Identification of Disaster-implicated Named Entities

Judit Ács[1,2], Dávid Márk Nemeskey[2], and András Kornai[2]

[1] Department of Automation and Applied Informatics, Budapest University of Technology and Economics
[2] Computer Science and Automation Research Institute, Hungarian Academy of Sciences
judit@aut.bme.hu, {ndavid,kornai}@sztaki.hu

Abstract. For disaster preparedness, a key aspect of the work is the identification, *ahead of time*, of potentially implicated locations (LOC), organizations (ORG), and persons (PER). Here we describe how static repositories of traditional news reports can be rapidly exploited to yield disaster- or accident-implicated named entities.

1 Introduction

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

Section 2 describes the current state of the Basic Emergency Vocabulary (BEV) that can be used as the basis of the classifiers we use to select emergency-related material in English and other languages. Section 3 outlines the main method of collecting emergency-implicated NERs, Section 4 describes an ad hoc evaluation, and Section 5 offers some conclusions.

2 Basic emergency vocabulary

The idea that there is a *basic* vocabulary composed of a few hundred or at most a few thousand elements goes back to the Renaissance – for a more detailed history see [1], for a contemporary list see [6]. The emergency vocabulary 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/newswire articles, situation reports, 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*[3] as 'chemical causing suffocation', we need to define *suffocation*, but not *chemical* as this item is 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. Proper nouns like *Jesus* are discussed further in [5], but we note here that they constitute a very small proportion (less than 6%) of the basic vocabulary. None of these basic entities, whose list is restricted to names of continents, countries, major cities, founders of religions, etc., are particularly implicated in disasters or accidents, so our method (for a theoretical justification see [7]) involves no seeding for the actual entity categories we wish to learn – our seed lists, minimal as they are, contain only common nouns, verbs, adjectives and adverbs.

At 1,200 items, the basic list was 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.

It is evident that many emergency words are not basic: examples include *jettisoning* and *typhoon*. To obtain a good sample, we analyzed the the glossary [2] 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

[3] One issue, not addressed in this paper, is reducing the morphological variability of these words: for example *asphyxiant* and *asphyxiate, radiological* and *radiology*, etc. shouldn't both appear in the final list. On the whole, we chose the shorter of competing forms, or when they had equal character length, as in *injure* and *injury* we chose the verbal form.

build a language-independent set of concepts, not something specific to American English. The Basic Emergency Vocabulary (see Appendix B) contains 350 emergency-related concepts – none of these are proper names.

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, *area authority care develop event exercise field general heat measure officer plan protection range search skin smoke team waste*, should also have been selected based on the above principles, especially the last (fifth) one.

The lesson from this is clear: the 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. In the next Section we describe a method of jointly bootstrapping the list and the emergency corpus.

3 Finding the NERs involved

Using the BEV as positive evidence [7], it is possible to select a small, emergency-related subset of a given corpus C of articles (we used 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 the BEV 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 D documents with a window size W, N can be found by looking at only $W \log_2(D)$ documents – in our case we only had to look at 80 documents of the entire corpus to select a core set E of about 2,000 emergency-related articles.

This is a noisy sample, only about 80% of the documents in it are actually emergency related, and we estimate 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 it does not unduly affect the logic of our enterprise: since a random document in the corpus C will be emergency related with probability $p = 0.0025$, 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, and looked at those 1,700 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 12 (135).

Since we have only 2k relatively short documents to consider, we ran the NER system from Stanford CoreNLP on these, and collected the results for all 1,700 words. Typically, words are classified unambiguously (label entropy is below 0.1

for over 82%), and by ignoring the rest we still obtain 1,398 words. The resulting file begins as follows:

word	Δ	DF	NER
hohenwutzen	5.53	110	LOC:110
platzeck	5.52	59	PER:57
slubice	5.50	90	LOC:88
oderbruch	5.50	196	LOC:189
dike	5.49	993	O:999
sandbag	5.45	362	O:334
oder	5.39	542	LOC:81,MISC:1,O:149,PER:272
popocatepetl	5.39	54	LOC:29,O:1,ORG:1,PER:23
pomes	5.30	78	PER:68
stolpe	5.29	57	PER:45
levee	5.28	229	O:188,ORG:1
floodwater	5.27	437	O:366
soufriere	5.23	62	LOC:48,O:5,PER:1
abancay	5.23	51	LOC:39
opole	5.21	105	LOC:79,O:2
forks	5.17	409	LOC:176,MISC:110,O:5,ORG:24
montserrat	5.15	268	LOC:267,ORG:18
hortense	5.13	399	LOC:1,MISC:1,O:56,ORG:4,PER:236
low-lying	5.13	299	O:214
flood-ravaged	5.10	74	MISC:1,O:57
nirmala	5.09	122	PER:87
godavari	5.08	128	LOC:78,O:12
falmouth	5.07	60	LOC:37
sodden	5.06	67	O:44
eruption	5.01	537	O:339
volcano	4.99	870	LOC:7,O:616,ORG:20,PER:1
hurricane-force	4.98	53	O:33
flood-stricken	4.98	62	O:40
evacuee	4.93	291	O:184
flood-hit	4.93	123	MISC:1,O:111
jarrell	4.92	81	LOC:19,O:5,ORG:4,PER:21
yosemite	4.92	100	LOC:52,O:3,ORG:1
bandarban	4.91	53	LOC:28
lava	4.90	136	O:76
mudslide	4.90	458	O:295

Two-thirds of the words in the list are emergency-related common nouns (e.g. *levee*, *floodwater*, *mudslide*). This number is so significant that we could in fact dispense with the BEV, and bootstrap the classifier starting with 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'. While the top of the list of words that are significantly more frequent in E' than in the background are not

quite as good as the actual BEV listed in Appendix B (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 practically identical, a matter we shall investigate more formally in Section 4 for Hungarian, and so are the lists of NERs.

To summarize what we have so far, we propose to identify 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 hundreds of thousands of locations in NewReuters, 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'.

4 Evaluation

It is clear that the precision of the system is reasonably high, even at the bottom of the range we get locations like *Key Biscayne* and good classifier words like *around-the-clock*. 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, *katasztrófa* 'catastrophy', *vészhelyzet* 'emergency situation', and *áldozat* 'victim, sacrifice'. (Hungarian doesn't have a word that could be used both as a noun and an adjective to denote emergency.)

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	niño	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			

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'. 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 2,200) and was dominated by false positives (valid Hungarian translations that corresponded to some sense of English keywords that were not emergency-related).

5 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. Here we considered 'emergency' to be a topic on a par with 'sports' or 'computer science' or any other topic in a well-established topic hierarchy, 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.

In the near future we plan to use the method for a systematic selection of a far larger emergency corpus (since we build linear classifiers, the time to do it is linear in the size of the corpus), with the expectation that not more than a quarter of a percent of the material in a static news corpus will be selected. Once the corpus is at hand, we can use standard NER techniques to designate persons, organizations, and locations as emergency-implicated. 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 [7] is outperformed by the slightly more complex Bi-Normal Separation method advocated in [3].

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. In better resourced languages, the iterative keyword selection method used here

can be compared to one based on word vectors [4] or we can hybridize the two – we leave this for further research.

Acknowledgment

We thank referee #2 for calling [3] to our attention.

References

1. Ács, J., Pajkossy, K., Kornai, A.: Building basic vocabulary across 40 languages. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora. pp. 52–58. Association for Computational Linguistics, Sofia, Bulgaria (2013)
2. Force, T.E.M.I.S.G.T.: Glossary and acronyms of emergency management terms. Office of Emergency Management, U.S. Department of Energy, third edn. (1999)
3. Forman, G.: An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3, 1289–1305 (2003)
4. Hashimoto, K., Kontonatsios, G., Miwa, M., Ananiadou, S.: Topic detection using paragraph vectors to support active learning in systematic reviews. Journal of Biomedical Informatics 62, 59–65 (2016)
5. Kornai, A.: The algebra of lexical semantics. In: Ebert, C., Jäger, G., Michaelis, J. (eds.) Proceedings of the 11th Mathematics of Language Workshop, pp. 174–199. LNAI 6149, Springer (2010)
6. Kornai, A.: Semantics. Springer Verlag (in press), http://kornai.com/Drafts/sem.pdf
7. Kornai, A., Krellenstein, M., Mulligan, M., Twomey, D., Veress, F., Wysoker, A.: Classifying the Hungarian Web. In: Copestake, A., Hajic, J. (eds.) Proceedings of the EACL. pp. 203–210 (2003)

Appendix A

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

Appendix B

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 can cancer carcinogen care catastrophy catch chemical civilian cloud cold combat

combustible compromised concentrated concern condition consequence containment contaminate cooling coordinate corrective could 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 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