared space across genres using manually crafted features, which requires great human effort. To effectively automate this process, we design a genre-separation network, which applies two encoders, one genre-independent and one genre-shared, to explicitly extract genre-specific and genre-agnostic features. Then we train a relation classifier using the genre-agnostic features on the source genre and directly apply to the target genre. Experiment results on three distinct genres of the ACE dataset show that our approach achieves up to 6.1% absolute F1-score gain compared to previous methods. By incorporating a set of external linguistic features, our approach outperforms the state-of-the-art by 1.7% absolute F1 gain. We make all programs of our model publicly available for research purpose [1] .

## 1 Introduction

Relation extraction aims to identify and categorize the semantic relation between two entity mentions based on the contexts within the sentence. Supervised learning approaches have shown to be effective on this task. However, as relation extraction highly depends on information about entities and their contexts, a supervised model trained in one genre suffers from dramatical performance decrease when applied to a new genre, due to the distinct contexts among various genres.

Previous studies (Plank and Moschitti, 2013; Nguyen and Grishman, 2014, 2015; Yu et al., 2015; Gormley et al., 2015) tackle this problem by manually crafting genre-agnostic features such as word clusters and word embeddings, to train a genre-shared relation extractor. These methods suffer from information loss due to the limited human knowledge to capture all genre-agnostic features. As depicted in Figure 1, where red rectangles are features shared by two genres, and blue and green triangles are source and target genre features respectively, *Feature Engineering* only captures a portion of the genre-agnostic features. Fu et al. (2017), depicted as *Feature Projection*, applies a domain adversarial neural network to automatically project the source and target genre features into one unified feature space. However, it unnecessarily introduces genre-specific features which undermine the overall performance.

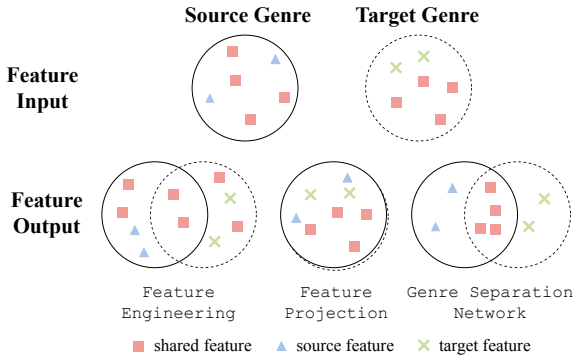


Figure 1: Comparison of Genre Separation Methods.

To address these problems, we propose a genre-separation network, which consists of two separate Convolutional Neural Networks (CNNs) to automatically separates genre-specific and genre-agnostic features for each genre, which is depicted as *Genre Separation Network* in Figure 1. To avoid information loss during feature encoding, we reconstruct the original input from the two separate feature spaces via a novel reconstruction loss. Then we use an adversarial similarity loss to limit the genre-agnostic features into one fea-

*Corresponding author

[1]We make all cleaned codes and resources publicly available at https://github.com/Garym713/Genre-Separation-Network-for-Relation-Extraction



*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1023
Brussels, Belgium, October 31 - November 4, 2018. 

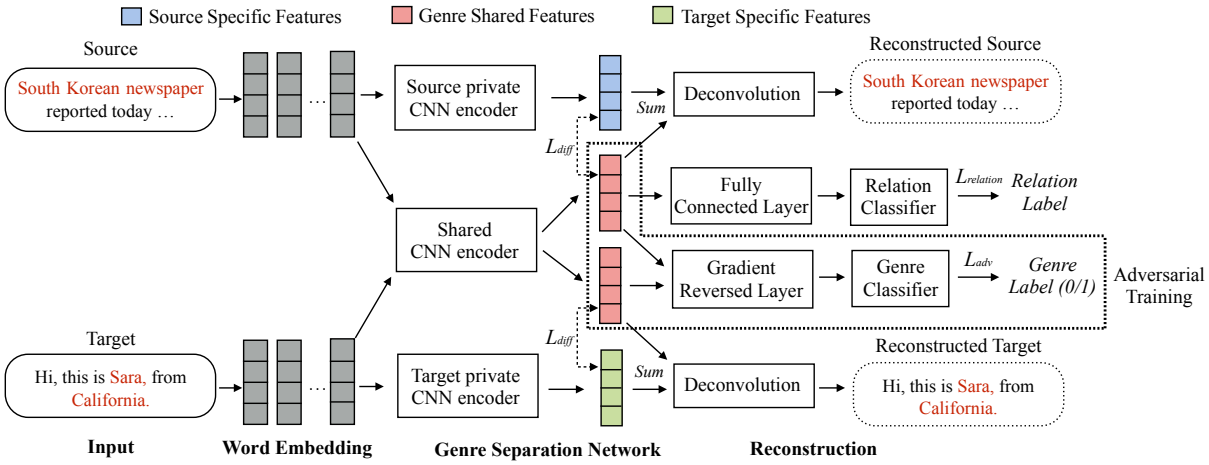


Figure 2: Overall genre separation framework for cross-genre relation extraction

ture space. The genre-agnostic features are finally used to predict entity relations in the source and target genres.

## 2 Approach

### 2.1 Overview

An overview of our framework is presented in Figure 2. We formulate the task as follows: given a labeled source genre corpus $S = \{(s_1, e_{11}, e_{12}, r_1), ..., (s_n, e_{n1}, e_{n2}, r_n)\}$, where $s_i = [w_{i1}, ..., w_{im}]$ denotes a sentence. $e_{i1}$ and $e_{i2}$ denote two entity mentions, and $r_i$ denotes the relation betwern $e_{i1}$ and $e_{i2}$, we build a relation extraction model on $S$ and apply it to a different target genre corpus $T = \{(\hat{s_1}, \hat{e_{11}}, \hat{e_{12}}), ..., (\hat{s_n}, \hat{e_{n1}}, \hat{e_{n2}})\}$.

### 2.2 Genre Separation Network (GSN)

As shown in Figure 1, our goal is to distinguish the genre-agnostic features (red rectangles) and genre-specific features (blue triangles and green crosses). Using source genre as an example, we apply a source private CNN encoder on the source sentence to generate the source-specific feature representation $f_s^p$, and a shared CNN encoder to generate genre-agnostic feature $f_s^c$. Similarly, we get $f_t^p$ and $f_t^c$ from the target private CNN encoder and the shared CNN encoder respectively. To separate $f_s^p$ from $f_s^c$ and separate $f_t^p$ from $f_t^c$, we introduce a *difference loss* following previous studies (Bousmalis et al., 2016; Liu et al., 2017). More details will be elaborated below.

Formally, given a source sentence $(s, e_1, e_2, r)$ where $s = [w_1, ..., w_m]$, for each word $w_{ik}$, we generate a multi-type embedding: $\tilde{v}_i = [v_i, p_i, \tilde{p}_i, t_i, \tilde{t}_i, c_i, \eta_i]$ where $v_i$ denotes a pre-trained word embedding. $p_i$ and $\tilde{p}_i$ are position embeddings (Al-Badrashiny et al., 2017) indicating the distance from $w_i$ to $e_1$ and $e_2$ respectively. $t_i$ and $\tilde{t}_i$ are entity type embedding (Ren et al., 2016; Huang et al., 2016) of $e_1$ and $e_2$. $c_i$ is the chunking embedding, and $\eta_i$ is a binary digit indicating whether the word is within the shortest dependency path between $e_1$ and $e_2$ (Bunescu and Mooney, 2005; Liu et al., 2015; Huang et al., 2017). All these embeddings except pre-trained word embedding are randomly initialized and optimized during training. Thus the input layer is a sequence of word representations $V = \{\tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_n\}$. We then apply the convolution weights $W$ to each sliding n-gram phrase $g_j$ with a biased vector $b$, i.e., $g_j^{'} = \tanh(W \cdot V) + b$. All n-gram representations $g_j^{'}$ are further used to get an overall vector representation $f$ by max-pooling.

Once we obtain $f_s^p$, $f_s^c$, $f_s^p$ and $f_s^c$, we compute the *difference loss*:

$$L_{diff} = ||f_s^{p\top} \cdot f_s^c + f_t^{p\top} \cdot f_t^c||_2^F$$

where $||.||_2^F$ represents the squared Frobenius norm.

To limit the genre-agnostic features from various genres into a shared feature space, we further design a genre adversarial training component. We take the genre-agnostic features from both source genre and target genre as input to a Gradient Reversed Layer(GRL) (Ganin et al., 2016), which acts as a general hidden layer in forward process and reverses the gradient in loss backward phase to confuse the genre classifier, so that it cannot distinguish the input features from the source genre

to the target genre:

$$L_{adv} = \sum_{i=0}^{N_s+N_t} d_i l \log(\hat{d_i}) + (1-d_i)\log(1-\hat{d_i})$$

where $d_i \in \{0, 1\}$ indicates the samples from the source genre or the target genre, and $N_s, N_t$ refer to the number of examples in the source genre and the target genre respectively. The term $\hat{d_i}$ represents the probability of the sample from the source genre, which is acquired by a linear function of the genre classifier.

### 2.3 Genre Reconstruction

Till now, we can separate the features of each genre into two separated feature spaces by optimizing $L_{diff}$ and $L_{adv}$. However, there is no guarantee that the separated feature spaces are actually meaningful. From equation $L_{diff}$, we can see that the $f_s^p, f_t^p$ would be easily optimized to zero if we did not place a constraint, in which case the model would fail to train. Therefore, we further reconstruct the input sentence from both genre-specific features and genre-agnostic features.

For each genre, e.g., the source genre, we first sum the genre-specific feature vector $f_s^p$ and genre-agnostic feature vector $f_s^c$, i.e., $f_s = f_s^p + f_s^c$. We take $f_s$ as input to an unpooling layer (Zeiler and Fergus, 2014) followed by a deconvolutional neural netowrk (DcNN) (Xu et al., 2014). The output of DcNN will include the same number of decoded vectors $V^* = \{\tilde{v}_1^*, \tilde{v}_2^*, ..., \tilde{v}_n^*\}$ as input $V = \{\tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_n\}$. We optimize the DcNN with the following reconstruction loss:

$$L_{rec} = 1 - \sum_{i=0}^{n} |cos(\tilde{v}_i, \tilde{v}_i^*)|$$

where $n$ indicates the total number of words in the input sentence, $\tilde{v}_i$ represents the input word representation described in Section 2.2, and $\tilde{v}_i^*$ is the corresponding reconstructed vector from DcNN.

### 2.4 Cross Genre Relation Extraction

We next utilize the genre-agnostic features from the source genre $f_s^c$ to train a relation classifier. We first feed $f_s^c$ into a fully connected layer and obtain a dense vector, then we use a linear projection function with a softmax as the relation classifier to determine the relation type

$$L_{relation} = \sum_{i=0}^{N_s} \sum_{k=0}^{K} -x_k \log(x_k)$$

where $K$ is the total number of relation types. $x_k$ represents the probability of entities being classified to category $k$.

We finally linearly combine all the loss functions and jointly optimize the model using SGD (Bottou, 2010).

$$L = L_{relation} + \alpha L_{diff} + \beta L_{rec} + \gamma L_{adv}$$

where $\alpha, \beta, \gamma$ denote the weights of various losses.[2]

## 3 Experiments

### 3.1 Data and Parameters

We evaluate our approach on the English portion of ACE2005 dataset (Walker et al., 2006; Ji et al., 2010; Hong et al., 2015; Yu et al., 2016). It covers 6 genres: Newswire (nw), Broadcast Conversation (bc), Broadcast News (bn), Telephone Speech (cts), Usenet Newsgroups (un), and Weblogs (wl) and 11 relation types. Following previous work (Yu et al., 2015; Nguyen and Grishman, 2015; Gormley et al., 2015), we use newswire and broadcast news (nw&bn) as training data, half of bc as development set, and test the model on the remaining half of bc, cts, wl. We conduct the same preprocessing steps as previous work and yield 43,497 entity pairs in total for training.

Table 3.1 shows the hyper-parameters that we use to train our model.

| Hyper-parameters | Value |
|---|---|
| # of Filters in Shared/Private CNN encoder | 800 |
| Filter Width | 3 |
| Hidden Size of Fully Connected Layer | 300 |
| Position Embedding Size | 25 |
| Entity Type/Chunking Embedding Size | 25 |
| Optimizer | SGD |
| Learning rate | 0.001 |
| Pre-trained Word Embedding | Glove-100[3] |

Table 1: Hyper-parameters

### 3.2 Baseline Models

We compare our approach with the following methods:

**FCM** (Gormley et al., 2015) is a feature combination model which composes word embeddings with traditional linguistic features.

**Hybrid FCM** (Gormley et al., 2015) incorporates many more selected linguistic features compared to FCM.

[2]We set $\alpha = 0.075, \beta = 0.01, \gamma = 0.25$ when the model performs the best on the development set.

**LRFCM** (Yu et al., 2015) is a feature compositional model which scales to more features and more labels.

**Log-linear & DNN** (Nguyen and Grishman, 2015) explores CNN, Bi-GRU, Forward GRU, Backward GRU, and log-linear model for relation extraction. We compare against the performances of individual models instead of assembled models.

**CNN+DANN** (Fu et al., 2017) utilizes domain adversarial training to automatically extract genre-agnostic features for source and target genre within one feature space.

### 3.3 Comparison and Analysis

Table 2 shows the cross-genre relation extraction performance among various methods. Our approach significantly outperforms all previous baselines by 1.2%-1.7% (F1). Table 3 presents the results without using extra linguistic features (only embedding based features), our approach achieves 2.9%-6.1% absolute gain over baselines. The ablation test by removing each component at a time justifies the contribution of each method. The difference and reconstruction components ensure the features to be separated into shared and private spaces, and they can remove redundant genre-specific features to some extent. That's why we got a significant F-score improvement when only utilizing these two components. The adversarial training component can further encourage the features of each genre from the shared encoder to be close to each other, thus the performance is further improved. We also conduct ablation experiments on each feature components. Among the linguistic features we used, the entity type and position features contribute the most to the performance. For example, the relation extraction performance decreases by about 8% if removing the entity type feature. We analyze the reasons and find that the entity type feature is vital to ensure the types of two entity mentions to be consistent with the hard entity type constraint of each relation type defined in ACE schema.

For the remaining errors, we notice that our model easily fails to predict relations between nested entity mentions. For example, in "*Our president has put homeland security in the hands of failed Republican hacks.*", our model mistakenly predicts the relationship between *Republican* and *failed Republican hacks* as *None* instead of *organization-affiliation*, due to the lack of context information. Besides, we also observe some failed cases where the two entities are separated with a extreme wide context, which suggests us to incorporate dependency path based deep neural networks into the framework.

| System | bc | cts | wl |
|---|---|---|---|
| FCM | 61.9 | 52.93 | 50.36 |
| Hybrid FCM | 63.48 | 56.12 | 55.17 |
| LRFCM | 59.4 | - | - |
| Log-linear | 57.83 | 53.14 | 53.06 |
| CNN | 63.26 | 55.63 | 53.91 |
| Bi-GRU | 63.07 | 56.47 | 53.65 |
| Forward GRU | 61.44 | 54.93 | 55.10 |
| Backward GRU | 60.82 | 56.03 | 51.78 |
| CNN+DANN | 65.16 | - | - |
| w/o Difference | 59.87 | 54.10 | 52.73 |
| w/o Adversarial | 64.42 | 57.32 | 55.63 |
| w/o Reconstruction | 59.12 | 53.48 | 53.17 |
| Our Approach | **66.38** | **57.92** | **56.84** |

Table 2: Cross Genre Relation Extraction Performances (Macro F-score %) on Various Genres. *w/o Difference* means to ablate the $L_{diff}$ loss. *w/o Adversarial* means to ablate the adversarial training component. *w/o Reconstruction* means to ablate the genre reconstruction component.

| System | bc | cts | wl |
|---|---|---|---|
| CNN | 46.3 | 40.8 | 35.8 |
| GRU | 45.2 | 40.2 | 35.1 |
| Bi-GRU | 46.7 | 41.2 | 36.5 |
| Our Approach | **52.8** | **45.3** | **39.4** |

Table 3: Cross Genre Relation Extraction Performances (Macro F-score %) on Various Genres (without linguistic features)

## 4 Related Work

Previous studies on cross-genre relation extraction either manually or automatically extract genre-agnostic features (Plank and Moschitti, 2013; Nguyen and Grishman, 2014; Yu et al., 2015; Gormley et al., 2015; Nguyen and Grishman, 2015), suffering from human labor and limited coverage of effective features, or automatically project source and target genres into one unified feature space and learn genre shared features (Fu et al., 2017), which inevitably introduces noise from genre specific features. Compared with these methods, our approach separates genre-specific features from genre-agnostic features first, and then automatically extracts meaningful features for cross-genre relation extraction.

Our work is also related to studies on domain separation networks (Bousmalis et al., 2016; Liu et al., 2017; Chen et al., 2017), which explicitly extracts features from two separate subspaces: domain-specific and domain-agnostic. We adopt

a similar framework for cross-genre relation extraction and introduce a novel reconstruction component which is proved to be suitable to relation extraction.

## 5 Conclusions

We propose a genre separation framework for cross-genre relation extraction. Without requiring human crafted features, this framework can effectively separate genre-specific features from genre-agnostic ones, and automatically extract meaningful features for the task. To ensure the separation of features within each genre and enforce the genre agnostic features from source genre and target genre to be in the same feature space, we design a difference loss and an adversarial training component. Experiments on various genres demonstrate the effectiveness of our framework. In the future, we will extend our framework to cross-lingual and cross-domain information extraction tasks.

## 6 Acknowledgements

We thank the anonymous reviewers for their valuable comments and suggestions. This work is supported by the National Key Research and Development Program of China No. 2017YFB1002101, the National Natural Science Foundation of China No. U1636203, China Scholarship Council CSC, No. 201706030131. RPI co-authors were supported by the U.S. DARPA AIDA Program No. FA8750-18-2-0014 and U.S. ARL NS-CTA No. W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

## References

Mohamed Al-Badrashiny, Jason Bolton, Arun Tejasvi Chaganty, Kevin Clark, Craig Harman, Lifu Huang, Matthew Lamm, Jinhao Lei, Di Lu, Xiaoman Pan, et al. 2017. Tinkerbell: Cross-lingual cold-start knowledge base construction. In *TAC*.

Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Advances in Neural Information Processing Systems*, pages 343–351.

Razvan C Bunescu and Raymond J Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 724–731. Association for Computational Linguistics.

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-criteria learning for chinese word segmentation. *arXiv preprint arXiv:1704.07556*.

Lisheng Fu, Thien Huu Nguyen, Bonan Min, and Ralph Grishman. 2017. Domain adaptation for relation extraction with domain adversarial neural network. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 425–429.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Matthew R Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. *arXiv preprint arXiv:1505.02419*.

Yu Hong, Di Lu, Dian Yu, Xiaoman Pan, Xiaobin Wang, Yadong Chen, Lifu Huang, and Heng Ji. 2015. Rpi blender tac-kbp2015 system description. In *Proc. Text Analysis Conference (TAC2015)*.

Lifu Huang, Jonathan May, Xiaoman Pan, and Heng Ji. 2016. Building a fine-grained entity typing system overnight for a new x (x= language, domain, genre). *arXiv preprint arXiv:1603.03112*.

Lifu Huang, Avirup Sil, Heng Ji, and Radu Florian. 2017. Improving slot filling performance with attentive neural networks on dependency structures. *arXiv preprint arXiv:1707.01075*.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *Third Text Analysis Conference (TAC 2010)*, volume 3, pages 3–3.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.

Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng Wang. 2015. A dependency-based neural network for relation classification. *arXiv preprint arXiv:1507.04646*.

Thien Huu Nguyen and Ralph Grishman. 2014. Employing word representations and regularization for domain adaptation of relation extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 68–74.

Thien Huu Nguyen and Ralph Grishman. 2015. Combining neural networks and log-linear models to improve relation extraction. *arXiv preprint arXiv:1511.05926*.

Barbara Plank and Alessandro Moschitti. 2013. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1498–1507.

Xiang Ren, Wenqi He, Meng Qu, Lifu Huang, Heng Ji, and Jiawei Han. 2016. Afet: Automatic fine-grained entity typing by hierarchical partial-label embedding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1378.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57.

Li Xu, Jimmy SJ Ren, Ce Liu, and Jiaya Jia. 2014. Deep convolutional neural network for image deconvolution. In *Advances in Neural Information Processing Systems*, pages 1790–1798.

Dian Yu, Xiaoman Pan, Boliang Zhang, Lifu Huang, Di Lu, Spencer Whitehead, and Heng Ji. 2016. Rpi blender tac-kbp2016 system description. In *Proceedings of the 2016 Text Analysis Conference (TAC2016)*.

Mo Yu, Matthew R Gormley, and Mark Dredze. 2015. Combining word embeddings and feature embeddings for fine-grained relation extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1374–1379.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.