Information Systems Frontiers (2018) 20:909–923
https://doi.org/10.1007/s10796-018-9843-x

Emergency Vocabulary

Dávid Márk Nemeskey[1] · **András Kornai**[1]

Published online: 26 March 2018

Abstract
For disaster preparedness, a key aspect of the work is the identification, *ahead of time*, of the vocabulary of emergency messages. Here we describe how static repositories of traditional news reports can be rapidly exploited to yield disaster- or accident-implicated words and named entities.

Keywords Information retrieval · Emergency · Vocabulary

1 Introduction

Our goal in this paper is to define an Emergency Vocabulary (EV), composed of words and phrases (n-grams), including named entities, that are highly characteristic of emergencies, without the use of expert or commonsense (crowdsourced) knowledge. The rest of this Introduction provides background and summarizes related work. Section 2 describes a series of pilot experiments performed on the NewReuters collection to see how a Basic Emergency Vocabulary (BEV) composed of unigrams can be iteratively refined for the purpose of building classifiers to select emergency-related material in English and other languages with a minimum amount of manual work. Section 3 describes experiments with a more sophisticated, semantics-based approach. Section 4 applies the lessons learned to a considerably larger corpus (CommonCrawl news) and adds n-grams. We evaluate the results in Section 5, and offer some conclusions.

1.1 Related Work

Our goals are very similar to those of the developers of CrisisLex (Olteanu et al. 2014) and, inevitably, our methods also show strong similarities. The main difference between their work and ours is that we avoid crowdsourcing at all stages, aiming at more automated discovery and testing – in this regard, our work is closer to Soni and Pal (2017) than Basu et al. (2017). A secondary difference is that CrisisLex, as well as the broad variety of systems surveyed in Imran et al. (2015), tend to operate on Twitter messages, whereas we work with more static collections of data such as New Reuters (Lewis et al. 2004) and CommonCrawl.[1]

In the normal course of events, emergencies like natural disasters, military escalation, epidemic outbreaks etc. are almost immediately followed by some response, such as containment and mitigation efforts, counterattack, quarantine, etc., often within minutes or hours, and much of the work on emergency response is concerned with exploiting this short-term dynamics and the messages (usually tweets) generated while the emergency is still unfolding (Phuvipadawat and Murata 2011).

Yet for preparedness, a key aspect of the work is the identification, *ahead of time*, of words characterizing emergency reports and of potentially implicated locations (LOC), organizations (ORG), and persons (PER). Here we describe how static repositories operating on a much slower (typically, daily) news cycle can be exploited to yield disaster- or accident-implicated words and named entities. Most of our results are on English data, but our bootstrap method, based on pseudo-relevance feedback (Buckley et al. 1995) works for any language, and to show this we evaluate our method on Hungarian as well.

Much of the work in this area involves the building and deployment of large-scale systems that ideally feed into emergency relief and response work in real time (Imran

Work supported by DARPA LORELEI 15-04.

✉ Dávid Márk Nemeskey
nemeskey.david@sztaki.mta.hu

András Kornai
andras@kornai.com

[1] HAS Institute of Computer Science, Kende u. 13-17, Budapest 1111, Hungary

[1] http://commoncrawl.org

et al. 2014; Imran 2017). We see our work as building preparatory infrastructure for such live systems, especially for the medium- and low-resource languages spoken in many disaster areas (Strassel et al. 2017).

2 Basic Emergency Vocabulary

The idea that there is a *basic vocabulary* (BV) composed of a few hundred or at most a few thousand elements goes back to the Renaissance – for a more detailed history, see Ács et al. (2013), for a contemporary list of 1200 items see Appendix 4.8 of Kornai (2018).[2] The basic emergency vocabulary (BEV) serves a dual purpose: first, these words are the English bindings for deep semantic (conceptual) representations that can be used as an interlingual pivot or as a direct hook into knowledge-based (inferential) systems; and second, these words act as a reasonably high-precision high-recall filter on documents that are deemed relevant for emergencies: newspaper/wire articles, situation reports, tweets, etc. In fact, rough translations of these words into a target language T can serve as a filter for emergency-specific text in T, a capability we evaluate on Hungarian in Section 4.

In terms of applications, the basic concept list promises a strategy of gradually extending the vocabulary from the simple to the more complex and conversely, reducing the complex to the more simple. Thus, to define *asphyxiant* as 'chemical causing suffocation', we need to define *suffocation*, but not *chemical* or *cause* as these items are already listed in the basic set. Since *suffocate* is defined as 'to lose one's life because of lack of air', by substitution we will obtain for asphyxiant the definition 'chemical causing loss of life because of lack of air' where all the key items *chemical*, *lose*, *life*, *because*, *lack*, *air* are part of the basic set.

Unfortunately, the list of emergency-related concepts, and topic-centric concept sets in general, are not closed definitionally: for example, the verb *decontaminate* is highly characteristic of nuclear and chemical emergencies, but the definition 'to remove a dangerous substance' is composed of parts that in isolation are not particularly emergency-related. Even *danger* is a normal part of many human activities from certain sports to industrial processes that do not, in themselves, constitute emergencies. A related problem is seen in proper nouns, where the assessment of Mount Fuji (LOC) or Kim Jong-un (PER) as a source of emergency is highly problematic. Proper nouns are discussed further in Kornai (2010), but we note here that they constitute a very small proportion (less than 6%) of the basic vocabulary. Since none of the basic NERs, whose list is restricted to names of continents, countries, major cities, founders of religions, etc., are particularly implicated in disasters or accidents, the methods discussed here involve no seeding for the actual entity categories we wish to learn. Our seed lists, minimal as they are, will contain only common nouns, verbs, adjectives and adverbs.

Table 1 Seed word list extracted from Wikipedia

airburst avalanche blizzard breach collapse collision crash derailment disaster drought earthquake epidemic eruption explode famine flood hailstorm heat impact landslide massacre mudslide prevention radiation riot shipwreck shutoff sinkhole spill terrorism thunderstorm tornado tragedy tsunami volcano wildfire wreck

2.1 Manual Extraction

Perhaps the simplest way of building the Basic Emergency Vocabulary (BEV) is by manual selection. At 1200 items, the basic list is small enough to permit manual selection of a seed emergency list, about 1/10th of the basic list, by the following principles. First, we included from the basic list every word that is, in and of itself, suggestive of emergency, such as *danger*, *harm*, or *pain*. Second, we selected all concepts that are likely causes of emergency, such as *accident*, *attack*, *volcano*, or *war*. Third, we selected all concepts that are concomitant with emergencies, such as *damage*, *Dr*, or *treatment*. Fourth, because of semantic decomposition, we added those concepts that signal emergencies only in the negative, such as *breathe* or *safe* (can't breathe, not safe, unsafe). Fifth, and final, we added those words that will, on our judgment, appear commonly in situation reports, news articles, or even tweets related to emergencies such as *calm*, *effort*, *equipment*, or *situation*. The full list of these manually selected entries is given in Appendix A.

The same criteria were applied in extracting a list from the section titles of the Wikipedia page on natural disasters.[3] This initial list was expanded with a few terms that refer to human-induced emergency situations, such as *terrorism* and *massacre*. See Table 1 for the full list.

While the simplicity of the manual selection method is attractive, the results are not very good. To evaluate precision and recall, we analyzed the glossary (TRADE Emergency Management Issues SIG Glossary Task Force 1999) using the same principles as above. This yielded another 267 words like *hazmat* or *thermonuclear*. These were taken, for the most part, from the definitions in the glossary, not the headwords, especially as the latter are often highly specific to the organization of US emergency response procedures, while our goal is to build a language-independent set of concepts, not something specific to

[2] http://bit.ly/2AejT0p

[3] https://en.wikipedia.org/wiki/Natural_disaster

American English. Of the 260 words found in the Glossary, only 41 appear on the basic list, and of these, only half (22) were found on the first manual pass over the basic list. In hindsight, it is clear that the remaining 19, given in italics at the end of Appendix A, should also have been selected based on the above principles, especially the last (fifth) one. The precision is 100%, as expected from a manual set, but recall is low (15.8%).

2.2 Corpus-Based, Manual Seed

The lesson from Section 2.1 is clear: the BEV list has to be built from emergency materials, rather than by human expertise. But there is something of a chicken and egg problem here: to have a good list, we need to have a good corpus of emergency materials, to have a good corpus, we need to build a good classifier, and to build a good classifier, we need a good list. Here we describe a method of jointly bootstrapping the list and the emergency corpus.

In a pilot experiment, we used the manually built list given in Appendix A as positive evidence (for a theoretical justification of ignoring negative evidence see Kornai et al. 2003) to select a small, emergency-related subset E of articles from a corpus C (in the experiment, the New Reuters collection of 806,791 news stories) by a simple, semi-automatic iterative process. First, the articles were indexed by a search engine, and Appendix A was used as the initial search query. Of the documents returned by the engine, only the most relevant N were retained. The threshold was selected in such a way that in a window of documents around it, about half should be emergency-related. A linear search from the top would have obviously been infeasible, but with a binary search among |C| documents with a window size W, N can be found by looking at only $W \log_2(|C|)$ documents – in our case we only had to look at 80 documents of the entire corpus to select a core set E of about 2000 emergency-related articles. Here binary search is made feasible by the empirical fact, broadly used in all forms of pseudo-relevance feedback since SMART (Buckley et al. 1995), that higher combined Δ scores yield more relevant documents.

This is a noisy sample, only about 80% of the documents in it are actually emergency related, and by sampling New Reuters we estimated recall also to be only about 80%, so there may still be about 400 further emergency-related articles in the corpus. An F-measure of .8 will not be impressive if our goal was detecting emergency-related articles in a live stream. But here even a more pessimistic estimate of missed documents, such as provided in Soni and Pal (2017) (perhaps more realistic for tweets than for the full news articles in our corpus) does not unduly affect the logic of our enterprize. Since a random document in the corpus C will be emergency related with probability $p = 0.0025$,[4] but in the subset E with probability $p = 0.8$, words in the subsample are far more likely to be emergency-prone. To quantify this, we computed log text frequency ratio $\Delta = \log(TF(w, E)/TF(w, C))$ for each word; the TF values were normalized as in Okapi BM25. We focused on the 1700 words where this exceed the expected zero log ratio by at least two natural orders of magnitude. Of these, the ratio is greater than 3 for about a quarter (472 items), and greater than 4 for about one in twelve (135).

Since in E we have only 2k relatively short documents to consider, we ran the NER system from Stanford CoreNLP (Manning et al. 2014) on these, and collected the results for all 1700 words. Typically, words are classified unambiguously (label entropy is below 0.1 for over 82%), and by ignoring the rest we still obtain 1398 words. Table 2 lists the highest ranked words.

Two-thirds of the words in the list are emergency-related common nouns (e.g. *levee*, *floodwater*, *mudslide*). This finding is so significant that in subsequent experiments we could in fact dispense with the manual selection method of Section 2.1 altogether, and bootstrap the classifier starting with only a handful of words – this will be described in Section 2.3.

Table 2 is also very promising in terms of identifying emergency-implicated NERs by searching for those NERs that occur in an emergency-related subcorpus considerably more frequently than in the corpus as a whole. Certainly, among the tens of thousands of locations in the NewReuters corpus, the method puts at the top Hohenwutzen, Slibice, and Oderbruch, still very much exposed to floods of the river Oder, and Popocatepetl, a volcano that has been implicated in half a dozen new eruptions since the corpus was collected. Among persons, the top choices are 'Matthias Platzeck, environment minister in the German state of Brandenburg' and 'government crisis committee spokesman Krzysztof Pomes'.

2.3 Minimum Seed

Instead of using the laboriously collected, yet still very incomplete, list of Appendix A, here we considered a seed list of only two words: *emergency* and *urgent*. Looking at the documents that contain at least one of these two words we can obtain an emergency-related corpus of documents E'. The top of the list of words that are significantly more frequent in E' than in the background are shown in Table 3.

While this list is not quite as good as the actual BEV (e.g. it has outright false positives like *nirmala*), it is good enough for further iteration. The emergency sets obtained from the BEV and from this skeletal list are far from identical, but

[4] $p = 0.01$ if we use the factor of four discovered in Soni and Pal (2017)

Table 2 Highest ranked words in emergency subcorpus

Word	Δ	DF	NER
Hohenwutzen	5.53	110	L:110
Platzeck	5.52	59	P:57
Slubice	5.50	90	L:88
Oderbruch	5.50	196	L:189
Dike	5.49	993	N:999
Sandbag	5.45	362	N:334
Oder	5.39	542	L:81,M:1,N:149,P:272
Popocatepetl	5.39	54	L:29,N:1,O:1,P:23
Pomes	5.30	78	P:68
Stolpe	5.29	57	P:45
Levee	5.28	229	N:188,O:1
Floodwater	5.27	437	N:366
Soufriere	5.23	62	L:48,N:5,P:1
Abancay	5.23	51	L:39
Opole	5.21	105	L:79,N:2
Forks	5.17	409	L:176,M:110,N:5,O:24
Montserrat	5.15	268	L:267,O:18
Hortense	5.13	399	L:1,M:1,N:56,O:4,P:236
Low-lying	5.13	299	N:214
Flood-ravaged	5.10	74	M:1,N:57
Nirmala	5.09	122	P:87
Godavari	5.08	128	L:78,N:12
Falmouth	5.07	60	L:37
Sodden	5.06	67	N:44
Eruption	5.01	537	N:339
Volcano	4.99	870	L:7,N:616,O:20,P:1
Hurricane-force	4.98	53	N:33
Flood-stricken	4.98	62	N:40
Evacuee	4.93	291	N:184
Flood-hit	4.93	123	M:1,N:111
Jarrell	4.92	81	L:19,N:5,O:4,P:21
Yosemite	4.92	100	L:52,N:3,O:1
Bandarban	4.91	53	L:28
Lava	4.90	136	N:76
Mudslide	4.90	458	N:295

Last column gives number of occurrences as P: PER; L: LOC; O: ORG; N: not NE

Table 3 Highest Δ words (ordered alphabetically) from seed *emergency*, *urgent*

ambulance anarchic anarchy angioplasty arsenc beachfront blizzard calamity coastline curfew cyclone devastation dike disaster dyke emergency eruption evacuate evacuation evacuee famine firefighter flood flood-hit flooding floodwater frantically gust gusty hard-hit hurricane impassable inaccessible insurgent insurrection inundate issues landfall levee low-lying malaria malnutrition melting meningitis mortuary mudslide nirmala non-essential ntsb overflow pilots preparedness rescuer rioter sandbag shelter stone-thrower submerge swollen tornado torrential tributary urgent volcanic volcano wildfire worst-hit

the iterative method remains attractive in a low resource setting (a matter we shall investigate more formally in Section 5 for Hungarian) especially in combination with the embedding-based lexicon expansion methods we now turn to.

3 Semantic Similarity

Since bootstrapping the vocabulary from very small seeds (three words or less) remains challenging, we experimented with word embeddings to enlarge our lexicon, especially as reasonably mature embeddings are often available even for those languages where no emergency (foreground) corpus is readily available. As the pilot study made clear, there are many cases where words are related morphologically but not semantically, e.g. *staging* – the sense associated to emergencies, 'staging area', has nothing to do with *stage*. Human coders deal with morphological variability automatically, but CoreNLP lemmatization (Manning et al. 2014) is noticeably imperfect in this regard, failing to deal with deeper morphology *fissionable/fissile*, typos and nonstandard spelling *catastrophy/catastrophe*, and the frequent variations in hyphenation *lifethreatening/life-threatening*. Such problems are of course rampant in less resourced languages, where morphological analysis and stemming often pose even more problems than in English.

To a remarkable extent, such issues are taken care of by the central method discussed here, the use of semantic vectors. In large corpora, spelling variants and typos appear dominantly in the same contexts as the standard form, so the vectors assigned to them will be highly similar. Here and in what follows we used the pre-trained GloVe (Pennington et al. 2014) embedding `840B.300d`.[5] Originally, the embedding contains 2,196,017 words with the associated 300-dimension vector trained on the Common Crawl dataset. Out of these, 2,196,013 words could be read correctly; filtering duplicates (which arise from Unicode whitespaces left in the data) leaves us with 2,195,893 tokens. These were lemmatized with CoreNLP, and words that differed from their lemmas were dropped. The reason for this is that the original embedding was highly redundant, with some lemmas represented by three or more surface forms (e.g. *tsunami*, *tsunamis* and *Tsunamis*), not counting typos. Filtering these forms brought the number of words down to 1,397,824. Finally, we eliminated numbers, proper nouns and punctuation marks by requiring that words start with a lower case latin letter, thus ending up with an embedding of 480,427 vectors – a little over 78% reduction from the original.

[5] http://nlp.stanford.edu/data/glove.840B.300d.zip

We ran an algorithm inspired by DBSCAN (Ester et al. 1996) to find lexicon enlargement candidates. DBSCAN proper is not applicable to the problem, since emergency-related words do not cluster together in the word vector space, and the DBSCAN notion of *core points* is not applicable. Our version of the algorithm starts from the words already in the lexicon, and it finds the candidates whose vectorial similarity to any of them is above a certain threshold. Next, candidates not "close enough" to the emergency terms are discarded, and the rest is added to the lexicon. By "close enough", we refer to the notion that it is not sufficient for a candidate word to be close to an emergency related term; its closest associations must also be emergency-related. We formalized this condition in two ways:

1. the closest neighbor of the candidate should be a word already in the lexicon;
2. the discounted cumulative gain (DCG) (Järvelin and Kekäläinen 2002) of the closest n neighbors, equating the "relevant" class with the emergency lexicon, is above a certain threshold.

It might not readily be obvious, but the second condition is a relaxation of the first. In fact, the first condition is a special case of the second where $n = 1$.

The algorithm is run repeatedly for a number of iterations, or until no more candidates can be found. We experimented with various hyperparameter settings (the condition used, the similarity threshold and n in DCG); Table 4 shows the result of successive steps in one of these settings. Clearly, most of the associations (which are not blocked by the morphological and spelling problems discussed above) are perfectly reasonable e.g. from *volcano* to *volcanic* to *lava* to *magma* to *plume*. The only problem is that by the time we get to *plume* or *caldera*, the sense of emergency is gone. The reason for this is twofold: first, as already discussed in Section 2, the lexicon is not closed under semantic similarity. Second, standard embeddings do not differentiate between the senses of homonymous and polysemous words, and even those that were created for this purpose leave a lot to be desired (Borbély et al. 2016).

To some extent, this is remedied by selecting a higher threshold of similarity. As the words in italics show, some of the more remote associations are dropped, but with the increased precision comes lower recall. This is especially painful because highly relevant terms, such as *ebola*, *oil-spill* or *shipwreck* are removed, while some unrelated and bogus terms, such as *caldera* and *plauge* (sic) are retained. In general, it is impossible to find a global threshold that cuts all association chains at the right places.

Another limitation of generic embeddings is that only unigrams are covered. On the one hand, it is possible to embed n-grams by averaging the vectors of their component words. On the other, as the Table 5 shows, the results are mixed: sometimes the nearest neighbors of the n-gram vector are dominated by one of the components (*hundred dead*); sometimes they are a reweighted union of the neighbor sets of the individual words (*tornado damage*); other times they reflect the compound meaning (*flood emergency*).

Even if we could reliably give meaningful vectors to n-grams, the algorithm would not fare any better, because of the thresholding problem. Furthermore, we cannot use the

Table 4 Nearest neighbors of some emergency keywords at cosine similarity > 0.4

Term	Iterations			
	1	2	3	4
Airburst	Air-burst			
Blizzard	Snowstorm	Snowfall	*Lake-effect*	*Snowbelt*
Collision	Head-on	*Headon*		
Crash	Accident	Incident		
		Mishap	*Malfunction*	
Earthquake	Magnitude			
	Quake	Aftershock	Temblor	*Seism*
Epidemic	Outbreak			
	Plague	Bubonic	*Ebola*	*Marburg*
			Plauge	
	Scourge	Menace		
Eruption	Erupt			
	Eruptive			
Explode	Burst			
	Implode			
Famine	Pestilence			
	Starvation	Deprivation		
		Starve		
Flood	Flooding			
Hailstorm	*Hailstone*			
	Windstorm			
Riot	Rioting			
Shutoff	Shut-off			
Spill	*Oil-spill*			
Terrorism	Terror			
	Terrorist			
Thunderstorm	Squall	Gale		
Tornado	*Mile-wide*			
	Twister			
Volcano	Crater	Caldera	Calderas	
	Volcanic	Lava	*Geyser*	
			Magma	*Plume*
Wildfire	Bushfire			
Wreck	*Shipwreck*	*Galleon*		
		Sunken		

Terms filtered by a similarity of 0.6 are in *italics*

Table 5 Bigram embeddings

Input term(s)	Nearest neighbors
Hundred	Fifty thousand twenty thirty eighty sixty forty seventy fifteen twenty-five
Dead	Alive death corpse kill man grave hell victim apparently murder
Hundred dead	Thousand fifty twenty thirty ten forty seventy sixty eighty fifteen
Tornado	Twister storm hurricane thunderstorm devastation earthquake tsunami aftermath cyclone flooding
Damage	Damaging damages harm cause destruction affect attack severe prevent injury
Tornado damage	Storm devastation hurricane damaging flooding damages lightning disaster severe earthquake
Alive	Dead forever survive life still never ever death soul alone
Rubble	Wreckage debris pile detritus bombed-out demolition devastation cinder concrete gravel
Alive rubble	Dead survive wreckage corpse remains unharmed forever bury still literally
Flood	Flooding deluge storm inundation hurricane tsunami levee disaster torrential floodwater
Emergency	Ambulance evacuation hospital urgent medical disaster assistance aid preparedness safety
Flood emergency	Flooding disaster evacuation storm fire hurricane preparedness catastrophic urgent tsunami

algorithm to find relevant n-grams, as no (tractable) algorithm exists that maps a certain vector back to a set of words.

Because of these difficulties, we did not use embedding-based query expansion beyond the unigram BEV, where we relied both on word vector similarity, and dictionary similarity (Ács and Kornai 2016), selecting only words that (i) were above a fixed cosine similarity threshold from the cluster center and (ii) contained, in their definition, some basic emergency word. Among the lessons learned, we had to give up our fourth principle (see Section 2.1), adding those concepts that signal emergencies only in the negative, such as *breathe* or *safe* (can't breathe, not safe). The issue can perhaps be reopened in conjunction with adding collocation analysis to the pipeline, but for now we kept only the single token entries such as *unsafe*. We also made some concessions to the current state of the art in morphological analysis, including pairs like *bioterror/bioterrorist* where the complex form should be (but is not) analyzed by most current lemmatizers.

The resulting BEV (see Appendix C), has 419 words, of which only 84 appear in BV. This version has only 181 words in common with the initial, manually selected version (Appendix B, 349 words), with 168 words dropped and 238 added. The final version is much better focused, with an average word frequency of 93,500 as measured on the UMBC WebBase (Han et al. 2013), compared to 204,800 in the initial version. We will compare the utility of the two versions for information retrieval in Section 5.

As a final step, we plotted a 2-dimensional projection of the basic vocabulary (green), the manually selected basic emergency vocabulary (Section 2.1 and Appendix A, red), and their intersection (yellow).

As Fig. 1 makes clear, the emergency vocabulary clusters remarkably well on the left side. To see the outliers, we considered those BEV words that fall closer to the center of BV than the center of BEV (computed on the original 300d vectors, not their 2d projection) and found only 13 words: *department Dr escort issue launch lost mine rad release rod secure site stolen*. Clearly, the emergency sense of e.g. *rad* 'unit of absorbed radiation dose' is overwhelmed by the non-emergency senses 'radian, measure of angle' and 'radical' (of style, politics, etc). While such words may literally meet our fifth criterion in Section 2.1 and occur frequently in emergency-related texts, their precision is low and they are omitted from the final version of BEV (Appendix C).

4 Large-Scale Experiments

Our key method for extending seed lexica is to take some foreground corpus E, a background corpus C, and

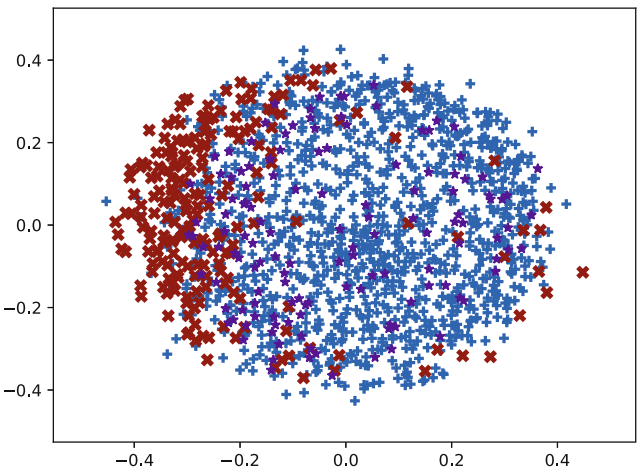

Fig. 1 2d projection of BEV (red ×), BV (blue +), and their intersection (purple star)

ranking words according to log TF ratio Δ as described in Section 2.2. Since we already exploited NewReuters in the pilot experiments discussed in Section 2, for the main experiments we used a similar, but much larger background corpus, the recently released news subset of Common Crawl (CC).[6] As we have seen in the previous section, unigrams participate in too many contexts to be good indicators for emergency-relatedness by themselves. Consequently, in the experiments below we also include bigrams.

In Section 4.1 we begin with a simple set of experiments in which the foreground collection was selected from the ReliefWeb collection.[7] Section 4.2 describes the preprocessing of CC News that resulted in a background corpus of nearly 3.9 billion words in over 14m documents. We experimented with several core term lists to see if it is possible to expand them and adapt them to a newswire corpus, including CrisisLex; the list of emergency-related terms extracted from Wikipedia given in Table 1; and BEV; see Section 4.3.

4.1 ReliefWeb

ReliefWeb is the result of a UN-sponsored effort to collect emergency relief-related materials on the web. It consists of 423,790 documents, out of which 110,932 have been classified for emergency type. There are a total of 21 types, and each emergency-related document is tagged with at least one of these (we ignore the rest and henceforth refer to the emergency-related subset by the name ReliefWeb). The number of documents in each category is reported in the second column of Table 6.

In the experiments, we took each disaster type subcorpus of ReliefWeb as foreground, computed Δ against the CC News background, and kept only those terms (unigrams and bigrams) that occurred at least 500 times in the background and at least 10 times in the foreground. We cut off the lists at $\Delta = 3$. The disastertype-specific lexica so created were compared to CrisisLex. Table 6 reports our findings: for each list, we record how many unigrams/bigrams from CrisisLex were *found* or were *missing* from the results, as well as the number of *new* emergency-related terms (not in CrisisLex) found (the difference between the *pre* and *post* conditions will be discussed in Section 4.2). Those categories given in italics contain very few documents (42–261) and are likely to be meaningless. We also performed the procedure for the entire ReliefWeb (the union of the disaster-specific types), these results are in the top panel. While the number of *found* tokens is small (evidence that CrisisLex and ReliefWeb are rather different), the number of *new* tokens is high, and manual inspection shows them to be good quality.

4.2 Preprocessing

The CC News dataset contains daily digests of news sites in many languages. When starting the experiments, we downloaded all files available at the time, which gave us a total of 1016 files to work with, dating from 26 August, 2016 to 29 June, 2017. The files were preprocessed as follows. First, boilerplate code was removed from the articles via the Boilerpipe library (Kohlschütter et al. 2010). In order to remove non-English materials, language identification was performed using the `CLD2`[8] and `langid.py`[9] (Lui and Baldwin 2012) libraries, the former via `cld2-cffi`[10]). The rest of the articles were lemmatized with Stanford CoreNLP (Manning et al. 2014). Duplicates then were filtered from the collection with the `datasketch`[11] package, with the following parameters. The documents were fingerprinted with a 128-permutation MinHash (Broder et al. 1998), computed from word (lemma) 5-grams. Locality sensitive hashing (Gionis et al. 1999; Indyk and Motwani 1998) was used to speed up fingerprint matching with an approximate Jaccard similarity threshold of 0.85. The deduplicated documents were filtered for stopwords. Tokens that contain no Latin letters or Arabic digits, and overly long tokens (above 30 characters), were also removed. Finally, words with low recall value, namely those that occurred in less then 10 documents, were removed as well.

The preprocessing code, as well as the components required to reproduce the experiments in this paper, is available as a GitHub repository.[12]

These settings yielded 14,163,517 documents comprising 3.9G word tokens in 8.4M types after lemmatization. Dropping low recall value words shrinks the number of types tenfold to 800k, but decreases the number of tokens only by 17 million. The average document length is thus 274 words.

In the experiments below, we include bigrams in the lexicon as well. While the above procedure is straightforward for the unigram case, filtering the bigram data can be done in two different ways. The linguistically correct way is to enumerate all bigrams present in the document first and then do the filtering, keeping only those bigrams whose component words were left intact. Olteanu et al. (2014), on the other hand, filtered the

[6] http://commoncrawl.org/2016/10/news-dataset-available/

[7] http://reliefweb.int

[8] https://github.com/CLD2Owners/cld2

[9] https://github.com/saffsd/langid.py

[10] https://pypi.python.org/pypi/cld2-cffi

[11] https://github.com/ekzhu/datasketch

[12] https://github.com/DavidNemeskey/cc_emergency_corpus

Table 6 CrisisLex coverage of the lexica generated using ReliefWeb as foreground

Category	Documents	Corpus	Unigrams				Bigrams			
			Found	Missing	New	Sum	Found	Missing	New	Sum
All	110,932	Pre	12	75	265	352	32	268	2479	2779
		Post	12	75	262	349	20	280	1258	1558
Cold wave	1460	Pre	4	79	176	263	15	270	936	1236
		Post	4	79	176	263	10	280	513	813
Drought	15,084	Pre	2	83	283	370	9	282	2781	3081
		Post	2	83	287	374	6	288	1405	1705
Earthquake	33,124	Pre	9	69	220	307	21	258	2484	2784
		Post	9	69	219	306	11	278	1225	1525
Epidemic	7212	Pre	1	85	287	374	8	284	2828	3128
		Post	1	85	293	380	5	290	1359	1659
Extratropical cyclone	243	Pre	11	65	90	177	12	276	155	455
		Post	11	65	85	172	10	280	111	411
Fire	42	Pre	5	77	7	94	0	300	0	300
		Post	5	77	7	94	0	300	0	300
Flash flood	8690	Pre	15	57	273	360	40	220	3188	3488
		Post	15	57	278	365	25	250	1565	1865
Flood	40,407	Pre	12	63	281	368	33	234	2831	3131
		Post	12	63	279	366	21	258	1356	1656
Heat wave	179	Pre	4	79	58	145	2	296	42	342
		Post	4	79	58	145	2	296	28	328
Insect infestation	1061	Pre	2	83	240	327	5	290	1021	1321
		Post	2	83	240	327	4	292	533	833
Land slide	8362	Pre	16	55	252	339	41	218	3097	3397
		Post	16	55	256	343	24	252	1510	1810
Mud slide	873	Pre	12	63	121	208	22	256	553	853
		Post	12	63	121	208	14	272	336	636
Other	8093	Pre	1	85	298	385	11	278	3202	3502
		Post	1	85	305	392	5	290	1596	1896
Severe local storm	1294	Pre	14	59	155	242	25	250	838	1138
		Post	14	59	153	240	17	266	465	765
Snow avalanche	261	Pre	5	77	76	163	4	292	114	414
		Post	5	77	76	163	3	294	78	378
Storm surge	2195	Pre	13	61	171	258	17	266	1333	1633
		Post	13	61	176	263	9	282	706	1006
Technological disaster	2004	Pre	5	77	182	269	15	270	834	1134
		Post	5	77	185	272	9	282	500	800
Tropical cyclone	25,595	Pre	12	63	266	353	36	228	2843	3143
		Post	11	65	261	348	23	254	1407	1707
Tsunami	10,697	Pre	6	75	229	316	18	264	2589	2889
		Post	6	75	229	316	13	274	1292	1592
Volcano	2228	Pre	7	73	222	309	12	276	1380	1680
		Post	7	73	219	306	7	286	743	1043
Wild fire	737	Pre	6	75	93	180	8	284	354	654
		Post	5	77	90	177	5	290	215	515

document first and created the bigrams based on the filtered content. Effectively, their lexicon contains unigrams and *skip-(bi)grams*. Since we wanted to use CrisisLex as another starting point for crisis term acquisition, we created two versions of our corpus: one where filtering was done after bigrams were collected (called *post*) and one that followed Olteanu et al. (2014), called *pre*. Since we don't use location-based corpora, we follow a more conservative term-culling approach: we drop words with a document frequency less than ten, while Olteanu et al. (2014) under 0.5%.

While the unigram statistics are the same, the two corpus variants have widely different bigram distributions. The *pre* setting has as many bigrams as unigrams, 3.9G in 141M types. The *post* variant, where filtering occurs after bigram creation, has about half that number: 1.9G and 61M types. Again, dropping the bigrams associated with the low recall unigrams does not affect the number of tokens much, but decreases the number of types to 119M and 49M, respectively. The *pre* corpus contains 293 bigrams from CrisisLex; as expected, *post* trails behind with 265. Naturally, both contain all 87 unigrams.

While the settings above seemed sensible at first, when applying our method, we have found that they did not reflect the realities of the dataset. The Jaccard similarity threshold of 0.85 leaves too many duplicates in the data, which leads to many expressions that are peculiar to certain news items creep into the results. After some experimentation, we had to use a similarity threshold as low as 0.1 to filter most of the near replicas. Such a low threshold most likely removes many false positives as well, negatively affecting document recall; however, the effect on the resulting emergency terms was markedly positive. We also applied a much bolder term frequency threshold of 100 to eliminate low quality uni- and bigrams from the data.

As anticipated, such an aggressive filtering has a huge effect on the corpus size. The number of unigram types falls to 156,975, and bigrams to 1M/2.7M (*post*/*pre*). The total number of bigrams is also visibly reduced to 1.65G and 3.2G, respectively. Finally, CrisisLex coverage also decreased greatly to 85 unigram and 140/218 bigram types.

4.3 Extending Seed Lexica

To deploy some seed lexicon on this corpus, we combine the expressions (words and n-grams) contained in the lexicon into one large query, and save the top ranking 10,000 documents according to the Okapi-BM25 scoring formula (Spärck-Jones et al. 2000). These documents form the presumably emergency-related subcorpus E. Thereafter, E is used as foreground against the entire CC News corpus C as background to compute the log TF ratios. Terms with too low DF in C or E are filtered out, and the resulting ranked list is cut off at $\Delta \leq 3$ (three natural orders of magnitude). The resulting lists are the iteratively refined lexica built on the seeds in question.

We applied this method to several seed lists, including CrisisLex, the manual (Wikipedia-based) list of Table 1, the 'minimum' seed list (which is actually longer, but ultimately it is based on just two words, see Table 3), and our BEV. The results are summarized in Table 7, and those words and bigrams that appeared in at least three of the four extensions are listed in Appendix D, showing the quality of the results.

As mentioned in Section 4.2, setting the corpus frequency threshold to 100 eliminated one third of the bigrams of CrisisLex from the *pre* corpus, and even more from the *post* variant. This indicates that many of the bigrams in CrisisLex are specific to Twitter and do not translate well to other domains. A few examples are listed below.

- Social-media specific: *donate tornado*, *retweet donate*, *txting redcross*
- Informal: *bombing saddened*, *storm amaze*
- Corpus language: *toxin flood*, *flood levy*

Out of these three, only the last one needs explanation. Many (skip-)bigrams in CrisisLex contain perfectly valid associations, yet are not found (enough times) in CC News. This points to a deeper divergence in the language models of the corpora than what shallow stylistic differences (formal–informal, etc) can explain.

5 Evaluation

It is not trivial to evaluate different lexica automatically. One option is to compare them against an already existing word list such as CrisisLex. Doing so not only allows us to assess our lexica, but it also gives us ideas about the quality of CrisisLex itself. In Section 4.1 we compared our methods to CrisisLex using the ReliefWeb subcorpora, and here we consider the accompanying humanitarian assistance and disaster relief topic lexicon (HADRTL) of 34,500 phrases. HADRTL is dominated by 22,380 bigrams and 7.818 trigrams and higher n-grams, leaving only 3860 unigrams for comparison. Needless to say, neither of these resources were used during any of the processing described so far.

HADRTL contains a Boolean "seed" field that was set by a manual procedure analogous to our Section 2.1 whenever a domain expert considered the term highly relevant to the topic. There are only 435 seeds and, remarkably, none of these are unigrams. In fact, the only unigrams we consider expertise-based in this lexicon are those derived from the names of the 25 topics used there. When these are higher n-grams (e.g. Violent Civil Unrest), we fall back on the headword (unrest). Some of the topic names are so generic (e.g. Volunteer or Professional Services, Money, Food, ...)

Table 7 CrisisLex coverage of the lexica generated by various seed word lists

Word list	Corpus	Unigrams				Bigrams			
		Found	Missing	New	Sum	Found	Missing	New	Sum
BEV	Pre	6	81	108	195	9	291	886	1186
BEV	Post	5	82	115	202	5	295	583	883
CrisisLex	Pre	30	57	60	147	65	235	416	716
CrisisLex	Post	30	57	69	156	41	259	241	541
Minimum	Pre	14	73	160	247	34	266	1250	1550
Minimum	Post	16	71	160	247	25	275	675	975
Wikipedia	Pre	13	73	149	236	26	274	680	980
Wikipedia	Post	13	74	152	239	17	283	380	680

that no reasonably emergency keyword could be derived from them. Altogether, we end up with 17 expert keywords: *cyclone drought earthquake evacuation flood heatwave infestation intervention landslide rescue sanitation shelter terrorism tsunami unrest violence wildfire*. Only three of these, *earthquake*, *shelter*, and *terrorism* appear on initial list (Appendix B), but all 17 are present in the final list (Appendix C), showing considerable improvement between B and C.

Another way to compare the effectiveness of the various vocabularies is by comparing the Δ values (log TF ratios) introduced in Section 2.2. In the denominator (background model) we use frequencies from the entire Common Crawl news corpus, and in the numerator the frequency counts are obtained from the ReliefWeb articles. Filtering out words with $DF < 10$ left a total of 2298 words from HADRTL, the initial, and the final BEV lists put together (condition *all* in Table 8 below). Of these, 60 have negative Δ (examples include *Republican*, *Democrat*, and *backbencher*) and are discarded in condition *pos*. Finally, in Section 2.2 we only considered those unigrams significant whose Δ exceeded the expected zero by at least two natural orders of magnitude, leaving 1665 words (condition *sig*). As is clear from Table 8, the final BEV outperforms the initial one under all conditions.

Recently, Gallagher et al. (2017) studied topic coherence, and automatically derived 50 topics based on the same ReliefWeb corpus. Their method, quite correctly, detects several topics that do not, in and of themselves, constitute emergencies: 10 (taliban); 16 (crops); 17 (medical); 18 (water); 20 (environmental) – all these are ranked higher (in terms of total correlation explained) than the central emergency topic (22). At the lower ranks, the number of non-emergency topic increases: 23 (military); 25 (transport); 26 (basin); 27 (criminal); 28 (public health); 29 (housing); 31 (training); 33 (flour) until we hit the disaster (35) and relief (36) clusters – overall about half of the clusters can be considered emergency-related. Extracting the unigrams from the topic descriptions yields an emergency vocabulary of 353 words that performs even better on ReliefWeb (average $\Delta = 3.813$), which is as expected given that it was developed specifically on ReliefWeb. In the intersection of this vocabulary with our initial BEV we find 32, with the final 46 words, again showing noticeable improvement, in spite of the fact that we held out ReliefWeb and HADRTL until BEV was complete.

Table 8 Intersection with HADRTL and log frequency ratios in the initial and final BEV

Condition → BEV version ↓	*all*		*pos*		*sig*	
	$\cap$	Δ	$\cap$	Δ	$\cap$	Δ
Initial (Appendix B)	329	2.536	325	2.569	221	3.145
Final (Appendix C)	406	2.672	394	2.768	277	3.385

If we select emergency corpora from Common Crawl by the method of Section 2.3, we obtain different results depending on what basic list we employ. Recall that the selection method (binary search among the documents with a window size $W = 10$) relies on the assumption that the Δ weights applied to the TFs will order the documents in decreasing order of emergency-ness. The more coherent the list, the more coherent the ordering, and the more documents will be found. In this regard, HADRTL does not work particularly well, finding only 40,012 emergency documents, while the much shorter initial BEV finds 54,504 and the final BEV 90,010. The corpora found by HADRTL and the initial BEV intersect only in 13,241 documents, whereas the final (90k) corpus contains more than half (23,138) of the documents selected based on HADRTL. All three corpora are available at our website.[13]

We also used the method of Section 2.2 to select emergency-implicated NERs, but have not run a formal evaluation. It is clear that the precision of the system is

[13] http://hlt.bme.hu/en/resources/emergency

Table 9 Keywords and Δ values for Hungarian

Word	Δ	TF	Word	Δ	TF	Word	Δ	TF
Goma	5.19	13	Richter-skála	3.91	34	Megáradt	3.68	16
Lávafolyam	4.95	15	Tűzhányó	3.91	17	Élelmiszercsomag	3.65	11
Ruanda	4.81	12	Földcsuszamlás	3.90	30	Aknamező	3.65	11
Vulkánkitörés	4.53	16	Földrengés	3.85	148	Láva	3.63	39
33-as	4.44	10	Monszun	3.81	14	Előrejelző	3.62	17
Ruandai	4.40	26	Károkozó	3.77	34	Epicentrum	3.56	24
Lőszerraktár	4.25	12	Ítéletidő	3.74	25	Rengés	3.48	52
Evakuál	4.10	21	Megrongálódik	3.74	20	Végigsöprő	3.48	13
Segélyszállítmány	4.10	14	Bozóttűz	3.71	36	Csernobil	3.48	13
Kongói	4.07	36	Nińo	3.71	31	Esőzés	3.47	132
Hóréteg	4.05	11	Hurrikán	3.69	42	Vulkanikus	3.41	14
Segélyszervezet	4.02	57	Tornádó	3.68	16			

reasonably high, even at the bottom of the range we get locations like *Key Biscayne*. To measure recall is much harder, and it would take manual analysis of larger samples to obtain significant figures. Therefore, we decided to validate the basic idea of iteratively bootstrapping the keyword- and the document-set on a different language, Hungarian. We use the MagyarHirlap collection of some 44,000 newspaper articles, and start with only three words, *vészhelyzet* 'emergency situation', *katasztrófa* 'catastrophe', and *áldozat* 'victim, sacrifice'. (Hungarian doesn't have a word that could be used both as a noun and an adjective to denote emergency.)

Based on these words, we found a small document set (170 documents) from which we repeated the process. The resulting wordlist required manual editing, primarily to take care of tokenization artifacts, but the top 35 words already show the same tendency, with several emergency-implicated locations (Goma, Rwanda, Chernobyl) and excellent keywords for a second pass such as *vulkánkitörés* 'volcanic eruption', *evakuál* 'evacuate', or *segélyszállítmány* 'relief supplies' (see Table 9).

There are also entries such as *33-as* '#33' which require local knowledge to understand (there was flooding along route 33 in Hungary at the time) and morphology is a much more serious issue: we see e.g. the locative adjectival form *ruandai* 'of or pertaining to Rwanda' along with the country name.

Although the TF values are really too small for this, we performed another iteration, obtaining a slightly longer document list, and a much longer wordlist, containing many excellent keywords that could not be obtained by translating the BEV to Hungarian, supporting the observation we already made in regards to English, that manual word selection has low recall. In fact, the wordlist we obtained by a dictionary-based translation of BEV had too many elements (over 2200) and was dominated by false positives (valid Hungarian translations that corresponded to some sense of English keywords that were not emergency-related).

In future work, we plan to investigate whether in the context of iterated keyword-weight bootstrap the simple recall-based ranking of selecting and weighing keywords advocated in Kornai et al. (2003) is outperformed by the slightly more complex Bi-Normal Separation method advocated in Forman (2003). Another attractive idea is to use active learning techniques (Hashimoto et al. 2016) as opposed to the simple word vector based filtering we presented in Section 3.

6 Conclusions

Faced with the problem of building a two-way classifier selecting a small class of emergency reports from a much larger set of other (non-emergency) texts, it is tempting to put the emphasis on non-textual features such as the snowballing of reports from the same area. Except for the experiments described in Section 4.1, we considered 'emergency' to be a single topic rather than a combination of smaller and better delineated topics such as Landslide or Tsunami, and assumed that reports coming in later will often have reference not just to the event, but to the response as well.

This assumption is clearly borne out by the vocabularies, not so much by the BEV (which was built by knowledge engineering, with the response assumption already built in), as by the lists built iteratively based on very small seeds (in English, two words, in Hungarian, three words). The first iteration already yields words like English *evacuee* or Hungarian *segélyszervezet* 'aid organization' that only makes sense in the context of some organized response.

We also used the method for the systematic selection of a larger (English) emergency corpus. At 90k articles,

our corpus stands halfway between the OSC corpus (17k language-filtered and deduplicated) and the ReliefWeb corpus (424k documents), but our selection criteria are more strict. For example OSC has articles that begin "Applicants for permanent residence status will have to wait a little longer to find out if they will receive it..."; "Two members of the Moro Islamic Liberation Front (MILF) peace panel who were invited to the International Meeting of Prayer for Peace in Sarajevo in September this year begged off from attending"; "Government defends removal of 'illegal' families from capital"; or "Indonesia's lesbian, gay, bisexual and transgender (LGBT) advocacy movement has come a long way since the 1960s". ReliefWeb is even worse, with less than 10% of the articles dealing with actual emergencies, the majority being devoted to the drawn-out political process dealing with the aftermath.

The key benefit of our proposal is that it only requires a collection of documents, typically easily obtained by web crawl even in less well resourced languages – everything else can be bootstrapped from minimal seeds of 2-3 words. Since we build linear classifiers, the process is linear in the size of the corpus, and does not involve the extensive (sometimes crowdsourced) curation effort that is very much part of other approaches. This is not to say that we completely automated the process, since the hyperparameters (thresholds) are still set manually, but it took manual inspection of less than 200 documents to select a 90k emergency corpus from over 14m Common Crawl documents.

Acknowledgments We thank Stephanie Strassel (LDC) for her support and encouragment, Graham Horwood (Leidos) for preparing some of the data used in the evaluation, and the anonymous referees for valuable suggestions that led to major improvements. Special thanks to Judit Ács who produced the original BV list that was the starting point of the entire work.

Appendix A: $BEV_0 \cap BV$

Dr able accident against alone angry arms army attack bad bite blood blow body bone break breathe burn calm can catch chemical cloud cold concern condition could crime crush damage danger dead destroy die dig drug effort end energy enough equipment escape explode extreme fail fall fault fight fire flesh food force frighten gas grain harm hospital hot hurt ice ill injure level lightning limit mass meal medical necessary offensive organization pain people police powerful problem protect public quick radio rain react report request risk rule safe sea serious shock shoot sick sink situation snow social soldier special speed stop strong surprise temperature tent thick thin travel treatment trouble vehicle violent volcano war weapon weather wind worry wound *area authority care develop event exercise field general heat measure officer plan protection range search skin smoke team waste*

Appendix B: BEV_0

Becquerel Bq Ci Curie Dr able absorb accident acute adverse affect against agency airborne alarm alert alone angry anomaly area arms army asphyxiant assistance assurance atomic attack authority avoid bad barrier bite blast blood blow body bomb bone boundary break breathe buffer burn burning calm cancer carcinogen care catastrophy catch chemical civilian cloud cold combat combustible compromised concentrated concern condition consequence containment contaminate cooling coordinate corrective counterterrorism crime crisis critical crush damage danger dangerous dead debris decay declaration decontaminate defective defense degrade demolition department designated destroy destruction deteriorate develop device diarrhea die dig disaster discharge disease disperse dose dosimeter downgrade drill drug earthquake effort embargo emergency emission end energy enough environment equipment error escape escort evacuate event exceed exclusion exercise explode explosion explosive expose exposure extreme facility fail failure fall fallout fatality fatigue fault field fight filter fire firefighter fission fissionable flammable flashpoint flesh food force frighten fuel fuse gas general grain grenade half-life harm hazard hazmat headquarters health heat hemorrhage herbicide hospital hot hotline hurricane hurt ice ignition ill illness impact inadequate inadvertent incident infect inflammation ingest inhale injure installation ionization issue jettison launch leak lethal level liason lifethreatening lightning limit loss lost malevolent malfunction management mass meal measure medical microorganism mine missile mitigate mobilize monitoring mortality nausea necessary notify nuclear offensive officer offsite operation organization pain parameter people perimeter pesticde plan plume plutonium poison police pollute pose powerful preparedness prevent problem procedure protect protection public quarantine quick rad radiation radio radioactive radiology rain range react reactor recovery reentry release rem report request resolution respiration responder response restoration risk rocket rod rule sabotage safe safeguard safety scenario sea search secure security serious severe shelter shield shock shoot sick sickness sink site situation skin smoke snow social soldier spark special speed spill stabilization staging stolen stop strike strong suffocation supply surprise symptom tank target team temperature tent terrorism thermal thermonuclear thick thin threat tornado toxic toxin travel treatment tritium trouble typhoon unconscious uncontrolled unexpected unintended unintentional

unstable uranium urgent vehicle victim violation violent vital volatile volcano vomiting vulnerability vulnerable war warhead waste weakness weapon weather wind worry wound zone

Appendix C: BEV

absorb accident acute adsorption against agony airlift alarm alert ambulance angry anti-tank antitoxin apocalypse army arsenal arsenic asphyxiate assault assurance atomic attack avoid bad barrage barrier battle biohazard biosafety bioterror bioterrorism bite blast blaze blizzard blockade bloodshed blow body bomb bomber bombing boundary breach break breathe brigade burn calamity cancer cannon care carnage catastrophe catch caustic chain-reaction civilian clot cold collapse combat commander concern condition conflagration constable contagious containment contaminate contamination cop corrode counter-measure counterattack crisis critical crossfire cyclone damage danger dangerous dead debilitate debris decay decontamination defibrillator defoliant degrade dehydrate despoil destroy detonate detonator device die disaster disaster-related disease disinfect disinfection disintegrate doctor drinking-water drought drown drug dynamite earthquake earthquake-prone effort embolism emergency emission end endanger enough environ epidemic epidemiologic erupt escape evacuate evacuation excrete explode explosion explosive exposure extinguish extreme fail failure fall fallout famine fatal fatality fatigue fault fight fighter fire fireball firefighter fireman firepower firestorm first-responder fissile fission flame flammable flare-up flash flesh flood flood-affected flooding force freezing fuel fuse gas general grave grenade gunner gunshot gust hail harm hazard hazardous health heat heatstroke heatwave high-priority hijack hit-and-run horrible hospital hot hurricane hurt hypothermia ignition ill ill-effects illness impact implode incapacitate incendiary incident incision inclement infantry infantryman infect infected infectious inferno infestation infirmary inflammable infraction inhale injure intervention invasion irradiate issue just-in-time kill landslide lava leak lethal life-saving life-threatening lifeless lift-off limit magnitude maim mass-casualty meal medic medical megaton meltdown military militia mine misfire mishap missile mitigate mitigation mobilize monsoon napalm navy neurotoxicity nuclear nuke nurse offensive officer operation ordnance outbreak pandemic panic parachute paramedic patrolman peril perish phosgene plague plume plutonium poison poisoning police policeman post-disaster post-earthquake postdisaster powerful precarious preparedness procedure projectile protect protection putrefy quake quake-hit quarantine racial-ethnic radiation radioactive radioactivity radiological radiology rain rainstorm range react reactor rebel rebuild recoil reconstruction recovery refugee rescue responder riot risk rocket rule safe safeguard safety sanctuary sanitation scorch secure security seismic sergeant serious severe shatter shelter shelter-in-place shield shoot shooting shot shotgun shrapnel sick sicken sickness siren skin sleet smash smoke snow snowfall snowstorm soldier spark spill squad stop stoppage stormy strangle stricken strike strong subdue suffocate surgeon surprise survive symptom tank temperature terror terrorism terrorize thermal thermonuclear thick thin threat thrombosis tinderbox toll toll-free tornado torpedo toxicant toxicity toxicological toxicologist toxicology traumatize treatment tremor triage troop trooper trouble tsunami unavailability uncontrollable uncoordinated unprotected unresponsiveness unrest unsafe unstable uranium venom victim violate violation violence violent violently virulent volatile volcano vulnerable waft war war-related warfare warhead waste weapon weather wide-spread wildfire wind windstorm worry wreckage

Appendix D: Terms Occurring in at Least Three of the Four List Increments

aerial.view affect.area affect.community affected.area affect.flood area.affect area.flood area.hit assess.damage bring.heavy burst.bank cause.flood cause.flooding cause.heavy cause.widespread cholera civil.defence civil.protection clear.debris clear.road coastal.area cyclone damage.destroy damage.house damage.report day.heavy death.toll debris deluge destroy.damage destroy.home devastation disaster disaster.agency disaster.area disaster.declaration disaster.management disaster.occur disaster.relief disaster.response disaster.risk disaster.say disaster.strike disaster.zone dozen.home due.flood due.flooding due.heavy emergency.management emergency.response emergency.worker evacuate evacuate.area evacuate.home evacuate.people evacuate.resident evacuation evacuation.order evacuation.plan evacuee extent.damage federal.disaster flash.flood flash.flooding flood flood.area flood.cause flood.damage flood.hit flood.home flooding flooding.cause flooding.landslide flood.kill flood.landslide flood.road flood.stage flood.street flood.victim flood.warning floodwater flood.water force.evacuation hardest-hit heavy.downpour heavy.rain heavy.rainfall higher.ground hit.area home.damage home.destroy home.flood house.damage house.destroy hundred.home hurricane inch.rain landslide leave.homeless low-lying low-lying.area major.disaster major.flooding management.official massive.landslide meteorologist monsoon.rain mud mud.debris mudslide national.disaster natural.disaster next.environment people.evacuate

people.miss people.strand people.trap plume powerful.earthquake powerful.storm rain.cause rain.expect rainfall rain.fall rain.trigger relief.effort relief.operation relief.supplies rescue.effort rescue.operation rescue.team resident.evacuate resident.flee response.team rise.water river.burst river.overflow rubble say.disaster say.flood say.rain say.storm seek.shelter severe.flooding severe.storm severe.thunderstorm severe.weather significant.damage storm.also storm.cause storm.pass storm.surge storm.system sweep.away take.shelter temporary.shelter thousand.home toll.could toll.rise tornado tornado.warning torrent torrential torrential.rain tree.power trigger.landslide tropical.cyclone twister typhoon uproot.tree wash.away water.recede water.rise weather.service widespread.damage widespread.flooding wildfire wildfire.burn worst.affect worst.flooding worst.hit worst-hit.

References

Ács, J., & Kornai, A. (2016). Evaluating embeddings on dictionary-based similarity. In Levy, O. (Ed.) *Proceedings of the first workshop on evaluating vector-space representations for NLP (RepEval)* (pp. 78–82).

Ács, J., Pajkossy, K., Kornai, A. (2013). Building basic vocabulary across 40 languages. In *Proceedings of the sixth workshop on building and using comparable corpora* (pp. 52–58). Sofia: Association for Computational Linguistics.

Basu, M., Roy, A., Ghosh, K., Bandyopadhyay, S., Ghosh, S. (2017). Microblog retrieval in a disaster situation: a new test collection for evaluation. In Moens, M.F., Jones, G., Ghosh, S., Ganguly, D., Chakraborty, T., Ghosh, K. (Eds.) *Proceedings of SMERP 2017*.

Borbély, G., Makrai, M., Nemeskey, D.M., Kornai, A. (2016). Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation. In *Proceedings of the 1st workshop on evaluating vector-space representations for NLP* (pp. 83–89). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2515. http://www.aclweb.org/anthology/W16-2515.

Broder, A.Z., Charikar, M., Frieze, A.M., Mitzenmacher, M. (1998). Min-wise independent permutations. In *Proceedings of the thirtieth annual ACM symposium on theory of computing* (pp. 327–336). ACM.

Buckley, C., Singhal, A., Mita, M. (1995). New retrieval approaches using SMART: TREC 4. In *Proceedings of TREC*, (Vol. 4 pp. 25–48).

Ester, M., Kriegel, H.P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, (Vol. 96 pp. 226–231).

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

Gallagher, R., Reing, K., Kale, D., Steeg, G.V. (2017). Anchored correlation explanation: topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, *5*, 529–542. https://transacl.org/ojs/index.php/tacl/article/view/1244.

Gionis, A., Indyk, P., Motwani, R., et al. (1999). Similarity search in high dimensions via hashing. In *VLDB*, (Vol. 6 pp. 518–529).

Han, L., Kashyap, A.L., Finin, T., Mayfield, J., Weese, J. (2013). UMBC_EBIQUITY-CORE: semantic textual similarity systems. In *Second joint conference on lexical and computational semantics (*SEM)* (pp. 44–52). Atlanta: Association for Computational Linguistics.

Hashimoto, K., Kontonatsios, G., Miwa, M., Ananiadou, S. (2016). Topic detection using paragraph vectors to support active learning in systematic reviews. *Journal of Biomedical Informatics*, *62*, 59–65.

Imran, M. (2017). Time-critical analysis of evolving social media streams during sudden-onset events. In Moens, M.F., Jones, G., Ghosh, S., Ganguly, D., Chakraborty, T., Ghosh, K. (Eds.) *Proceedings of SMERP 2017*.

Imran, M., Castillo, C., Lucas, J., Meier, P., Vieweg, S. (2014). AIDR: artificial intelligence for disaster response. In *Proceedings of WWW (companion) IW3C2* (pp. 159–162).

Imran, M., Castillo, C., Diaz, F., Vieweg, S. (2015). Processing social media messages in mass emergency: a survey. *ACM Computing Surveys (CSUR)*, *47*, 1–38.

Indyk, P., & Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on theory of computing* (pp. 604–613). ACM.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, *20*(4), 422–446.

Kohlschütter, C., Fankhauser, P., Nejdl, W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 441–450). ACM.

Kornai, A. (2010). The algebra of lexical semantics. In Ebert, C., Jäger, G., Michaelis, J. (Eds.) *Proceedings of the 11th mathematics of language workshop, LNAI 6149* (pp. 174–199). Springer.

Kornai, A. (2018). *Semantics*. Springer. http://kornai.com/Drafts/sem.pdf.

Kornai, A., Krellenstein, M., Mulligan, M., Twomey, D., Veress, F., Wysoker, A. (2003). Classifying the Hungarian Web. In Copestake, A., & Hajic, J. (Eds.) *Proceedings of the EACL* (pp. 203–210).

Lewis, D., Yang, Y., Rose, T., Li, F. (2004). RCV1: a new benchmark collection for text categorization research. *Journal of Machine Learning Research*, *5*, 361–397.

Lui, M., & Baldwin, T. (2012). langid.py: an off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations* (pp. 25–30). Association for Computational Linguistics.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for computational linguistics (ACL) system demonstrations* (pp. 55–60). http://www.aclweb.org/anthology/P14-5010.

Olteanu, A., Castillo, C., Diaz, F., Vieweg, S. (2014). CrisisLex: a lexicon for collecting and filtering microblogged communications in crises. In *Proceedings of the AAAI conference on weblogs and social media (ICWSM'14)*. AAAI Press.

Pennington, J., Socher, R., Manning, C. (2014). Glove: global vectors for word representation. In *Conference on empirical methods in natural language processing (EMNLP 2014)*.

Phuvipadawat, S., & Murata, T. (2011). Detecting a multi-level content similarity from microblogs based on community structures and named entities. *Journal of Emerging Technologies in Web Intelligence*, *3*(1), 11–19.

Soni, R., & Pal, S. (2017). Microblog retrieval for disaster relief: how to create ground truths? In Moens, M.F., Jones, G., Ghosh, S., Ganguly, D., Chakraborty, T., Ghosh, K. (Eds.) *Proceedings of SMERP 2017*.

Spärck-Jones, K., Walker, S., Robertson, S.E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing & Management*, *36*(6), 779–840.

Strassel, S., Bies, A., Tracey, J. (2017). Situational awareness for low resource languages: the LORELEI situation frame annotation task. In Moens, M.F., Jones, G., Ghosh, S., Ganguly, D., Chakraborty, T., Ghosh, K. (Eds.) *Proceedings of SMERP 2017*.

TRADE Emergency Management Issues SIG Glossary Task Force (1999). *Glossary and acronyms of emergency management terms*, 3rd edn. Office of Emergency Management, U.S. Department of Energy.

Dávid Márk Nemeskey is research associate at the Hungarian Academy of Sciences Institute of Computer Science and a PhD student at the Loránd Eötvös University Faculty of Informatics doctoral programme, where his advisors are András Benczúr and András Kornai. He has published work on a variety of information retrieval and natural language processing problems including recognizing textual entailment, named entity recognition, and multi-modal information retrieval. Currently he is working on statistical language modeling. His homepage is at http://hlt.bme.hu/en/david.

András Kornai is senior scientific advisor at the HAS Institute of Computer Science and a professor at the Department of Algebra, Budapest University of Technology and Economics. He has published extensively on all aspects of mathematical linguistics including natural language processing, optical character recognition, speech recognition, signal processing, formal reasoning, information retrieval, information extraction, and statistical data analysis. His latest monograph, Semantics, is forthcoming from Springer in 2018. His homepage is at http://kornai.com.