Information Retrieval for Humanitarian Crises via a Semantically Classified Word Embedding

Aladdin Shamoug
University of Otago
aladdin.shamoug@postgrad.otago.ac.nz

Stephen Cranefield
University of Otago
stephen.cranefield@otago.ac.nz

Grant Dick
University of Otago
grant.dick@otago.ac.nz

ABSTRACT

Decision-makers in humanitarian crises need information to guide them in making critical decisions. Finding information in such environments is a challenging task. Therefore, decision-makers rely on domain experts who possess experience and knowledge from previous humanitarian crises to provide them with the information they need. In this paper, we explore the ability of the existing computing technologies to augment the capabilities of those experts and help decision-makers to make faster and better decisions. In this paper, we train a word embedding model using word2vec, transform words from news archive to entities in domain ontology, annotate them with their equivalent concepts from upper ontologies, and reason about them using semantic similarity and semantic matching, to represent and retrieve knowledge, and answer questions of interest to decision-makers in humanitarian crises. The approach was evaluated by comparing the use of word embeddings with and without semantic classification for the retrieval of information about the current humanitarian crisis in Syria.

Keywords

Ontologies, Word Embedding, Information Retrieval, Humanitarian Crisis, Humanitarian Response.

INTRODUCTION

In response to emerging humanitarian crises, humanitarian organisations use their expert knowledge to make decisions on the challenges they encounter in their day-to-day work. This expert knowledge is accumulated knowledge carried-over from previous crises by domain experts who are considered the guardians of humanitarian knowledge. Domain experts have various job descriptions and functional titles, yet all of them are hired and retained for one reason: to provide information to support decision-makers. Domain experts help humanitarian organisations in providing answers to critical questions, supporting the decision-making process, and sustaining the institutional memory of those organisations (Clark et al. 2015; Widener et al. 2017).

Domain experts collect, integrate and reason about historical humanitarian information and records using their knowledge and experience to inform the decision-making process in response to emerging crises. Domain experts must have three essential qualities to be able to answer questions and provide information for decision-makers: analytical skills, solid knowledge, and practical experience (ACAPS 2017). However, this approach of using human reasoning, knowledge, and experience is time and resource intensive.

We believe that existing computing technologies can augment the role of domain experts and help decision-makers obtain better, more accurate, and higher quality information. This raises the question: how can we automate knowledge retrieval in humanitarian contexts to augment domain experts' capabilities? To answer such a question, we need to explore the ability of existing technologies to help decision-makers in (a) transforming and representing the collected information in a machine-friendly format, and (b) developing a computing model to reason about that information and provide answers needed by decision-makers. Addressing these two issues is the motivation of this research.

The contribution of this paper is a technique to represent, classify, and retrieve information from historical data to answer a range of information-seeking questions of interest to decision makers in humanitarian crises. This technique transforms an archived text corpus into: 1) a vector space allowing semantic similarities between words to be discovered, and 2) a domain ontology where the meanings and classes of a subset of the words are harvested from upper ontologies and stored in a locally stored ontology. We combine these two models to find similarities in the vector space between query words and those from corpus documents, and to filter the results using semantic matching. This results in increased relevance of query results compared to the use of similarity in the vector space alone.

The remainder of this paper is organised as follows: in the next section we review related work. Then explore the existing problems in the current situation and identify the research questions in section 3. In section 4 we present a real-life scenario to demonstrate the impact of humanitarian crises. Then, in section 5 we propose an approach to address the research problems. In section 6 we demonstrate a solution to address the problems identified in section 3, and then discuss the implementation and evaluation of the solution in section 7. Finally, in section 8, we review the work presented in this paper and discuss the future steps.

RELATED WORK

Using word embedding in semantic web domain has been attracting the attention of the research community in recent years. Word embedding has been explored in many studies, such as matching ontological concepts from different ontologies based on vector distances (Zhang et al. 2014), enhancing search results by adding semantics to similarities in word embedding (Maas and Ng 2010), embedding similarities and semantics to word image recognition and retrieval (Gordo et al. 2015), aligning documents to domain ontologies based on semantic similarity from word2vec (Albukhitan et al. 2018), and enhancing conversational dialogue systems using knowledge graphs and word embedding to enrich queries (Celikyilmaz et al. 2015; Chen and Rudnicky 2014).

Word embedding is a Natural Language Processing (NLP) technique that represent words as vectors in a metric space. Using word co-occurrence measures to generate a word embedding can result in embeddings in which semantically similar words map to vectors that are close to each other (Roy et al. 2017).

Among many word embedding techniques, the word2vec is one of the most prominent models (Albukhitan et al. 2018; Othman et al. 2017; Pilehvar and Collier 2016; Zhang et al. 2014). Word2vec has two techniques to train a word vector embedding: 1) the continuous-bag-of-words (CBOW), which learns the embedding through maximising the ability of each word to be predicted from its set of context words using vector similarity, and 2) the skip-gram with negative sampling (SGNS), which works in the opposite direction of CBOW by maximising the ability of target words to predict context words (Mikolov et al. 2013).

Despite being an influential model, word2vec has its limitations when it comes to relational and categorical similarities (Xu et al. 2014) and ordinal semantic knowledge (Liu et al. 2015). For instance, although a well-trained word2vec model is able to find similarity between “London” and “Paris”, it often fails to find similarity between “London” and its relational and categorical attributes, such as “City” or “Capital”.

Liu et al. (2015) note that word2vec has three fundamental limitations when it comes to inequality in semantic ordinal: synonym antonym, semantic category, and semantic hierarchy. They propose a framework to enrich word embedding using semantic rules to gather ordinal semantic knowledge from lexical knowledge resources and prevent ordinal inequality. Having similar rules in our model might improve similarities as well, given that similarities in our model depend on cosine distances only, where the external knowledge in our model is not used to improve similarities but rather used to improve relevancy of results.

To address limitations in word2vec, Xu et al. (2014) propose a model to improve similarity in word embedding by adding relational and categorical knowledge to word embeddings. This model addresses the lack of semantics in word2vec by embedding knowledge from knowledge graphs and upper ontologies into word embedding and improve information retrieval by enriching entities using relational and categorical attributes.

The findings in (Ling et al. 2017), come very close to (Xu et al. 2014) in terms of using upper ontologies to improve search results in word embeddings. However, the way each exploited knowledge embedding is different. In contrast to our work, both Ling et al. (2017) and Xu et al. (2014) focused on improving similarity between entities in word embeddings, while our method focuses on the relevance of results. Semantics from upper ontologies remain distinctive in our model and only used in classifying entities to allow extraction of similar concepts based on their categorical relations to concepts in upper ontologies.

Similarly, Roy et al. (2017) propose a novel approach to enrich word embedding by inserting domain-specific terms as text annotation during model training. The problem with domain-specific terms is that they usually have lower similarity rates – due to their sparsity and scarcity in text corpora – despite their importance to the

domain. The proposed model enhances search results, but, again, once the model is trained, the domain knowledge is blended in and those terms are treated as part of the word embedding space. Therefore, it is not possible to locate terms and extract results based on their semantics.

In the work of Yao et al. (2018) word embedding has been used to discover the changing semantics of words over time. Lack of time-awareness is one of the challenges we encountered in our work, where real-life concept classifications, keep changing over time and so their semantic annotation do as well. We presume that using dynamic word embedding for evolving semantic discovery might address this issue in our future works.

The work of Othman et al. (2017) uses the Yahoo QA archive to train a word2vec model. The aim of their experiment is to find semantically similar questions and use this similarity to find related answers from QA database, where each question in QA database has associated answer(s) in the same database. Using such techniques to answer questions might not be feasible in the humanitarian domain due to lack of such large-scale QA databases, though it could be feasible to explore alternative applications of this method in our future work.

Nevertheless, word embedding is one among many approaches used in humanitarian knowledge representation and retrieval. For instance, the work of Malizia et al. (2010) propose a different approach for knowledge representation using an ontology-based solution to enhance knowledge representation, sharing and communication that integrates information from emergency notification systems during crisis situations.

Meanwhile, Imran et al. (2015) conducted a study to explore and review existing solutions that use information from social media to improve situation awareness, event detection and semantic enrichment during response to large-scale natural crises. Similarly, Yin et al. (2012) uses various data mining techniques to classify, cluster, geotag, and visualise information captured from social media during natural disasters in order to enhance emergency situation awareness.

PROBLEM STATEMENT

According to the United Nations, the world experienced 2,911 humanitarian crises since 1981, with 49 of them are currently ongoing (ReliefWeb 2018). This volume of crises produces a huge amount of data, which are neither integrated nor related to each other. Most of the information related to current and previous crises is widely available online in human-friendly formats i.e. tabular, graphical, and raw texts (Widener et al. 2017).

In such environments, decision-makers must make critical decisions that impact the lives of millions of people in light of limited information, and under the pressure of limited time and resources. Those decision-makers need reliable information, which is usually provided by domain experts who: 1) have accumulated knowledge and experience from previous crises, and 2) are hired by humanitarian organisations to provide such information to support the decision-making process (ACAPS 2017).

To support decision making in a newly emerging crisis, domain experts collect data from scattered data sources, and use their prior knowledge and practical experience to reason about such data and provide information and answers to support decision-making. Dependence on human memory and reasoning skills, in absence of knowledge representation and retrieval tools, might affect the decision-making process and costs humanitarian organisations precious time and resources during the response to crises (Widener et al. 2017). Reliance on human capacities to preserve and retrieve knowledge in humanitarian crises has three disadvantages:

- Slow response: a reliance on ad-hoc data collection usually consumes a lot of time in searching for reliable sources, extracting and analysing data, and providing answers. To save this time, it might be better to transform such information into a format that allows machines to retrieve, reason about, and retrieve information in accordance with the questions asked by decision-makers.
- High cost: also it is expensive to hire domain experts to reason about collected data in every crisis. The cost of hiring and retaining such experts is expensive in comparison to intelligent computing models, which can augment decision-making through reasoning about millions of records, texts, tweets, websites, databases, and blogs in a fraction of the time and cost of the domain experts.
- Low quality: the urgent nature of such requests does not provide data collectors enough time for quality assurance. Domain experts are forced to find relevant data in the shortest possible time, and sacrifice quality in favour of speed, i.e. to choose fast and low-quality results over slow and high-quality ones.

We presume that collecting, transforming, integrating, and reasoning about data generated from previous crises will generate adequate knowledge to help decision-makers in making better decisions in emerging crises, in light of answers derived from the knowledge that can be semantically inferred from public data sources.

In this paper, we consider the following research question: how can we combine semantic information with free

text from open sources to build a model that can help decision-makers to answer critical questions in humanitarian crises?

APPLICATION SCENARIO

The ongoing humanitarian crisis in Syria has been regarded as the most violent, ruthless, and devastating crisis since World War II. The cost of the Syrian crisis has exceeded $226 billion US dollars, and has immense human impact: 27% of the housing stock are impacted, 53% of education facilities are damaged, almost half of the population are forcibly displaced, 4.9 million of them live in refugee camps, and more than 400,000 are estimated to have died (World Bank 2017).

To reduce the impact of this tragedy, the international community spent billions of US dollars on humanitarian assistance, most of which went to life-saving interventions (such as food distribution, water supplies, healthcare, hygiene and sanitation, shelter, and nutrition) led by different agencies of the United Nations, and other International Organisations, to save and protect the lives of Syrian people (OCHA 2017). Such a huge amount of humanitarian assistance cannot be delivered to affected areas without having enough background information.

It is likely that in the past seven years, a vast number of "small" questions have been asked on a daily basis, such as: In what countries do Syrian refugees seek asylum? Which organisations provide medical aid to Syrian people? What happened in Eastern Aleppo in 2016? Where did chemical attacks take place in Syria? And what does the United Nations do in Syria? Such questions have been asked by decision-makers in the United Nations, NGOs, media, governments, donor countries and agencies, grassroots organisations, affected populations, and many more entities. The answers to such questions lie somewhere in "abandoned" media reports, tweets, blogs, datasets, websites, and other data sources. The main problem is that we don't currently possess a proper computer model to collect relevant data from those sources, integrate and model them in a machine-friendly format, and reason about them to produce new knowledge and provide answers to decision-makers.

THE APPROACH

To answer the research question, we need to explore existing technologies to collect related information from open data sources, integrate and align them together in a unified format, and semantically reason about them to retrieve information and provide answers for questions that are often asked by decision-makers in crisis contexts. Figure 1 shows how different technologies (such as word embedding, data transformation, semantic classification, and semantic matching) are utilised in this work to build a framework that can augment the role of domain experts in information retrieval problems:

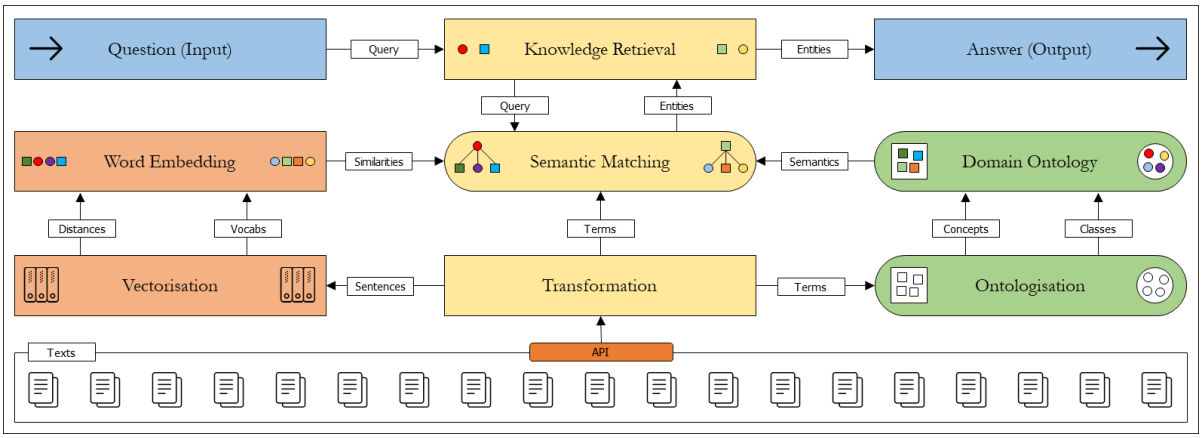

Figure 1. Our Framework

Data Collection

Data can be collected from different data sources of heterogeneous nature. These data sources can be found in structured or unstructured format. Given that most humanitarian data are produced for the consumption of humans rather than machines, we focus in this paper on the second format: unstructured data sources, which usually come as raw text from social media, news portals, blogs, and documents. To harvest content from such sources, we may use an application programming interface (API). In absence of dedicated APIs, then text mining and web scraping techniques can be applied (Atanasova et al. 2010) in order to collect and locally store

such data for further processing.

Data Transformation

For data transformation, we use natural language processing to clean and normalise collected data and transform them from human-readable to machine-readable format. Data transformation starts with text cleaning, where noise and errors are eliminated from collected data to make them ready for normalisation. Text cleaning involves the following steps: expressive lengthening, emoticon handling, HTML tags removal, slang handling, punctuation handling, and stopwords removal (Mhatre et al. 2017). In data normalisation, we tokenise and tag the textual contents, and transform them into unique terms. Tokenisation aims to break down the text into words and punctuation, while tagging is the process of classifying words into their parts-of-speech and labelling them accordingly (Bird et al. 2009).

Knowledge Representation

In our approach we represent knowledge using two techniques: vectorisation and ontologisation.

- For vectorisation, we use the CBOW word2vec approach (Mikolov et al. 2013) to "encode continuous similarities between words as distance or angle between word vectors in a high-dimensional space" (Maas and Ng 2010 p. 1). We train a word embedding model in which we have all terms used in the text corpus transformed into vectors, where each vector represent the exact location, i.e. coordinates of each term in a vector space.
- For ontologisation, we construct the ontology and annotate the concepts as explained below:
  - Ontology construction aims to convert transformed concepts into an ontology, which is a set of "basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary" (Neches et al. 1991 p. 40). For ontology construction we use the Web Ontology Language (OWL), which is "the basis for [a] machine to understand resources, define the concepts and relations of things, and describes knowledge with useful keywords and structures in a both human and machine-understandable way" (Huang and Dai 2017 p. 6671).
  - Concept annotation aims to map the normalised terms to the most relevant concepts in upper ontologies and ensure the entities asserted into domain ontology are aligned to equivalent concepts in upper ontologies. Upper ontologies are defined as vocabularies for top-level generic concepts that "serve as common ground for communication across different domains" (Grabarske and Heutelbeck 2012 p. 1128).

Having the text corpus vectorised, ontologised and semantically annotated into upper ontologies allows concepts in the text corpus to benefit from similarity measurement and embedded semantics to reason about them and produce answers to the end-user questions, as explained in the next section.

Knowledge Retrieval

Knowledge retrieval is the last step in our approach that aims to extract information from data repositories in term of semantic relations. We use acquired knowledge from previous steps to retrieve knowledge that helps to answer the questions asked by decision-makers. For knowledge retrieval, we use similarity measurement and semantic matching to produce the required results and answers.

For results extraction, we measure semantic similarities between concepts using cosine distance in the vector embedding, which helps in discovering the hidden similarities between different concepts in big datasets and text corpora. For results optimisation, we use semantic matching: a process of improving search results through determining the matching degree between a given question and possible answers (Li and Xu 2014). We use semantic matching in our model to classify results obtained through similarity measurement and make them more informative, accurate, inclusive, and relevant (Ling et al. 2017).

THE SOLUTION

The solution illustrated in this section aims to: answer the research question, address the above-mentioned scenario, and examine the validity of the approach. It has been divided into two interlinked modules:

- *The knowledge representation module*, in which we extract terms from a text corpus, semantically

annotate them using upper ontologies, locally store them in a Domain Ontology (DO_ONT), embed the terms in a vector space using word2vec (W2V), and then merge the results into a semantically classified word embedding model (SC-WE). Knowledge representation entails the following processes:

- Extracting concepts and classes (from DO_ONT) and vectors (from W2V).
- Storing concepts (from DO_ONT) and terms (from W2V) as entities (in SC-WE).
- Annotating classes (from DO_ONT) and aligning vectors (from W2V) to entities (in SC-WE).

❖ *The knowledge retrieval module* uses classes and vectors to retrieve entities from SC-WE. In this module, end-users enter keywords (e.g. "provide medical assistance to affected population") and the class of results they are looking for (e.g. organisation). This module performs the following processes:

- Measuring the similarities between keywords (from user query) and entities (in SC-WE) using cosine distance between keywords and entities.
- Matching the semantics of classes (from user query) with classes of similar entities (SE).

Knowledge Representation Module

We have transformed our data into: 1) a Domain Ontology (DO_ONT), and 2) a Word2Vec Model (W2V), where DO_ONT hosts the extracted terms from data sources and upper ontologies, stores them as concepts using OWL language, and annotates them with equivalent concepts from upper ontologies. Meanwhile, W2V hosts the same set of terms as vectors in a word embedding space. Figure 2 shows how information from the text corpus has been vectorised, ontologised and transformed into the knowledge representative format.

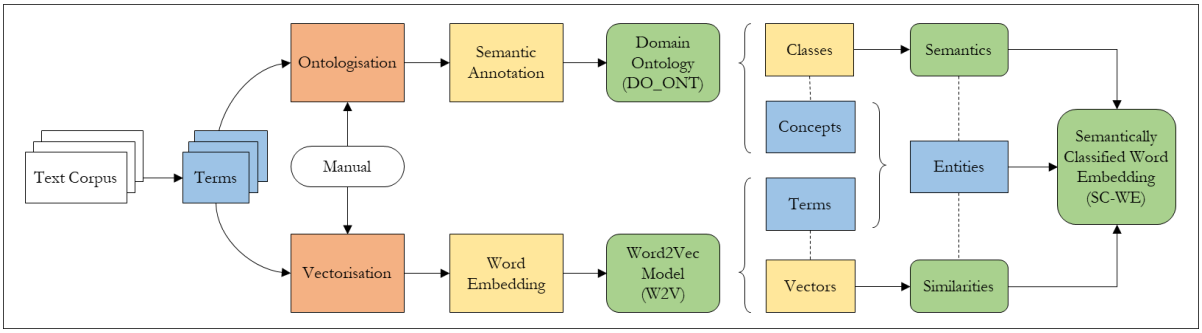

Figure 2. Knowledge Representation Module

Word Embedding

Word embedding has been attracting increasing attention from researchers in both neural language processing and semantic web fields. Using word embedding to transform large text corpora into vector spaces has allowed researchers to find hidden similarities between words, phrases, sentences, and documents.

In this paper, we used the Guardian Open Platform (GOP) API [2] to collect published materials about the protracted crisis in Syria. We ended up with a text corpus that contains 7,053 documents, covering the period from the first day of the crisis i.e. 15/03/2011 until 14/03/2018, in which seven years of news, reports, op-eds, materials, articles, and opinions have been collected, cleaned, collated and locally stored.

The corpus obtained from GOP has 6,627,078 words, which have been used to train a word embedding space through the continuous-bag-of-words (CBOW) technique using the Gensim library (Řehůřek and Sojka 2010). The result was a vector space of 94,104 terms $\{T_1, \ldots, T_{94,104}\}$. Each term has associated coordinates in a vector space with 200 dimensions[1] $\{D_1, \ldots, D_{200}\}$. These coordinates represent the location of the terms in the space, as shown in Figure 3.

1) It has been recommended that word embeddings should have between 100 to 300 dimensions (Albukhitan et al. 2018; Celikyilmaz et al. 2015; Chen and Rudnicky 2014; Liu et al. 2015; Othman et al. 2017; Xu et al. 2014).

Semantic Annotation

The same corpus has been used to annotate terms using semantics collected from upper ontologies. We used three upper ontologies (DBpedia, Freebase, and WordNet) and an ontology construction tool (OwlReady) to transform, classify, ontologise, and store semantic concepts extracted from the text corpus. The purpose of this process is to semantically represent the terms extracted from GOP and semantically annotate them with their equivalent concepts in upper ontologies. We stored the harvested semantics, from upper ontologies, in DO_ON.

Semantic Classification

In the semantic classification, we created the SC-WE model by transforming terms (from W2V) and concepts (from DO_ONT) into entities (in SC-WE). Next, we used vectors (from W2V) and semantics (from DO_ONT) to classify entities in the SC-WE model. Semantic classification adds semantics to entities in SC-WE and categorises them in classes, such as organisations, locations, actions, sectors, and persons, as shown in Figure 3.

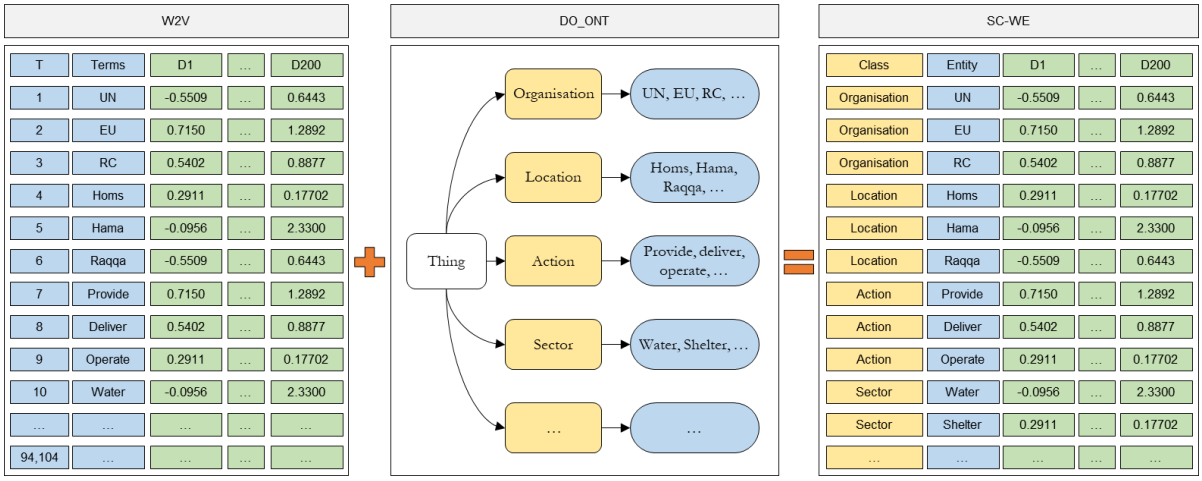

Figure 3. Semantically Classified Word Embedding (SC-WE) Model

Out of 94,104 unique terms in W2V we managed to annotate 19,387 of them with entities in SC-WE using DO_ONT. The remainder of the terms (74,717) were either generic terms of no significant importance or expressions that cannot be classified under DO_ONT.

Knowledge Retrieval Module

Searching for similarities in the word2vec model needs a rich and accurate selection of keywords (KW), where the more precise the keywords are, the more accurate are the results obtained. Using only keywords to find similarities in SC-WE yields generic results, while users are usually looking for a specific type of results, such as organisations responsible for an action, locations where events happen, actions taken by certain actors, or persons in charge of some actions. To address this issue, the class (CL) of the result desired, must be identified.

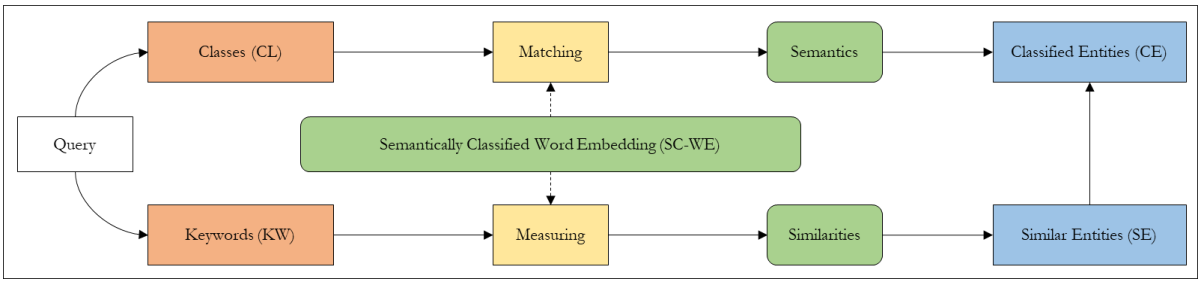

Figure 4. Knowledge Retrieval Module

Similarity Measurement

To measure similarity between any given set of keywords and entities in SC-WE we used cosine distance to measure similarities between keywords and entities, then sorted all the 94,104 entities in SC-WE, based on their

distance from the keywords, from most-similar (shortest cosine distance) on top of the list, to least-similar (longest cosine distance) at the bottom. The results are stored as "similar entities" (SE) (Figure 4).

Semantic Matching

In semantic matching, we use classes to filter SE and only keep those which have semantics matching with classes. For example, if the end-user chose "Organisation" as a class, we eliminate all entities from SE which are not labelled as "Organisation" and keep only those that are considered organisations in SC-WE. The results of the matching process are stored as "classified entities" (CE) (Figure 4, above).

Results Retrieval

We developed a query language for SC-WE, we call it "Semantically Classified Word Embedding Query Language" (ScWeQL), to express our information retrieval requests. ScWeQL queries consist of three clauses: Using, Find, and For. The Using clause is used to select the source of semantics and classes, i.e. the upper ontology (UO); the end-user can choose any upper ontology among the three upper ontologies we linked DO_ONT classes to. The Find clause is used to select the class of results (CL) desired by the user. In the For clause we inform SC-WE about the keywords (KW) we are looking for. Figure 5 shows the syntax of ScWeQL.

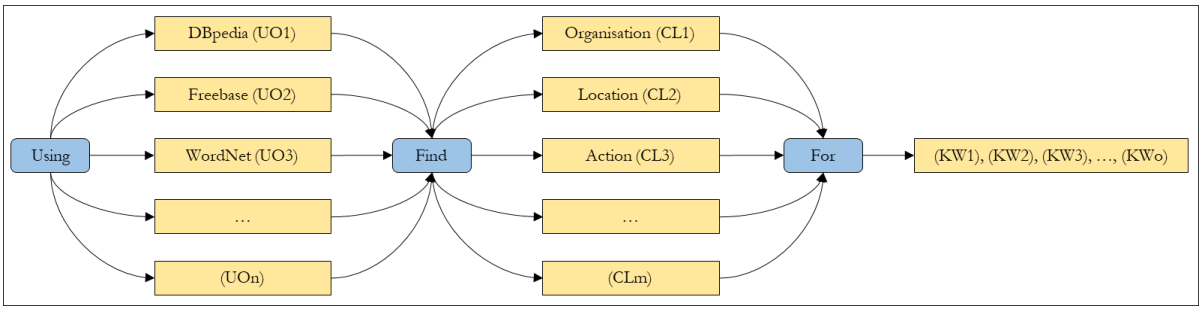

Figure 5. Semantically Classified Word Embedding Query Language (ScWeQL)

We implemented a query processor that performs the above-described processes (similarity measurement and semantic matching). This processor measures similarity, semantically match entitles, picks and delivers the top results from SE and CE (i.e. results with highest similarities and matching semantic attributes) to the end-user. ScWeQL queries are solely designed to extract results from SC-WE.

IMPLEMENTATION AND EVALUATION

Implementation

To test our approach, we used ScWeQL to ask the five questions in Table 1. The keywords have been extracted from those questions and the class has been assigned to each one of them:

Table 1. Examples of Questions Asked by Decision-makers in Humanitarian Crises

	Question	Keywords (KW)	Classes (CL)
1	What countries do Syrian refugees seek asylum in?	Countries, Syrian, Refugees, Seek, Asylum	Location
2	Which organisations provide medical aid to Syrian people?	Organisation, Provide, Medical Aid, Syrian	Organisation
3	What happened in Eastern Aleppo in 2016?	Happened, Eastern Aleppo, 2016	Action
4	Where did chemical attacks take place in Syria?	Chemical, Attack, Took Place, Syria	Location
5	What does the United Nations do in Syria?	United Nations, Doing, Syria	Action

We used the keywords and classes identified in Table 1 to conduct five tasks and get results from SC-WE using ScWeQL. Below, for each task, we show the result entities (both similar and classified), and similarity between keywords and each result entity. The similarity varies between 0 and 1, where 1 means the resulting entity is

identical to the keywords, while 0 means no similarity at all between the resulting entity and the keywords. To estimate the weight of each result, we calculated the frequency of the terms in the original text corpus.

Task 1: Asylum Seekers

We consider the query: What countries do Syrian refugees seek asylum in? The following keywords have been extracted from the question: Countries, Syrian, Refugees, Seek, and Asylum; and the following class has been assigned to the query: Location. ScWeQL has been used to reason about entities in SC-WE twice: in the first part we used only keywords to find similarities using ScWeQL statement to retrieve results:

```
ScWeQL >>> Using (DBpedia) Find (*) For (Countries, Syrian, Refugees, Seek, Asylum) Limit (5)
```

ScWeQL returned the five most similar entities (SE): Syrians, Syrian Refugees, Asylum Seekers, Migrants, and Seekers. We now add the class to classify the findings using the following ScWeQL query:

```
ScWeQL >>> Using (DBpedia) Find (Location) For (Countries, Syrian, Refugees, Seek, Asylum) Limit (5)
```

This time, ScWeQL returned the five most similar classified entities (CE): Greece, Germany, Turkey, Europe, and Italy. The results of both queries in this task are in Table 2.

Table 2. Results of Query 1

SE (KW)	Similarity	Frequency	CE (KW + CL)	Similarity	Frequency
Syrians	0.64	3661	Greece	0.55	815
Syrian Refugees	0.60	1715	Germany	0.46	1403
Asylum Seekers	0.58	493	Turkey	0.45	7358
Migrants	0.57	940	Europe	0.45	3485
Seekers	0.56	8	Italy	0.44	490

The results of this task show that Greece, Germany, Turkey and Italy are on the top destinations for Syrian refugees and asylum seekers (Amnesty International 2018).

Task 2: Medical Aid Providers

We consider the query: Which organisations provide medical aid to Syrian people? The following keywords have been extracted from the question: Organisation, Provide, Medical Aid, and Syrian; and the following class has been assigned to the query: Organisation. ScWeQL has been used to reason about entities in SC-WE twice: in the first part we used only keywords to find similarities using the following ScWeQL query:

```
ScWeQL >>> Using (DBpedia) Find (*) For (Organisation, Provide, Medical Aid, Syrian) Limit (5)
```

ScWeQL returned the five most similar entities (SE): Providing Humanitarian, Communications Equipment, Humanitarian Organisations, Relief Organisations, and Child Protection. We now add the class to classify the findings using the following ScWeQL query:

```
ScWeQL >>> Using (DBpedia) Find (Organisation) For (Organisation, Provide, Medical Aid, Syrian) Limit (5)
```

This time, ScWeQL returned the five most similar classified entities (CE): SARC, UNRWA, Red Cross, Works Agency, and International Committee. The results of both queries in this task are in Table 3 below:

Table 3. Results of Query 2

SE (KW)	Similarity	Frequency	CE (KW + CL)	Similarity	Frequency
Providing Humanitarian	0.58	44	SARC[1]	0.50	51
Communications Equipment	0.56	21	UNRWA[2]	0.50	109
Humanitarian Organisations	0.56	83	Red Cross	0.49	352
Relief Organisations	0.56	16	Works Agency	0.49	25
Child Protection	0.56	31	International Committee[3]	0.48	149

1) Syrian Arab Red Crescent

2) The United Nations Relief and Works Agency for Palestine Refugees in the Near East

3) International Committee of the Red Cross (ICRC)

The results of this task show that SARC, UNRWA, Red Cross, and ICRC were providing medical aid to affected population in Syria, which aligns with what is observed in the Syrian crisis (OCHA 2017).

Task 3: Eastern Aleppo in 2016

We consider the query: What happened in Eastern Aleppo in 2016? The following keywords have been extracted from the question: Happened, Eastern Aleppo, and 2016; and the following class has been assigned to the query: Action. ScWeQL has been used to reason about entities in SC-WE twice: in the first part we used only keywords to find similarities using the following ScWeQL query to retrieve results:

```
ScWeQL >>> Using (DBpedia) Find (*) For (Happened, Eastern Aleppo, 2016) Limit (5)
```

ScWeQL returned the five most similar entities (SE): This Year, Eastern Ghouta, East Aleppo, Next Year, and March 2011. We now add the class to classify the findings using the following ScWeQL query:

```
ScWeQL >>> Using (DBpedia) Find (Action) For (Happened, Eastern Aleppo, 2016) Limit (5)
```

This time, ScWeQL returned the five most similar classified entities (CE): Clashes Erupted, Heavy Fighting, Reconvening, Be Replicated, and Heavy Shelling. The results of both queries in this task are in Table 4 below:

Table 4. Results of Query 3

SE (KW)	Similarity	Frequency	CE (KW + CL)	Similarity	Frequency
This Year	0.58	996	Clashes Erupted	0.50	12
Eastern Ghouta	0.57	249	Heavy Fighting	0.49	86
East Aleppo	0.57	237	Reconvening	0.49	3
Next Year	0.57	210	Be Replicated	0.47	11
March 2011	0.55	216	Heavy Shelling	0.46	48

The results of this task show that actions in CE were happening in Eastern Aleppo in 2016 (Humud et al. 2018).

Task 4: Chemical Attacks

We consider the query: Where did chemical attacks take place in Syria? The following keywords have been extracted from the question: Chemical, Attack, Took Place, and Syria; and the following class has been assigned to the query: Location. ScWeQL has been used to reason about entities in SC-WE twice: in the first part we used only keywords to find similarities using the following ScWeQL query to retrieve results:

```
ScWeQL >>> Using (DBpedia) Find (*) For (Chemical, Attack, Took Place, Syria) Limit (5)
```

ScWeQL returned the five most similar entities (SE): 21 August, Chemical Attack, Gas Attack, Chlorine, and Civilian Areas. We now add the class to classify the findings using the following ScWeQL query:

```
ScWeQL >>> Using (DBpedia) Find (Location) For (Chemical, Attack, Took Place, Syria) Limit (5)
```

This time, ScWeQL returned the five most similar classified entities (CE): Ghouta, Houla, Eastern Ghouta, Khan Sheikhun, and Damascus Suburbs. The results of both queries in this task are in Table 4 below:

Table 5. Results of Query 4

SE (KW)	Similarity	Frequency	CE (KW + CL)	Similarity	Frequency
21 August	0.65	266	Ghouta[1]	0.56	283
Chemical Attack	0.64	437	Houla[2]	0.53	175
Gas Attack	0.59	74	Eastern Ghouta	0.53	249
Chlorine	0.56	143	Khan Sheikhun[3]	0.52	120
Civilian Areas	0.56	127	Damascus Suburbs	0.51	103

The results of this task align with known evidence that that all the locations in CE except Houla were theatres of chemical attacks in the past seven years (Smith and Brooke-Holland 2018).

1) The countryside and suburban area surrounds the city of Damascus.

2) An area consisting of three villages in the Homs Governorate of central Syria.

3) A town in southern Idlib Governorate of northwestern Syria.

Task 5: United Nations Interventions

We consider the query: What does the United Nations do in Syria? The following keywords have been extracted from the question: United Nations, Doing, and Syria; and the following class has been assigned to the query: Action. ScWeQL has been used to reason about entities in SC-WE twice: in the first part we used only keywords to find similarities using the following ScWeQL query to retrieve results:

```
ScWeQL >>> Using (DBpedia) Find (*) For (United Nations, Doing, Syria) Limit (5)
```

ScWeQL returned the five most similar entities (SE): Inside Syria, Humanitarian, Humanitarian Aid, Facilitating, and European Union. We now add the class to classify the findings using the following query:

```
ScWeQL >>> Using (DBpedia) Find (Action) For (United Nations, Doing, Syria) Limit (5)
```

This time, ScWeQL returned the five most similar classified entities (CE): Facilitating, Securing, Coordinating, Establishing, and Ongoing. The results of both queries in this task are in Table 6 below:

Table 6. Results of Query 5

SE (KW)	Similarity	Frequency	CE (KW + CL)	Similarity	Frequency
Inside Syria	0.42	753	Facilitating	0.38	53
Humanitarian	0.39	1813	Securing	0.35	214
Humanitarian Aid	0.39	496	Coordinating	0.35	97
Facilitating	0.38	53	Establishing	0.35	209
European Union	0.37	368	Ongoing	0.35	527

Again, the results of this task align with actions, such as facilitating, securing, coordinating, establishing, and ongoing that are associated with the United Nations in Syria (OCHA 2017).

Evaluation

Our solution addresses the two motivations of this study using a semantically classified word embedding model to augment a domain expert's role in humanitarian crises through collecting, transforming, reasoning about historical information to provide information to decision-makers. The implementation phase shows that our semantically classified model is not only able to retrieve similar concepts but also provides correct and relevant answers as well. Out of the 25 answers (classified entities), obtained in the five tasks, there were no wrong or irrelevant answers returned, except the second result in the fourth task "Houla", where there was no chemical attack reported there in the past seven years. Those answers have been cross-checked with external sources (Amnesty International 2018; Humud et al. 2018; OCHA 2017; Smith and Brooke-Holland 2018) and found that they are highly relevant.

The five tasks show that our results were retrieved based on their similarity to keywords and matching with classes, while the frequency of words is not a dominant factor in retrieving those results. Our approach in this study is different from other approaches, which have been reviewed in the related work section. It worth mentioning that we did not use any pre-trained model in our work but rather trained our model using a text corpus that we extracted from news archive. The methodology introduced in this paper can be applied to many other fields such as healthcare, education, agriculture, security, and many more.

On the other hand, our work has room for future enhancements. The text corpus used for model training is a single-source, aligned to The Guardian editorial policy and perspective. We need to use other sources from different orientations in future to ensure results neutrality and non-alignment. We might need, also, to use domain-specific text corpora and domain-specific upper ontologies to make sure that the results we obtain are of higher relevancy to the humanitarian domain. Using the archive of The Guardian to train our model produced promising results, though having more specialised sources of data would significantly improve those results.

In comparison to existing approaches, which are heavily dependent on human intervention, our work in this paper provides faster and cheaper decision support mechanism by eliminating the slow and expensive human involvement in knowledge retrieval, adapting open source computing technologies to automate knowledge representation, and using publicly – and freely – available text corpora to train our model. However, our solution has a few disadvantages such as exclusive reliance on secondary data sources, territorially limited geographical coverage, and broad categories used in semantic classification.

CONCLUSION

In this paper, we present a technique that uses word embeddings and semantic web technologies to build a

semantically classified word embedding model. The proposed technique used 6,627,078 words, in 7,053 documents, extracted from the archive of The Guardian newspaper to train a word2vec model using continuous-bag-of-words (CBOW), making a vector space of 200 dimensions and 94,104 vectors. We used semantic web techniques to classify terms in this model by harvesting semantics from upper ontologies and assign them to terms in word2vec model. The results of the previous two processes (i.e. vectorisation and ontologisation) were locally stored in a semantically classified word embedding (SC-WE) model. We also developed a query language to retrieve information and find answers to the questions of the decision-makers in humanitarian crises, and use cosine distance to measure similarities and semantic matching to filter the results. The results of implementation show that using semantic matching to classify word embedding model yields more relevant results to the queries of the end-users. We plan to use the findings of this study with other data formats, bigger data sets, more diverse data sources, more sophisticated ontologies, and better training models to improve humanitarian information retrieval in the future.

REFERENCES

ACAPS. (2017). "Expert judgment - The use of expert judgment in humanitarian analysis: Theory, methods and applications." *The Assessment Capacities Project*, https://www.acaps.org/library/assessment#resource-930 (Jun. 11, 2018).

Albukhitan, S., Alnazer, A., and Helmy, T. (2018). "Semantic Annotation of Arabic Web Documents using Deep Learning." *Procedia Computer Science*, 130, 589–596.

Amnesty International. (2018). *No safe refuge: Asylum-seekers and refugees denied effective protection in Turkey*. Amnesty International, London WC1X 0DW, UK.

Atanasova, T., Kasheva, M., Sulova, S., and Vasilev, J. (2010). "Analysis of the possible application of Data Mining, Text Mining and Web Mining in business intelligent systems." *Proceedings of the 33rd International Convention MIPRO*, IEEE Conference, 1294–1297.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Celikyilmaz, A., Hakkani-Tür, D., Pasupat, P., and Sarikaya, R. (2015). "Enriching Word Embeddings Using Knowledge Graph for Semantic Tagging in Conversational Dialog Systems." *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium Series*, 39–42.

Chen, Y. N., and Rudnicky, A. I. (2014). "Dynamically supporting unexplored domains in conversational interactions by enriching semantics with neural word embeddings." *Proceedings of the Workshop on Spoken Language Technology (SLT)*, 590–595.

Clark, T., Kessler, C., and Purohit, H. (2015). "Feasibility of Information Interoperability in the Humanitarian Domain." *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium Series*, 2–6.

Gordo, A., Almazan, J., Murray, N., and Perronnin, F. (2015). "LEWIS: Latent Embeddings for Word Images and their Semantics." http://arxiv.org/abs/1509.06243 (Jun. 6, 2018).

Grabarske, J., and Heutelbeck, D. (2012). "An Upper Ontology for the Social Web." *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 1128–1131.

Huang, W., and Dai, W. (2017). "Knowledge storage and acquisition for industrial cyber-physical systems based on non-relational database." *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, 6671–6676.

Humud, C. E., Blanchard, C. M., and Nikitin, M. B. D. (2018). *Armed Conflict in Syria: Overview and U.S. Response*. Congressional Research Service, United States Congress, Washington, D.C.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). "Processing Social Media Messages in Mass Emergency: A Survey." *ACM Comput. Surv.*, 47(4), 67:1–67:38.

Li, H., and Xu, J. (2014). *Semantic Matching in Search*. Now Publishers Inc, Boston, Mass.

Ling, Y., An, Y., Liu, M., Hasan, S. A., Fan, Y., and Hu, X. (2017). "Integrating extra knowledge into word embedding models for biomedical NLP tasks." *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 968–975.

Liu, Q., Jiang, H., Wei, S., Ling, Z.-H., and Hu, Y. (2015). "Learning Semantic Word Embeddings based on Ordinal Knowledge Constraints." *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1, 1501–1511.

Maas, A. L., and Ng, A. Y. (2010). "A Probabilistic Model for Semantic Word Vectors." *Proceedings of the Workshop on Deep Learning and Unsupervised Feature Learning*, 10.

Malizia, A., Onorati, T., Diaz, P., Aedo, I., and Astorga-Paliza, F. (2010). "SEMA4A: An ontology for emergency notification systems accessibility." *Expert Systems with Applications*, 37(4), 3380–3391.

Mhatre, M., Phondekar, D., Kadam, P., Chawathe, A., and Ghag, K. (2017). "Dimensionality reduction for

sentiment analysis using pre-processing techniques." *Proceedings of the International Conference on Computing Methodologies and Communication (ICCMC)*, 16–21.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space." http://arxiv.org/abs/1301.3781 (Jun. 6, 2018).

Neches, R., Fikes, R. E., Finin, T., Gruber, T., Patil, R., Senator, T., and Swartout, W. R. (1991). "Enabling Technology for Knowledge Sharing." *AI Magazine*, 12(3), 36.

OCHA. (2017). "Syrian Arab Republic: 2017 Humanitarian Response Plan." *United Nations Office for the Coordination of Humanitarian Affairs*, https://www.humanitarianresponse.info/en/operations/whole-of-syria/document/2017-syrian-arab-republic-humanitarian-response-plan (Jun. 11, 2018).

Othman, N., Faiz, R., and Smaili, K. (2017). "A Word Embedding based Method for Question Retrieval in Community Question Answering." *Proceedings of the International Conference on Natural Language, Signal and Speech Processing*.

Pilehvar, M. T., and Collier, N. (2016). "Improved Semantic Representation for Domain-Specific Entities." *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, 12–16.

Řehůřek, R., and Sojka, P. (2010). "Software Framework for Topic Modelling with Large Corpora." *Proceedings of the Workshop on New Challenges for NLP Frameworks*, ELRA, 45–50.

ReliefWeb. (2018). "Ongoing disasters actively monitored by ReliefWeb." *ReliefWeb*, https://reliefweb.int/disasters (Jun. 11, 2018).

Roy, A., Park, Y., and Pan, Sh. (2017). "Learning Domain-Specific Word Embeddings from Sparse Cybersecurity Texts." *arXiv:1709.07470 [cs]*.

Smith, B., and Brooke-Holland, L. (2018). *Syria and chemical weapons - in brief*. Briefing Paper, The House of Commons, London, UK.

Widener, D. V., Mazzuchi, T. A., and Sarkani, S. (2017). "Simplifying humanitarian assistance/disaster relief analytic models using activity-based intelligence: Syrian refugee crisis as a case study." *Disaster Prevention and Management: An International Journal*, 27(1), 60–73.

World Bank. (2017). *The Toll of War: The Economic and Social Consequences of the Conflict in Syria*. Report, World Bank Group, Washington, DC.

Xu, C., Bai, Y., Bian, J., Gao, B., Wang, G., Liu, X., and Liu, T.-Y. (2014). "RC-NET: A General Framework for Incorporating Knowledge into Word Representations." *Microsoft Research*, https://www.microsoft.com/en-us/research/publication/rc-net-a-general-framework-for-incorporating-knowledge-into-word-representations/ (Jun. 6, 2018).

Yao, Z., Sun, Y., Ding, W., Rao, N., and Xiong, H. (2018). "Dynamic Word Embeddings for Evolving Semantic Discovery." *Proceedings of the 11th Association for Computer Machinery (ACM) International Conference on Web Search and Data Mining*, WSDM '18, 673–681.

Yin, J., Lampert, A., Cameron, M., Robinson, B., and Power, R. (2012). "Using Social Media to Enhance Emergency Situation Awareness." *IEEE Intelligent Systems*, 27(6), 52–59.

Zhang, Y., Wang, X., Lai, S., He, S., Liu, K., Zhao, J., and Lv, X. (2014). "Ontology Matching with Word Embeddings." *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, Lecture Notes in Computer Science, Springer, Cham, 34–45.