Discovering Latent Information from Noisy Sources in the Cultural Heritage Domain*

Fabrizio Scarrone

Università degli studi di Torino

Abstract. Today, there are many publicly available data sources, such as online museum catalogues, Wikipedia, and social media, in the cultural heritage domain. Yet, the data is heterogeneous and complex (diverse, multi-modal, sparse, and noisy). In particular, availability of *social media (such as Twitter messages) is both a boon and a curse* in this domain: social media messages can potentially provide information not available otherwise, yet such messages are short, noisy, and are dominated by grammatical and linguistic errors. The key claim of this research is that the availability of publicly available information related to the cultural heritage domain can be improved with tools capable of signaling to the various classes of users (such as the public, local governments, researchers) the entities that make up the domain and the relationships existing among them. To achieve this goal, I focus on developing novel algorithms, techniques, and tools for leveraging multi-modal, sparse, and noisy data available from multiple public sources to integrate and enrich useful information available to the public in the cultural heritage domain. In particular, research aims to develop novel models that take advantage of multi-modal features extracted by deep neural models to improve the performance for various underlying tasks.

Keywords: Multi-modal · information extraction · information integration and latent information discovery · entity recognition · neural-networks · attention models · cultural heritage domain

1 Introduction

Today, there are many publicly available data sources, such as online museum catalogues(e.g. [17]), Wikipedia (e.g. [18]), and social media (e.g. [19]), in the cultural heritage domain. The key premise of the research is that the availability of publicly available information related to the cultural heritage domain can be significantly improved with tools capable of signaling to the various classes of users (such as the public, local governments, researchers) the entities that make up the domain and the relationships existing among them. Yet, realizing this idea is non-trivial due to the complexity (multi-modality, sparsity, and noise) of the data available in this domain (Figure 1).

* Results presented in this paper were obtained using the Chameleon testbed.

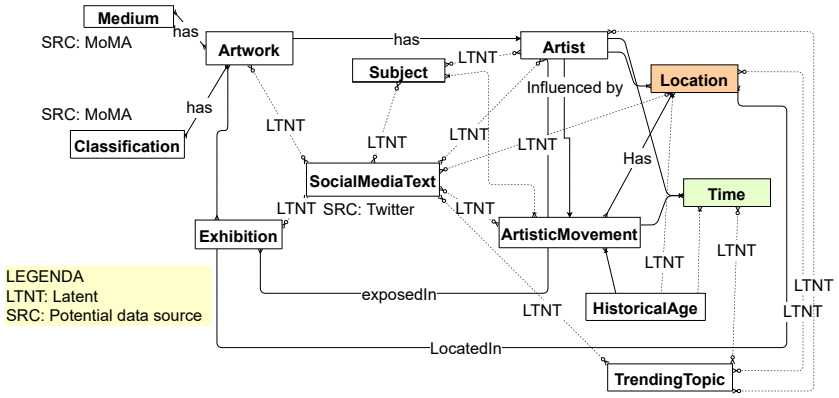

Fig. 1. A simplified ER-diagram outlining some of the data sources and their contributions to the information available in the cultural heritage domain – here any edge marked as "LTNT" represent a latent relationship that needs to be extracted

Problem Statement Given the above context, in this thesis, I focus on developing novel algorithms, techniques, and tools for leveraging multi-modal, sparse, and noisy data available from multiple public sources to integrate and enrich useful information available to the public in the cultural heritage domain.

2 Methodology

Figure 1 includes an ER-diagram outlining some of the data sources and their contributions to the information available in the cultural heritage domain. For example, one can observe that a number of entities (such as *artwork*, *artist*, *exhibition*, *artistic movement*) are available from various sources. Some of the relationships among this entities can be explicit in these sources (such as an *artwork* is being associated to an *artist* or an *artist* being included in an *exhibition*), while some other relationships, such as an *artist* being influenced by an *artistic movement* or an *artist* being interested on a particular *subject* may be latent – in Figure 1, such latent relationships are highlighted with edges marked as "LTNT" and they represent areas of interest for this research.

2.1 Social Media to our Help (?)

Availability of *social media (such as Twitter messages) is both a boon and a curse* in this domain:

– *How can social media help in this domain?:* Social media messages can potentially provide information not available otherwise: many museums or collections post regular Twitter messages about *artists*, *artworks*, or *exhibitions*; in addition, there may be online communities that discuss a particular *artistic movement*. These messages may not only provide data points that are not available otherwise, but can also help contextualize available data or help discover latent relationships among entities in the domain of interest.

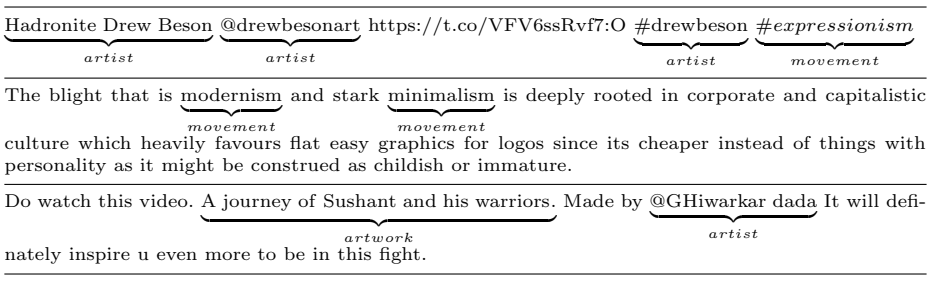

Fig. 2. Sample Twitter messages and associated entity types

– *Why is leveraging social media in this domain challenging?:* Yet, despite the opportunities they provide, leveraging social media data in this context is not trivial: Twitter messages are short, noisy, and are largely dominated by text full of grammatical and linguistic errors. Moreover, many messages in this domain are automatically generated through APIs and their structures are not linguistically based. Being extremely short, they often lack context to help interpret their content. In fact, our experience with Mechanical Turk [20], has shown that even manually labeling portions of the messages for a supervised methodology is difficult due to the underlying ambiguities.

2.2 A Multi-Modal Approach

It is important to note that social media, in this context, is often multi-modal in that many messages in this domain are accompanied by visuals – our experience has shown that roughly 30% of the messages have one or more associated images. Recently, joint learning from sources with different underlying modalities has become an emerging research interest and, with the rise of deep learning techniques, such approaches have found increasing use in diverse domains, such image understanding/annotation [21], and natural language processing [5].

While in theory, these visuals can also help provide context necessary to better interpret these social media messages for effective information extraction and integration, in practice, these visuals can be very diverse (such as the visual representation of the artwork, picture of an artist, a snapshot of the exhibit venue, or an announcement flyer) and they themselves lack descriptive labels. Consequently, leveraging such visual data to implement a multi-modal approach is not possible with existing techniques. We next illustrate this with a specific learning task: *entity and entity type recognition*.

2.3 A Specific Task: Entity and Entity Type Recognition

Consider the task of recognizing where entities of various types, such as *artists* (A), *artworks* (W), *movements* (M), or *venues* (V), occur in a given Twitter as visualized in Figure 2. As the examples in this figure illustrate, this is a highly difficult task, primarily because, the tweets are short, many times poorly organized, and they lack context to help identify that entities they may contain.

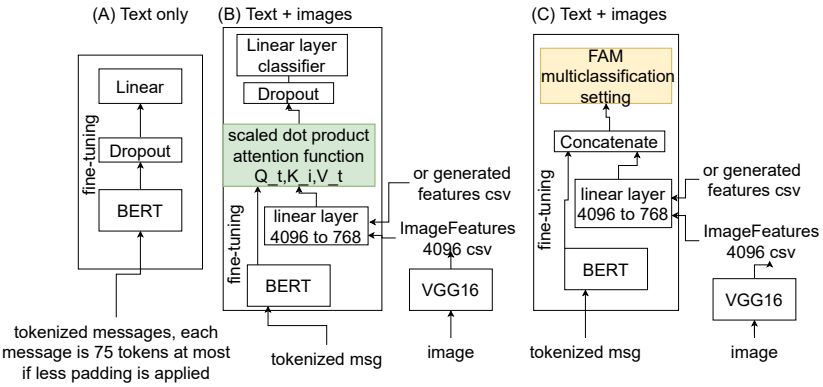

Fig. 3. Three different models

The closest literature to this task is on the *named entity recognition* (NER) problem, where one aims to discover entities, such as persons, locations, that are referred to by their names in a given text document corpus [11, 9, 10]. Existing approaches to this problem are often studied in four major groups: 1) rule-based approaches rely on hand-crafted rules [12]; unsupervised learning approaches [22] attempt to solve the problem without the help of hand-labeled training examples; 3) feature-based supervised learning techniques rely on with careful feature engineering; and 4) more recent deep-learning based approaches aim to automatically discover latent representations needed for the classification and/or detection task from the raw input [15, 1, 2, 6, 5]. This latter approach can benefit from non-linear transformations and eliminate (or reduce) the efforts needed for designing hand crafted features for this task. However, existing deep-learning based approaches require significant amount of contextual information to help interpret the text and perform poorly when provided out-of-context and short Twitter messages as input data. Therefore, I am developing novel, multi-modal approach to addressing this extremely difficult task.

Data Set Using the Twitter API, we are building a Cultural Heritage dataset collecting related tweets with their associated images. The entities in the text messages are manually labeled by Mechanical Turkers, using the BIO tagging scheme [7][1] Latent embeddings for the textual part of the messages are obtained using BERT [1] a model Trained on a large corpus of unlabelled text including the entire Wikipedia and Book Corpus. Bert generates context related embedding at sub-word level. Latent embeddings for the images associated the to messages, on the other hand, are obtained from VGG16 [3] a convolutional model trained using ImageNet 15 million labeled images in 22,000 categories.

Novel Multi-Modal Attention Mechanisms As we mentioned above, a particular challenge in addressing the entity and entity type recognition challenge

[1] We will make this data set publicly available to the research community.

F1 Score, 1000 samples

Entity	text only	text+images	ideal fts
A	0,597	0,458	0,598
M	0,734	0,449	0,792
V	0,667	0,071	0,714
W	0,612	0,255	0,619

F1 Score, 1500 samples

Entity	text only	text+images	ideal fts
A	0,678	0,545	0,728
M	0,827	0,675	0,895
V	0,538	0,456	0,690
W	0,621	0,449	0,708

Table 1. F1 score (75% training, 15% validation, and 10% testing); each model has been trained for 600 epochs with a batch size of 32 and AdamW optimizer [8] – here "ideal features" corresponds to a scenario where text messages are paired with synthetically-constructed ideal visual features to assess the maximum possible improvements we can expect from FAM

is the lack of context when interpreting the messages. I argue that multi-modal attention, if used effectively, can help alleviate this problem. Attention as a mechanism by which a network can capture the interaction among elements of a given data object (or across multiple data objects) to discover features weights and use this weighting to help improve the network performance [13, 14, 2]. Since our goal is to leverage visual information to help us provide context to recognize entities in short text messages, we are especially interested in cross-attention mechanisms, such as [21]. Figure 3 depicts, the possible models: Model (A) uses only the text messages and applies a linear layer after the Bert model in order to classify the tokens in the available classes. Model (B) uses text messages and images embeddings – Bert embeddings and visual features are combined in the attention module and the result goes through a linear model for classification. The attention applied in this case is similar to that the one present in [2].

I am, however, proposing a novel cross-attention mechanism, depicted as Model (C): in this model textual and visual embeddings are passed to a novel factorizing attention module (FAM) module for a multi-classification setting, inspired by factorization machines [4] – FAM accounts for the second order interactions within and across textual and visual features while analyzing the combined sparse feature space for feature weights that will inform the search for entities in short text. As shown in Figure 1, FAM has the potential to provide significant improvements in F1 scores for all entity types when provided idealized features. While the results are less accurate when provided real images, the difference quickly improves as the number of samples increases from 1000 to 1500, which indicates that with a reasonable training corpus, the provided FAM will surpass the text-only accuracies.

3 Conclusions and Future Work

Here, I provided an overview of my PhD work on improving information availability in the cultural heritage domain. My research is focusing on novel algorithms and techniques these can include novel multi-modal, cross-attention mechanism but also instance and n-shot learning, for leveraging multi-modal, sparse, and noisy data available from multiple public sources to improve information quality. Future work will include building a sufficiently large corpus to train and evaluate algorithms and tackle latent information extraction tasks outlined in Figure 1.

References

1. Devlin, J., Chang, M., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding, NAACL-HLT, (2019)
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Illia Polosukhin, Attention Is All You Need, In Advances in Neural Information Processing Systems, pages 6000–6010, (2017).
3. Karen Simonyan, Andrew Zisserman: Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR 2015
4. Rendle, S., Factorization Machines, In , 2010 IEEE International Conference on Data Mining (2010)
5. Asgari-Chenaghlu, M., Feizi-Derakhshi, M. R., Farzinvash, L., Balafar, M. A., and Motamed, C.: A multimodal deep learning approach for named entity recognition from social media, CoRR, (), (2020).
6. Zhang, Q., Fu, J., Liu, X., and Huang, X.: Adaptive co-attention network for named entity recognition in tweets, In , AAAI (2018).
7. https://en.wikipedia.org/wiki/Inside_outside_beginning_(tagging)
8. Loshchilov, I., and Hutter, F., Decoupled Weight Decay Regularization, ICLR (2019).
9. Li, J., Sun, A., Han, J., & Li, C., A survey on deep learning for named entity recognition, CoRR, (), (2018).
10. Yadav, V., & Bethard, S., A survey on recent advances in named entity recognition from deep learning models, CoRR, (), (2019).
11. Grishman, R. and Sundheim, B. 1996. Message understanding conference-6: A brief history. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, volume 1. (1996)
12. Sekine, S. and Nobata, C., "Definition, dictionaries and tagger for extended named entity hierarchy." in LREC, pp. 1977–1980, (2004)
13. Bahdanau, D., Cho, K., and Bengio, Y., Neural machine translation by jointly learning to align and translate, CoRR, (), (2014).
14. Luong, T., Pham, H., and Manning, C. D., Effective approaches to attention-based neural machine translation, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, (), (2015). http://dx.doi.org/10.18653/v1/d15-1166
15. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," J. Mach. Learn. Res., vol. 12, no. Aug, pp. 2493–2537, 2011
16. Baltrusaitis, T., Chaitanya Ahuja and Louis-Philippe Morency. "Multimodal Machine Learning: A Survey and Taxonomy." IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2019): 423-443.
17. MuseumofModernArt, GitHub repository, https://github.com/MuseumofModernArt/collection
18. Wikipedia, The Free Encyclopedia, "List of art movements", https://en.wikipedia.org/wiki/List_of_art_movements, (accessed July 21, 2021)
19. MoMA The Museum of Modern Art [@@MuseumModernArt].Tweets [Twitter profile]. Twitter. https://twitter.com/museummodernart, (Accessed July 21, 2021)
20. Amazon Mechanical Turk, https://www.mturk.com/, (Accessed July 21, 2021)
21. Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, Feng Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10941-10950
22. D. Nadeau, P. D. Turney, and S. Matwin, "Unsupervised named entity recognition: Generating gazetteers and resolving ambiguity," in CSCSI, 2006, pp. 266–277.