Towards Machine Reading for Interventions from Humanitarian-Assistance Program Literature

Bonan Min, Yee Seng Chan, Haoling Qiu, and **Joshua Fasching**
Raytheon BBN Technologies, Cambridge, Massachusetts
{bonan.min,yeeseng.chan,haoling.qiu,joshua.fasching}@raytheon.com

Abstract

Solving long-lasting problems such as food insecurity requires a comprehensive understanding of interventions applied by governments and international humanitarian assistance organizations, and their results and consequences. Towards achieving this grand goal, a crucial first step is to extract past interventions and when and where they have been applied, from hundreds of thousands of reports automatically. In this paper, we developed a corpus annotated with interventions to foster research, and developed an information extraction system for extracting interventions and their location and time from text. We demonstrate early, very encouraging results on extracting interventions.

1 Introduction

The world is a complex socio-political system: there are long lasting problems such as food insecurity, global warming and diseases affecting much of the world's population, as well as bursting, extreme events such as war, natural disasters and financial crisis. Recently, there has been growing interests in applying event extraction to provide better situation awareness ("what happened"), but rarely in terms of offering insight into providing guidance on what humanitarian assistance interventions have been applied in the past and how they influence the situation.

Furthermore, interventions may have intended outcomes and unintended consequences. An example of an unintended consequence is that free food distribution depresses prices for local produce, and creates a disincentive for farmers. It is extremely useful to automatically extract interventions, their outcome and consequences from hundreds of thousands of articles, to provide to decision makers in governments or humanitarian aid organizations a comprehensive understanding of what intervention options are available when crisis happened, and what outcomes to expect when applying each of them.

This paper is a first step in this direction, starting with developing Information Extraction (IE) techniques to automatically extract interventions from text including academic studies, program/project guidance and evaluation documents from non-profit or international organizations.

We view interventions as a (series of) event(s) [1] with time and space dimensions. For example:

S1: *WFP is scaling up its* **food assistance** *activities in Baghdad, Anbar, Dohuk and Ninewa governorates in 2018.*

The intervention "Food Assistance" is happening in the year 2018 in locations "Baghdad, Anbar, Dohuk and Ninewa governorates". Being able to extract interventions and their location and time is a first step towards enabling comprehensive understanding of their effects and consequences.

In this paper, we develop IE techniques towards reading for interventions. The basic methodology is to treat intervention extraction as an event extraction problem, and read intentional and factual statements about interventions in existing project documentation and evaluations literature. Our contributions are three fold:

- We construct a new corpus, annotated with interventions to foster research.
- We develop an IE algorithm for extracting interventions and their locations and time.
- Experiments show the effectiveness of our approach.

We discuss related work in the next section and describe our intervention extraction models in Section 3. In Section 4, we describe our ontology

[1] The main difference between an intervention and other events is that an intervention is a deliberate policy choice, whereas famine or a terrorist attack, for example, is not.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6444–6448, Hong Kong, China, November 3–7, 2019.

of intervention types and the intervention dataset. We present experiment results in Section 5, before concluding in Section 6. The intervention corpus and source code are available at https://github.com/BBN-E/mr-intervention.

2 Related Work

Event extraction is often formulated as a multi-stage (Ahn, 2006) classification (trigger classification then argument identification) problem. Prior works either use high-level features (Huang and Riloff, 2012; Ji and Grishman, 2008) or are Neural Network models (Chen et al., 2015). Nguyen (2016) propose joint event extraction using recurrent neural networks.

In need of labeled datasets for training models and evaluation, datasets such as MUC (Grishman and Sundheim, 1996), ACE (Doddington et al., 2004) and Situation Frames (Strassel et al., 2017) have been developed. There are also datasets created for specific domains. An example is the GENIA biomedical event annotation (Kim et al., 2008). Our work is the latest continuation along this path: creating a dataset to foster research in automatically extracting interventions from text, and demonstrating encouraging results.

3 Extraction Models

We model interventions as events. Given a sentence, we perform intervention extraction using a two-stage process:

- Trigger classification: Labeling words with their predicted intervention type (if any). For instance, in sentence S1, the extraction system should label "food **assistance**" [2] as a trigger of an intervention type *provide food*.
- Argument classification: If a sentence contains predicted triggers $\{t_i\}$, we pair each t_i with each entity and time mention $\{m_j\}$ in the sentence to generate candidate event arguments. Given a candidate argument (t_i, m_j), the system predicts its associated role (if any). For instance, given the candidate argument ("food assistance", "Baghdad"), it predicts the role $Place$.

To perform event trigger and argument classification, we developed two convolution neural network (CNN) models: one for performing trigger extraction, and one for performing argument extraction. We show the argument model in Figure 1. These models are based on the work of Chen et al. (2015), which achieve competitive performance for event extraction.

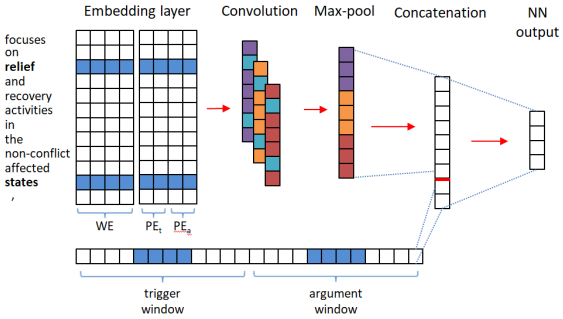

Figure 1: A CNN based model for event argument classification. WE is word embeddings. PE_t and PE_a are position embeddings, capturing a token's distance to the candidate trigger and argument respectively. These position embeddings are randomly initialized and learnt during training.

Our trigger model uses pre-trained word embeddings[3] (Baroni et al., 2014), and learns position embeddings during training (to represent relative distance of each word in the sentence to the candidate trigger). Our argument model uses these, as well as position embeddings relative to the candidate argument, and event embeddings (to represent event type of predicted candidate trigger). The position and event embeddings are randomly initialized and learnt during training.

4 Intervention Ontology and Dataset

In this section, we first present the types of interventions that we focus on, then describe a corpus annotated with intervention instances.

4.1 Intervention ontology

We focus on modeling interventions or humanitarian assistances that are meant to alleviate mass suffering, improve socioeconomic conditions, and maintain human dignity. We list our intervention ontology in Table 1. Types include promotion of *anti-retroviral* healthcare, promoting respect of *human rights*, ensuring *children friendly learning spaces*, *management of sexual violence*, *therapeutic feeding* of the severely malnourished, *vector control* of insects and pests, and *provision of various humanitarian aid* such as cash, food, etc.

[2] Our model makes predictions on single words, which we then automatically expand to phrases.

[3] EN-wform.w.5.cbow.neg10.400.subsmpl.txt.gz embeddings from the "Don't count, predict" project at http://clic.cimec.unitn.it/composes/semantic-vectors.html

Intervention Type	Example Snippet
anti-retroviral treatment	postpartum ARV *drugs* may also be given to infants
capacity building human rights	mission personnel are also engaged in building the capacity of national authorities to promote and *respect* human rights
child friendly learning spaces	promotes quality education for indigenous girls and boys through child-friendly learning *environments*
provision of goods and services	
• provide cash	cash *distributions* during emergencies
• provide delivery kit	*distributing* a home delivery kit to every pregnant woman
• provide education kit	developing and freely *distributing* education materials
• provide farming tool	the scope of the program encompasses *provision* of fertilizer
• provide fishing tool	restoration of livelihoods through *provision* of fishing boats and fishing equipment
• provide food	food *aid* is often supplied in emergency situations together with seed aid
• provide hygiene tool	respond to humanitarian emergencies always aim to *distribute* soap routinely
• provide livestock feed	where they were *provided* with fodder
• provide seed	food aid is often supplied in emergency situations together with seed *aid*
• provide veterinary service	*providing* free or subsidized animal health services
sexual violence management	health professionals expected to provide *post-rape care*
therapeutic feeding or treating	therapeutic food *provided* in supplementary feeding centers
vector control	Malathion is commonly used to *control* mosquitoes

Table 1: Types of interventions with example text snippets, where event triggers are italicized.

4.2 An intervention corpus

State-of-the-art event extraction systems adopt a supervised approach where they learn from a corpus of manually labeled examples that are specific to a predefined event ontology. For instance, the Automatic Content Extraction (ACE) (Doddington et al., 2004) corpus contains more than 500 documents manually annotated with examples for 33 event types. We similarly take a supervised learning approach by collecting and annotating examples for training extraction systems.

Humanitarian assistance programs are associated with various documentation: project proposals, guidances, progress reports, and evaluation reports on program execution. These documents are ideal for mining intervention instances. We collected several hundred documents from the following sources:

- Literature reviews, e.g., the REFANI review[4]. which reviews Cash Transfer Programmes and their impact on malnutrition in humanitarian contexts.
- Programme/project guidelines, e.g., the Sphere Handbook, which lists universal standards in core areas of humanitarian response.
- Evaluation documents, which range from thorough external evaluation of intervention operations[5], to brief presentations of results in programme documents, and post-intervention summary articles.
- Programme documents by implementers[6].
- Academic studies, such as quasi-experimental studies from ebrary.ifpri.org.

4.3 Annotating intervention instances

We provided definitions and text examples for the intervention types [7] to two annotators, and then asked them to identify and annotate intervention instances for each document. Annotators are provided with a User Interface (UI) (Chan et al., 2019) which allows them to search for examples efficiently. 30 documents are annotated by two annotators, resulting in an inter-annotator agreement of 0.83.

A total of 976 intervention instances (triggers) are found for the target intervention types. The "Count" column of Table 2 shows numbers of examples for each intervention type.

5 Experiments

In this section, we first present experiments in extracting interventions (triggers), and then describe early results on extracting locations and time for interventions.

As shown in Table 1, a large number of the intervention types (e.g. *provide cash*, *provide delivery kit*) have to do with *provision of goods and services*. Although we have kept the labeling of these interventions separate during the annotation process, so that we could optionally perform fine-grained evaluation (and indeed we will later in this section), we found that these interventions share

[4] www.actionagainsthunger.org/sites/default/files/publications/REFANI-lit-review-2015_0.pdf

[5] E.g. bmcnutr.biomedcentral.com/articles/10.1186/s40795-016-0102-6

[6] E.g. one.wfp.org/operations/current_operations/project_docs/200275.pdf

[7] A portion is shown in Table 1.

Intervention Type	Count	F1-score
anti-retroviral treatment	114	0.59
capacity building human rights	65	0.68
child friendly learning spaces	39	0.30
provision of goods and services	432	0.77
sexual violence management	131	0.45
therapeutic feeding or treating	49	0.52
vector control	146	0.67
Aggregate	976	0.68

Table 2: Intervention types with number of trigger examples and F1-scores based on 5-fold cross validation.

common trigger words (e.g. "provide", "provision", "distribute", etc.) and really rely on their arguments (e.g. "cash", "fertilizer", "fishing boats") for disambiguation. Hence, we perform two sets of trigger classification evaluation: coarse-grained and fine-grained.

5.1 Coarse-grained trigger classification

Our annotated trigger examples are spread across 240 documents. We perform 5-fold cross validation to evaluate trigger classification, over the 7 intervention types shown in Table 2. In each fold, we use 20% of the documents as test data and the remainder as training data. We performed minimal hyper-parameters tuning, using 30 epochs and batch size of 40. These were found to achieve good performance in preliminary experiments where we had further split the training data into training and development. We follow (Chen et al., 2015) for the values of the remaining hyper-parameters, e.g. CNN filter size of 3, position and event embeddings of length 5, etc. In our evaluation, a trigger is correctly classified if its intervention event type and offsets match those of a reference trigger. We show the coarse-grained trigger classification scores in the column "F1-score" of Table 2. We obtained a micro-averaged F1-score of 0.68 from the cross validation experiments.

Analysis on the decoding results show that examples vary greatly in terms of difficulty in extracting them. For example, for "sexual violence management", some triggers are phrases such as "post-rape care" that are straightforward for a classifier to recognize, if given sufficient training data. However, there is a long tail of examples where long range dependencies need to be resolved in order to type them correctly. For example, in "**counseling** ... sexual violence" and "clinical **management** ... sexual abuse", the trigger words are often more than 5 tokens far away from additional contextual clues that indicate the target type. We leave modeling these as future work.

5.2 Fine-grained trigger classification

As mentioned earlier in this section, we propose that the intervention type *provision of goods and services* rely on their event arguments (artifacts involved) for disambiguation into the finer-grained interventions listed in Table 1. To enable this, we first need to detect mentions of different goods/services in text. We adopt a simple list-based approach, where we manually compiled lists [8] of descriptors for each category. For instance, we use the descriptors ("livestock feed", "fodder", "hay", etc.) for the category *livestock feed*.

Then, when we note that our coarse-grained trigger model had predicted a trigger instance of *provision of goods and services* in a sentence, we check the trigger's surrounding context (5 token window) for mentions of *livestock feed*, *farming tool*, *fishing tool*, etc. We thus deterministically relabel *provision of goods and services* into the appropriate finer-grained intervention type, depending on which category of descriptor is present in the trigger's context window. As shown in Table 2, the coarse-grained F1-score of *provision of goods and services* is 0.77. After performing the deterministic re-labeling into finer-grained intervention types, we obtain an aggregate F1-score of 0.56 when evaluating against our fine-grained trigger labels. Recall misses such as those resulting from incomplete descriptor lists, and precision misses resulting from multiple descriptor categories being present within a trigger's surrounding context, contributed to the drop in F1-score.

5.3 Extracting locations and time

We leverage the ACE corpus, which contains annotations of *Place* and *Time* event arguments, to train an event type independent *Place/Time* argument classifier, based on the neural architecture described in Section 3. In our evaluation, an argument is correctly classified if its event type, event argument role, and offsets match any of the reference event arguments.

Africa countries are often the focus sites of humanitarian programs and agencies, such as the World Food Programme (WFP). Hence, to evaluate the performance of our argument model for intervention events, we randomly selected 250 documents from around 6,000 documents collected from `allafrica.com`.

[8] These lists are available at `https://github.com/BBN-E/mr-intervention`

We first apply our coarse-grained trigger classifier on these documents. We then ask annotators to evaluate trigger predictions and retain only correct ones (188 triggers), which we subsequently use to evaluate our argument classifier. We focus on using correct triggers to evaluate argument classification, to avoid error propagation (from erroneous trigger predictions) from muddling a fair assessment of the argument classifier.

Our annotators assigned a total of 15 *Time* arguments and 77 Place arguments to the 188 event triggers. Our argument classifier predicted a total of 12 Time arguments, giving a precision, recall, and F1 of 0.92, 0.73, and 0.81 respectively. Our argument classifier predicted 30 *Place* arguments, giving a precision, recall, and F1 of 0.93, 0.36, and 0.52 respectively.

6 Conclusion and Future Work

In this paper, we introduced a new corpus annotated with intervention events, and presented a system that achieves encouraging results.

Our next step is to annotate more documents and make them available to the research community to foster research in this area. We also plan to add *Location* and *Time* annotation on top of the intervention annotation.

Acknowledgements

The authors would like to thank Ben Watkins and colleagues at Kimetrica LLC for the inspiring discussions on interventions, and providing initial definitions of intervention types in Table 1. The authors also would like to thank the anonymous reviewers for their insightful comments, which helped us to improve the final version of the paper.

This work was supported by DARPA/I2O and U.S. Army Research Office Contract No. W911NF-18-C-0003 under the World Modelers program. The views, opinions, and/or findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the Department of Defense or the U.S. Government. This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

References

David Ahn. 2006. The stages of event extraction. In *ARTE*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Marco Baroni, Georgiana Dinu, and German Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL-2014*, pages 238–247, Baltimore, USA.

Yee Seng Chan, Joshua Fasching, Haoling Qiu, and Bonan Min. 2019. Rapid customization for event extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 31–36, Florence, Italy. Association for Computational Linguistics.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *ACL-IJCNLP2-2015*, pages 167–176, Beijing, China. Association for Computational Linguistics.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel. 2004. The automatic content extraction (ace) program - tasks, data, and evaluation. In *LREC*. European Language Resources Association.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, volume 1.

Ruihong Huang and Ellen Riloff. 2012. Modeling textual cohesion for event extraction. In *AAAI-CAI*, AAAI'12, pages 1664–1670. AAAI Press.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *ACL-HLT-2008*, pages 254–262, Columbus, Ohio. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC bioinformatics*, 9(1):10.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *NAACL-HLT-2016*, pages 300–309, San Diego, California. Association for Computational Linguistics.

Stephanie M Strassel, Ann Bies, and Jennifer Tracey. 2017. Situational awareness for low resource languages: the lorelei situation frame annotation task. In *SMERP@ ECIR*, pages 32–41.