A Study in Domain-Independent Information Extraction for Disaster Management

Lars Döhling, Jirka Lewandowski, Ulf Leser

Department of Computer Science, Humboldt-Universität zu Berlin
Unter den Linden 6, 10099 Berlin, Germany
doehling, lewandow, leser@informatik.hu-berlin.de

Abstract

During and after natural disasters, detailed information about their impact is a key for successful relief operations. In the 21st century, such information can be found on the Web, traditionally provided by news agencies and recently through social media by affected people themselves. Manual information acquisition from such texts requires ongoing reading and analyzing, a costly process with very limited scalability. Automatic extraction offers fast information acquisition, but usually requires specifically trained extraction models based on annotated data. Due to changes in the language used, switching domains like from earthquake to flood requires training a new model in many approaches. Retraining in turn demands annotated data for the new domain. In this work, we study the cross-domain robustness of models for extracting casualty numbers from disaster reports. Our models are based on dictionaries, regular expressions, and patterns in dependency graphs. We provide an evaluation on extraction robustness across two disaster types – earthquakes and floods. It shows that applying extra-domain models without retraining gives a relative F1 decrease of solely 9 %. This is a fairly small drop compared to previous results for similar complex extraction tasks.

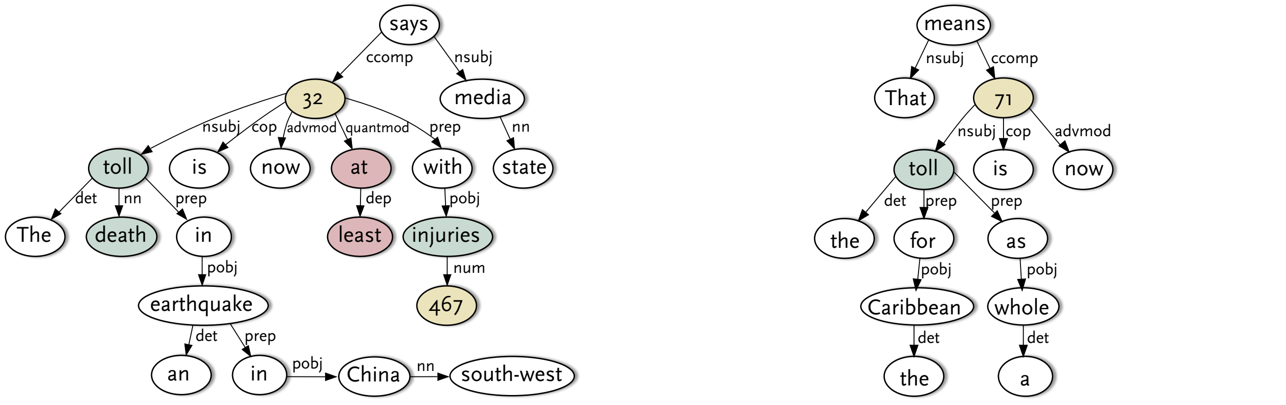
Keywords: natural language processing, cross-domain learning, n-ary relationship extraction, dependency patterns

1. Introduction

Crisis events like earthquakes or disease outbreaks are striking humankind regularly. In the aftermath, decision makers require precise and timely information to assess damages and to coordinate relief operations (Guha-Sapi and Lechat, 1986). Understanding "the big picture" in emergency situations is obviously essential for effective responses. Today, the Internet plays an important role for information acquisition, especially if no on-site contact is available. Supportive information are published both in conventional sources like newspapers (Döhling and Leser, 2011) as well as in social media, e.g. web forums (Qu et al., 2009) or microblogs (Vieweg et al., 2010). These sources offer the most details available, but searching and analyzing them manually is a time-consuming and therefore costly task.

Information Extraction (IE) studies the problem of automatically extracting structured information from given unstructured text (Sarawagi, 2008). By offering fast and on-line information acquisition, IE methods may help to mitigate disaster effects. Developing IE applications typically involves supervised learning, i.e. requires annotated training data to generate and test extraction models. Examples are CRFs for entity recognition (McCallum, 2003) or SVMs for classification tasks (Cortes and Vapnik, 1995). To achieve high quality results, such training data must be sourced from the same domain as the extraction will be applied to. Here, the term 'domain' refers to the texts used, especially their type (news article, tweet) and topic (earthquake, flood). As a consequence, generated models from these training data are domain-specific as well. While achieving optimal results in the original domain, they often perform poorly when applied to different, even closely related domains. For instance, (McClosky, 2010) evaluated the cross performance of PCFG-based parsing models for corpora from diverse domains. Syntactical sentence parsing is a prerequisite for many state-of-the-art IE methods, including the one presented in this paper. For closely related domains, the F1 score decreased relatively by 3 %, while for more distant domains it dropped by 10 %. (Jakob and Gurevych, 2010) studied recognizing opinion targets (what the opinion is about) in user-generated reviews. They observed a relative F1 decrease of 12 % on average when applying CRF-based models across four topics. (Tikk et al., 2010) analyzed an even more complex IE task, (binary) relationship extraction. They measured the cross-corpus performance of SVM-based models for extracting protein-protein interactions. Although all corpora consisted of biomedical texts, their experiments revealed a relative F1 decrease of 24 % on average. To prevent such performance losses, applying extraction methods in new domains mostly requires retraining appropriate models. Retraining in turn demands new annotated data and annotating is an expensive and cumbersome manual task. An alternative approach is to use extraction methods based on robust models performing well across domains.

In this paper, we present such models for extract casualty numbers from disaster reports in multiple domains. Casualty numbers are an indicator for the scale of damage, determining the required extend of relief operations. Our extraction models (Section 2) are based on dictionaries, regular expressions, and patterns in dependency graphs. Cross-domain evaluation results for earthquake and flood reports are given in Section 3 and discussed in Section 4.



(a) ’The death toll in an earthquake in south-west China is now at least 32, with 467 injuries, state media says.’[1]

(b) ’That means the toll for the Caribbean as a whole is now 71.’[2]

Figure 1: Dependency graphs in the (a) earthquake and (b) flood domain. All relevant entities are colored.

2. Information Extraction

The information extraction procedure is based on our method presented in (Döhling and Leser, 2011). It allows to extract arbitrary facts from texts, formalized as n-ary relationships. Here, n denotes the number of entities – single words or word groups – used to express the fact. For instance, the sentence ’The death toll [. . .] is now at least 32, with 467 injuries [. . .]’[1] contains two facts: $\geq$ 32 killed and 467 injured. We formalize these facts as 4-tuples [modifier, quantity, subject, type], resulting in [at least, 32, –, death toll] and [–, 467, –, injuries]. Each 4-tuple is defined by:

- Modifier: modifies quantity values, e.g. ’at least’, ’about’, or ’more than’
- Quantity: numbers casualties and consists of two subtypes: cardinal (’12’, ’ten’, ’no’, ’a’) and vague (’many’, ’hundreds’, ’some’)
- Subject: characterizes casualties explicitly, e.g. ’people’, ’villagers’, or ’students’
- Type: describes the type of damage and consists of multiple subtypes, e.g. injured or trapped

2.1. Extraction Pipeline and Model

The automatic fact extraction consists of three steps. First, we recognize all entities (relevant words or word groups) of the targeted relationship, e.g. ’32’, ’at least’, or ’injuries’. Cardinal quantities are recognized by a domain-independent regular expression, all other entities by a dictionary derived from training data. In contrast to more sophisticated extraction models, e.g. CRFs, dictionaries carry very little contextual information. Consequently, they are potentially more robust when applied across domains. In addition to (Döhling and Leser, 2011), we enhance this step by two optional post filters for cardinal quantities: M-Filter and A-Filter. As the regular expression identifying cardinals does not encode the context, the M-Filter removes potentially false positives by revoking those surrounded by units of measurement, e.g. ’ft’, ’km’, ’$’, or ’%’. The A-Filter withdraws all ’a’/’an’ annotations, as the majority of these terms refer to the indefinite article and not to the cardinal 1, resulting in many false positives. Next, we infer semantic relationships between pairs of entities by matching patterns in dependency graphs. Dependency graphs (Figure 1) model the syntactical relationships between the words of a sentence as typed, directed edges between them. By offering direct access to sentence structures, they often reveal relations between words far apart more easily than if modeled as lists of words (Fundel et al., 2007). Given the example in Figure 1a, the distance on the surface level between the related entities ’death toll’ and ’32’ is ten words, whereas they are directly connected in the corresponding dependency representation. We use the shortest paths between two entities as patterns, collected from training data. (Bunescu and Mooney, 2005) showed that these paths are well suited to capture relationships between entities within sentences. Similar to dictionary entries, shortest paths carry minimal contextual information, again supporting cross-domain robustness. For instance, although Figure 1a’s sentence contains ’earthquake’, this domain-specific keyword is not part of the shortest paths (Figure 2). In addition, the shortest path in Figure 1b is the same as in Figure 1a, emphasizing potential domain-independence of patterns. The pattern matching is adjustable by optionally ignoring the dependency type or direction, or the entity subtype. Ignoring the entity subtype originates from the observation that subtypes are often interchangeable within sentences, e.g. ’injured 13 people’ vs. ’buried many people’.

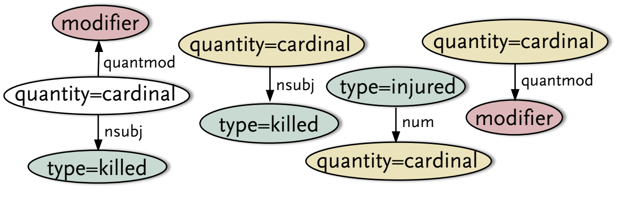


Figure 2: Shortest paths patterns, derived from Figure 1’s dependency graphs.

[1] news.bbc.co.uk/2/hi/asia-pacific/7591152.stm

[2] au.news.yahoo.com/world/a/15270957/haiti-storm/

The results of the pattern matching step are entity graphs with edges between all pairs of related entities (Figure 3). By finding maximal cliques (McDonald et al., 2005) in these entity graphs, the binary relationships are finally merged into tuples of the desired n-ary relationship. In this paper, that is the 4-ary relationship modeling reported casualties. Compared to (Döhling and Leser, 2011), we enhance the last step by an optional, domain-independent post filter for handling enumerations of facts within one sentence. Given '[. . .] killed X and injured Y [. . .]', our method also extracts false tuples like [–, X, –, injured] due to similar dependency structures compared to true tuples. As each quantity belongs to only one type, the Enum-Filter investigates all tuples sharing the same quantity and keeps only the most probable one. Its decision is based on the sentence's token sequence level. It considers distances between entities as well as linguistic hints, such as 'and'.

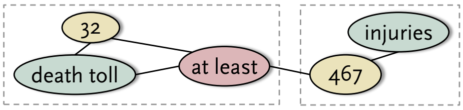


Figure 3: An entity graph, derived from Figure 1a's dependency graph. The rectangles mark all contained maximal and valid cliques, i.e. having one type entity.

3. Evaluation and Results

We evaluated the domain independence of acquired extraction models by comparing their intra-domain performance against cross-domain results. Each model consists of (1) the entity dictionary, (2) the dependency patterns, and (3) the pipeline configuration, i.e. matching and filter switches.

3.1. Data sets

We used two data sets, consisting of news articles collected from the web reporting on earthquakes and floods, respectively (Table 1). The earthquake articles were sourced from BBC and Yahoo! News in 2009/10. The flood articles were selected from various search engine results in 2012. Each article was manually annotated with the 4-ary relationship, covering six casualty types: injured, killed, homeless, affected, missing, and trapped. Both corpora are available on request. We also examined the inter-annotator agreement on corpus samples, which was 82 % on average. We partitioned each corpus into a training (2/3) and an evaluation set (1/3) by stratified random sampling on the sentence level.

	Earthquake	Flood
Documents	245	412
Sentences	4795	8616
Tokens	100 303	187 894
Relationship tuples	1277	1860
size=2	483 (38 %)	570 (31 %)
=3	507 (40 %)	900 (48 %)
=4	287 (22 %)	390 (21 %)
type=killed	825 (65 %)	1362 (73 %)
=injured	224 (17 %)	63 (4 %)
=trapped	74 (6 %)	7 (0 %)
=missing	81 (6 %)	166 (9 %)
=homeless	49 (4 %)	88 (5 %)
=affected	24 (2 %)	174 (9 %)

Table 1: Corpus statistics; size refers the number of set entities within tuples.

Parameter	Earthquake	Flood
M-Filter	enabled	
A-Filter	enabled	disabled
Ignore dependency type	true	
Ignore dependency direction	false	
Ignore entity subtype	true	
Enum-Filter	enabled	

Table 2: Best extraction pipeline configurations per corpus, F1-optimized at the final relationship level.

3.2. Experiments

For the intra-domain experiments (source=target), the models were trained on the training set and evaluated on the evaluation set within the same domain. For the cross-domain experiments (source≠target), the models were trained on the extra-domain training set and evaluated on the evaluation set. permitting fair comparisons. The extraction configuration was derived by 5-fold cross-validation on the respective training set, maximizing the average F1 score (Table 2). Our experiments showed a relative F1 decrease of 9.0 % on average (geometric mean) if applying models across domains (Table 3, top). Both recall (−12.6 %) as well as precision (−4.5 %) declined.

	Target	
Source	Earthquake	Flood
Earthquake	78/317/114 .803/.735/**.768**	153/437/202 .741/.684/**.711**
Flood	94/285/146 .752/.661/**.704**	159/518/121 .765/.811/**.787**
Enhanced		
Earthquake	117/333/98 .740/.773/**.756**	–
Flood	–	221/526/113 .704/.823/**.759**

Table 3: Intra-/cross-domain evaluation results (top) and enhanced intra-domain results (bottom); Numbers are false positives/true positives/false negatives and precision/recall/F1 at the final relationship level.

Encouraged by the low decrease in performance, we further analyzed the potential benefit of adding extra-domain data in the intra-domain setting. For each domain, we enhanced the intra-domain training set by including the complete extra-domain data. We kept the extraction configuration. The resulting mixed-domain models were evaluated on the intra-domain evaluation set as before. Both domains showed a small relative increase in recall (+2.8 %), but a significant decrease in precision (−7.9 %) (Table 3, bottom). The resulting F1 scores were slightly lower than those without extra-domain data (−2.3 %).

4. Discussion and Conclusion

We studied the cross-domain robustness of models for extracting casualty numbers from disaster reports. Our evaluation showed that applying models across disaster types results in only 9 % F1 relative performance decrease. This is a small drop compared to, for instance, the 24 % observed for extracting protein-protein interactions (Tikk et al., 2010), a similar complex extraction task.

By checking the trained models and comparing the underlying data sets, we identified two main reasons for the observed domain independence. Both are connected to each other and equally important. (1) Sentences reporting on casualties use similar structures and wordings to express facts, independent of the domain. (2) Most entries of the acquired entity dictionaries and pattern catalogues are comprised of domain-unspecific words. The dictionaries overlap by approximately 44 % of their entries. Of all entries, only about 3 % are domain-specific, e.g. 'quake toll' or 'drownings'. These figures can be observed in the patterns as well, having an overlap of around 32 % and a domain specificity of roughly 4 %.

We further investigated the potential positive effect of additional extra-domain data on the intra-domain performance. Due to larger dictionaries and pattern catalogues, we observed a slight increase in recall at the cost of a clear decrease in precision. So extra-domain data introduced more false than true positives, favoring single-domain over mixed-domain models.

5. Acknowledgements

We kindly thank Christoph Fischer for contributing to the annotations. We also thank Mariana Neves for providing valuable feedback on a draft of this article.

6. References

R.C. Bunescu and R.J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *HLT/EMNLP 2005.*

C. Cortes and V. Vapnik. 1995. Support-Vector Networks. *Machine Learning*, 20(3).

L. Döhling and U. Leser. 2011. EquatorNLP: Pattern-based Information Extraction for Disaster Response. In *Terra Cognita 2011.*

K. Fundel, R. Küffner, and R. Zimmer. 2007. RelEx - Relation extraction using dependency parse trees. *Bioinformatics*, 23(3).

D. Guha-Sapi and M.F. Lechat. 1986. Information systems and needs assessment in natural disasters: An approach for better disaster relief management. *Disasters*, 10(3).

N. Jakob and I. Gurevych. 2010. Extracting Opinion Targets in a Single- and Cross-domain Setting with Conditional Random Fields. In *EMNLP '10.*

A. McCallum. 2003. Efficiently Inducing Features of Conditional Random Fields. In *UAI '03.*

D. McClosky. 2010. *Any Domain Parsing: Automatic Domain Adapatation for Parsing.* Ph.D. thesis, Brown.

R. McDonald, F. Pereira, S. Kulick, S. Winters, Y. Jin, and P. White. 2005. Simple algorithms for complex relation extraction with applications to biomedical IE. In *ACL '05.*

Y. Qu, P.F. Wu, and X. Wang. 2009. Online Community Response to Major Disaster: A Study of Tianya Forum in the 2008 Sichuan Earthq. In *HICSS'09.*

S. Sarawagi. 2008. Information Extraction. *Foundations and trends in databases*, 1(3).

D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, and U. Leser. 2010. A Comprehensive Benchmark of Kernel Methods to Extract Protein-Protein Interactions from Literature. *PLoS Comput Biol*, 6(7).

S. Vieweg, A.L. Hughes, K. Starbird, and L. Palen. 2010. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. In *CHI 2010.*