SC-WE: A Semantically Classified Word Embedding for Information Retrieval in Humanitarian Crises

Aladdin Shamoug, Stephen Cranefield and Grant Dick
Department of Information Science
UNIVERSITY OF OTAGO (Dunedin, New Zealand)

Overview

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. 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. We train a word embedding model using word2vec, transform words and terms from news archive to entities in domain ontology, annotate those entities 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. 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.

Problem Statement

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.
- ❖ 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.
- ❖ 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.

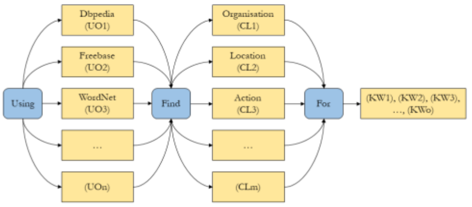
The Solution

- ❖ 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).

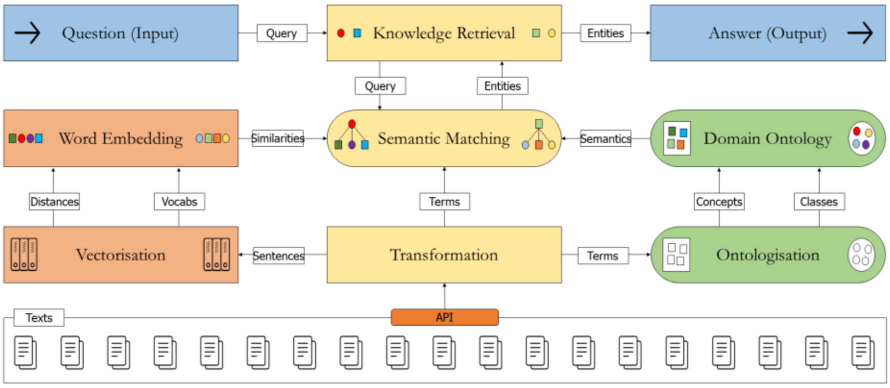
The Results

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 our 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 support 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.

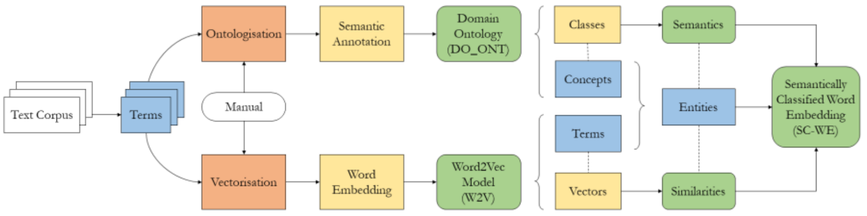
Semantically Classified Word Embedding Query Language



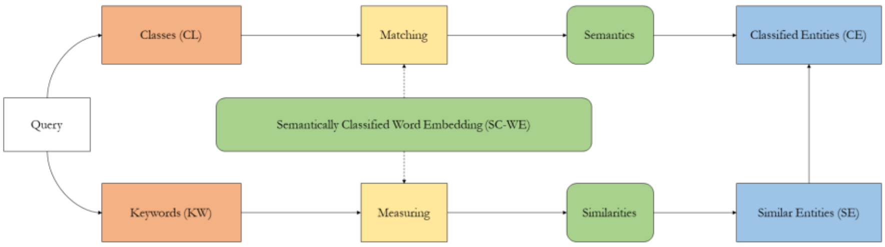
The Framework



Knowledge Representation Model



Knowledge Retrieval Model



Queries | Similar Entities ☆ | Classified Entities ◇

Task 1: Asylum Seekers

Question	Keywords (KW)	Classes (CL)
What countries do Syrian refugees seek asylum in?	Countries, Syrian, Refugees, Seek, Asylum	Location

```
SCWEQL >>> USING (DBPEDIA) FIND (*) FOR (COUNTRIES, SYRIAN, REFUGEES, SEEK, ASYLUM) LIMIT (5) ☆
SCWEQL >>> USING (DBPEDIA) FIND (LOCATION) FOR (COUNTRIES, SYRIAN, REFUGEES, SEEK, ASYLUM) LIMIT (5) ◇
```

SE (Using Keywords)	Similarity	Frequency
Syrians	0.64	3661
Syrian Refugees	0.60	1715
Asylum Seekers	0.58	493
Migrants	0.57	940
Seekers	0.56	8

CE (Keywords + Classes)	Similarity	Frequency
Greece	0.55	815
Germany	0.46	1403
Turkey	0.45	7358
Europe	0.45	3485
Italy	0.44	490

Task 2: Medical Aid Providers

Question	Keywords (KW)	Classes (CL)
Which organisations provide medical aid to Syrian people?	Organisation, Provide, Medical Aid, Syrian	Organisation

```
SCWEQL >>> USING (DBPEDIA) FIND (*) FOR (ORGANISATION, PROVIDE, MEDICAL AID, SYRIAN) LIMIT (5) ☆
SCWEQL >>> USING (DBPEDIA) FIND (ORGANISATION) FOR (ORGANISATION, PROVIDE, MEDICAL AID, SYRIAN) LIMIT (5) ◇
```

SE (Using Keywords)	Similarity	Frequency
Providing Humanitarian	0.58	44
Communications Equipment	0.56	21
Humanitarian Organisations	0.56	83
Relief Organisations	0.56	16
Child Protection	0.56	31

CE (Keywords + Classes)	Similarity	Frequency
SARC	0.50	51
UNRWA	0.50	109
Red Cross	0.49	352
Works Agency	0.49	25
International Committee	0.48	149

Task 3: Eastern Aleppo in 2016

Question	Keywords (KW)	Classes (CL)
What happened in Eastern Aleppo in 2016?	Happened, Eastern Aleppo, 2016	Action

```
SCWEQL >>> USING (DBPEDIA) FIND (*) FOR (HAPPENED, EASTERN ALEPPO, 2016) LIMIT (5) ☆
SCWEQL >>> USING (DBPEDIA) FIND (ACTION) FOR (HAPPENED, EASTERN ALEPPO, 2016) LIMIT (5) ◇
```

SE (Using Keywords)	Similarity	Frequency
This Year	0.58	996
Eastern Ghouta	0.57	249
East Aleppo	0.57	237
Next Year	0.57	210
March 2011	0.55	216

CE (Keywords + Classes)	Similarity	Frequency
Clashes Erupted	0.50	12
Heavy Fighting	0.49	86
Reconvening	0.49	3
Be Replicated	0.47	11
Heavy Shelling	0.46	48

Task 4: Chemical Attacks

Question	Keywords (KW)	Classes (CL)
Where did chemical attacks take place in Syria?	Chemical, Attack, Took Place, Syria	Location

```
SCWEQL >>> USING (DBPEDIA) FIND (*) FOR (CHEMICAL, ATTACK, TOOK PLACE, SYRIA) LIMIT (5) ☆
SCWEQL >>> USING (DBPEDIA) FIND (LOCATION) FOR (CHEMICAL, ATTACK, TOOK PLACE, SYRIA) LIMIT (5) ◇
```

SE (Using Keywords)	Similarity	Frequency
21 August	0.65	266
Chemical Attack	0.64	437
Gas Attack	0.59	74
Chlorine	0.56	143
Civilian Areas	0.56	127

CE (Keywords + Classes)	Similarity	Frequency
Ghouta	0.56	283
Houla	0.53	175
Eastern Ghouta	0.53	249
Khan Sheikhun	0.52	120
Damascus Suburbs	0.51	103

Task 5: United Nations Interventions

Question	Keywords (KW)	Classes (CL)
What does the United Nations do in Syria?	United Nations, Doing, Syria	Action

```
SCWEQL >>> USING (DBPEDIA) FIND (*) FOR (UNITED NATIONS, DOING, SYRIA) LIMIT (5) ☆
SCWEQL >>> USING (DBPEDIA) FIND (ACTION) FOR (UNITED NATIONS, DOING, SYRIA) LIMIT (5) ◇
```

SE (Using Keywords)	Similarity	Frequency
Inside Syria	0.42	753
Humanitarian	0.39	1813
Humanitarian Aid	0.39	496
Facilitating	0.38	53
European Union	0.37	368

CE (Keywords + Classes)	Similarity	Frequency
Facilitating	0.38	53
Securing	0.35	214
Coordinating	0.35	97
Establishing	0.35	209
Ongoing	0.35	527