Searching for Concepts in Natural Language Part of Fire Service Reports*

Kamil Bąk[1], Adam Krasuski[1], and Marcin Szczuka[2]

[1] Chair of Computer Science, The Main School of Fire Service, Poland
[2] Institute of Mathematics, The University of Warsaw, Poland
krasuski@inf.sgsp.edu.pl,szczuka@mimuw.edu.pl

Abstract. In the article we present the comparison of the information retrieval approaches focused on a searching of specific concepts in a Natural Language part of Fire Service reports. The comparison comprise of searching with use of regular expressions, Latent Semantic Indexing and pre-defined set of search terms. As a case study we selected three concepts which may not be explicitly mentioned in reports, have various meanings (homonymy), or may be replaced by synonyms.
Keywords: natural language processing, information retrieval, search, fire service.

1 Introduction

The Public Security Services in any country are charged with maintaining public safety and emergency assistance. In Poland a large part of public security and safety tasks is the responsibility of the State Fire Service (PSP – from *Państwowa Straż Pożarna* in Polish). As a primary emergency response service the PSP not only deals with fires, but is also charged with technical rescue (e.g., during road collisions, building collapses), chemical emergency response (chemical spills, hazardous material handling), natural disaster response (floods, wildfire, storms and so on) as well as tasks such as removing beehives or inspecting security measures in buildings.

Every time a fire fighting team is dispatched a report of activity shall be created by the commander at the scene. These reports are prepared in a particular, regulated manner and stored in EWID – a computerized incident data reporting system (IDRS) built for this purpose. Each of approximately 500 Fire and Rescue Units (JRG) of the PSP conducts around 3 fire & rescue actions a day. Since after every action a report is created, the total number of reports in EWID is currently around six million.

* This work was partially supported by the Polish National Science Centre grants 2011/01/B/ST6/03867 and 2012/05/B/ST6/03215, and by the Polish National Centre for Research and Development (NCBiR) - grant O ROB/0010/03/001 under Defence and Security Programmes and Projects: "Modern engineering tools for decision support for commanders of the State Fire Service of Poland during Fire&Rescue operations in buildings".

The EWID reporting system is an unparalleled source of information and knowledge about fire&rescue (F&R) operations. Ability to process and analyze this data could help in development of new procedures and protocols as well as aid the optimization of existing ones [1]. The knowledge derived from EWID may be also very helpful in firefighters' training process. For example, if we can retrieve a reference set of descriptions of similar situations from EWID we can apply techniques based on Conversational Case Based Reasoning (CCBR, see [2]) to decide the course of actions for the new situation. In order to use information contained in EWID efficiently and effectively we need to be able to search and summarize reports according to various, possibly changing requirements.

In this paper we focus on one of the particular tasks associated with identification of EWID records that fulfill certain criteria. This corresponds to identification (retrieval) of action reports that describe situation involving a predefined elements (concepts) such as "Hymenoptera insects","mini-bus" or "carbon monoxide". An important factor in that the concept we look for may not be explicitly mentioned in the record. As EWID record comprise of numerical part and Natural Language (NL) description part, we are particularly interested in finding records related to a preset concept even though they do not have corresponding numerical indicators set and the description part is not clearly listing these concepts. We describe a set of techniques that make it possible to cleanse and filter EWID records, most importantly their description part, in such a way that the search/identification is efficient. This involves overcoming typical problems associated with inconsistencies, vagueness and imprecisions that are commonplace in EWID records. Yet another type of problems that we have to overcome is associated with the very nature of NL data. Notions (words) we are looking for may have various meanings (homonymy) or may be replaced by synonyms.

While it is possible to obtain good results using classical search techniques, their application to description part of EWID records is not always viable in practical applications. In a nutshell, they require a person in front of the computer, who is able to resolve inconsistencies (e.g. homonymy), identify meanings and tune filters. In order to ease some of this manual load and extend search scope while retaining acceptable quality of retrieved information we propose to use a combination of language processing and data analysis tools. In our approach texts from the description parts of EWID records are converted to different representation with use of a method known as Latent Semantic Analysis (LSA). Then, a clustering technique is used to find groups of semantically similar concepts. This grouping is then a basis for constructing search and retrieval algorithm. The quality of retrieved result is compared with straightforward manual filtering by means of standard measures from the field of Information Retrieval (IR - see [3]) such as *recall*, *precision*, and *F-measure*[3].

[3] `http://en.wikipedia.org/wiki/F1_score`

In the paper we first introduce the data we work with (Section 2), then we describe the methodology behind our approach (Section 3). The application of the proposed method and results obtained this way are presented in the Section 4. We finish with discussion of results and conclusions in Section 5.

2 Description of Data

Our data set consists of 291 683 F&R reports extracted from the EWID system. They contain information about incidents to which PSP responded in the period between 1992 and 2011. The data is limited to incidents that happened in the City of Warsaw and its surroundings. Out of 291,683 cases in this dataset 136,856 reports represent fires, 123,139 local threats, and 31,688 false alarms.

As already mentioned, each report consists of a numerical attribute section and a natural language *description* part. The attribute section consists of 506 attributes describing various types of incidents. However, depending on the category of incident, the number of attributes that are actually present (have a non-zero value) varies from 120 to 180 per report. Most of the numerical attributes are boolean (True/False), but there are also some numerical values like fire area or amount of water used to extinguish the fire.

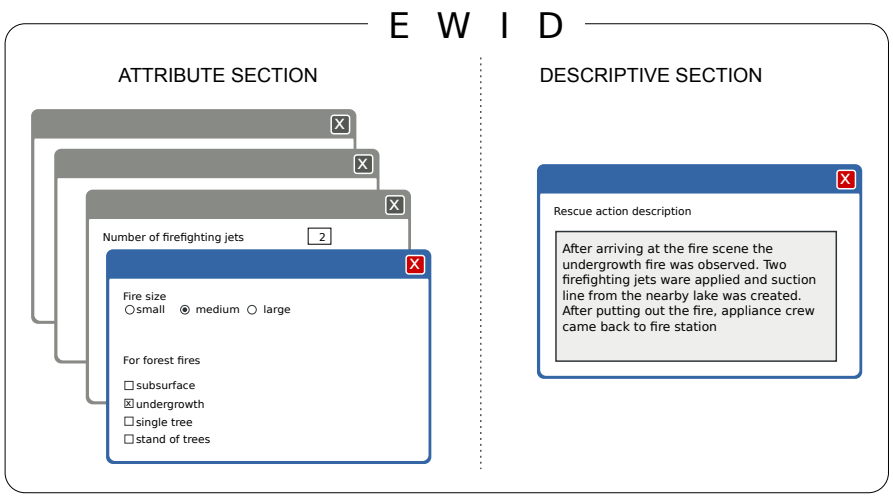

Fig. 1. Representation of a report in EWID database.

The natural language description (NL) part is an extension of the attribute part. It was designed to store information which cannot be represented in the form of a predefined set of attributes. Unfortunately, there are neither clear regulations what should be written in the description part nor any strict guidelines

regarding the format of this part. Therefore, in this part a full spectrum of information can be found. Some descriptions contain detailed information including the precise timeline of events while others are very brief and general. On average, the NL part contains approximately three sentences that describe the situation at the fire ground, actions taken, and weather conditions. Figure 1 depicts the idea of a report representation in EWID database.

In terms of factual aspects the data stored in the EWID contains information about persons, objects involved in the incident, and methods used to eliminate threats that have arisen.

For the purpose of this study we decided to concentrate on three types of incidents that are of some importance to overall management of Fire Service. These are:

1. Incidents where carbon monoxide was present. This mostly concerns fires in residential buildings as carbon monoxide poisoning is one of the major threats in such incidents and the cause of major part of fatalities.
2. Incidents with insects of the Hymenoptera order such as honeybees, hornets, bumblebees, wasps. These incidents fall into the local threats category. Even though they are rarely a major problem, these incidents require relatively high amount of manpower and involvement of specialized equipment.
3. Road collisions involving mini-buses. This category of road accidents is somewhat special. A *mini-bus* in the Polish terminology is a vehicle that is registered to carry between 7 and 12 persons. Such vehicle can be driven by a person with a regular (non-professional) driving permit. In the recent years the accidents involving mini-buses became a major issue in Poland. From security services' point of view they are important, as they may involve many more casualties that "regular" road collisions, and hence require much larger resources to respond to.

In order to perform our experiments we selected from the original data set a sample of 4 135 reports. The records in our subset consist only of NL description part. We extracted this subset using a two-fold procedure. First, using the attribute part we selected the reports which we suspected to contain the kinds of incidents that are of interest to us. Then, using a greedy algorithm based on searching for regular expressions in NL part, we narrowed down the number of previously selected reports to 2 135. In the second step, we selected at random a sample of 2 000 reports, regardless of their kind as a reference sample. Then, we merged this two subsets into one data set for experiments with 4 135 reports in it.

We are fully aware that our data subset may not be sufficiently representative as part of it was not properly, randomly sampled. However, the fully random sample contains too few interesting reports. Therefore, we opted for a compromise combining the fully random sample with the preselected bunch of reports.

The next phase of data preparation involved inspecting (reading) the selected reports one by one and labeling them manually. This step is tantamount to injection of the expert knowledge into the system. We assigned the report to a

category (carbon monoxide, hymenoptera, mini-bus) if it contains the information about a corresponding type of incident. Our final, partly labeled data set contains 82 reports with carbon monoxide intoxication, 167 with road accidents involving mini-buses, and 1557 incidents with Hymenoptera.

3 Methods

Our methodology involves four approaches. In the first approach we adopt a traditional search with use of regular expressions. The user inserts a term or terms which express his information need. He/she defines it using exact or fuzzy search with wild-cards. For example, while searching for reports which describe incidents with *carbon monoxide intoxication* the query can be defined as: "*carbon monoxide*" or "\s CO \s", where *CO* is a chemical symbol for carbon monoxide.

In the second approach the experts define a set of concepts which are related to the defined problem. We transformed these concepts into set of lexemes, i.e., search terms. For example, the problem of finding the reports with carbon monoxide intoxication was defined by the following set of terms: *carbon monoxide, CO, oxide, afterdump, choke-dump, asphyxiate, intoxication.*

In the third approach we we transformed the reports to Latent Semantic Space and performed search using the cosine similarity measure between the query and each of the reports. The fourth approach was similar to the third, but the transformation to LSA was extended by clustering. LSA representations of reports were clustered in order to identify groups of similar incidents.

All the approaches were compared using standard information retrieval measures (recall, precision, and F-measure). In the following subsections we provide some details of our approaches, except for the first one, as it is quite common and simple.

3.1 Search with a set of predefined terms

For all three classes of EWID reports (carbon monoxide, Hymenoptera and mini-buses) we asked domain experts (firefighters) to define the concepts which are related to these problems. They have created a list of concepts which, in their opinion, can express the problem, are associated with it, or occur very often at the emergency scene while responding the particular type of incident. Then, we transformed these concepts into a set of terms. Namely, for the problem of searching carbon monoxide intoxication we defined the following set: *carbon monoxide, co, oxide, afterdump, choke-dump, asphyxiate, intoxicate.* The information need for Hymenoptera-related incidents was defined by the set of terms: *wasp, bee, hornet, bumblebee, insect, cocoon, swarm, ergotizm, gastight clothing.* The terms corresponding to road accidents with mini-buses were: *mini-bus, dostawczy (Polish-specific word), courier.*

As Polish is a fusional language we had to deal with problems posed by inflexion of words. To do that we lemmatized the words in reports from our data

set, creating the non-inflexed form. The lemmatizations were performed with use of the Morfologik software [4].

For each of three incident types we ran a query against the data set using all the terms associated with the given type. The terms were combined in the query using OR operator. The experimental results with this approach are presented in Sections 4 and 5.

3.2 Search using LSA space representation

This approach is based on the transformation of the reports using *Latent Semantic Analysis* (LSA) [5, 6]. The basic idea of LSA is to create the concepts for the given text corpus and then assign each single word from a document (report) to a corresponding concept. The result is that reports can be expressed in Latent Semantic Space as vectors of corresponding concepts' weights. The advantages of LSA representation are that it is considerably more compact than original one and makes it possible to find indirect similarities between reports or between reports and queries.

The reports were lemmatized and transformed into LSA space with use of the R system's [7] library *lsa*. As a result of the transformation we obtained three matrices: *report – concept* matrix (concept in the sense of LSA), *term – concept* matrix and *eigenvalues* matrix. The number of LSA dimensions was established experimentally, based on the values of final measures (precision, recall, F-measure). We found out that the best number of dimensions in this case is 50.

In this approach the search for relevant reports was performed as follows. First, the query (e.g. carbon monoxide) was converted to vector (bag-of-words) form and multiplied by the term-concept matrix in order to obtain its LSA representation. Then, using the cosine similarity measure we find in the report-concept matrix the reports which are similar, as vectors, to the query. The threshold for similarity (cosine between query and report vectors) was established experimentally at 0.7.

3.3 Search using LSA representation and clustering

We have found the results obtained through the usage of the LSA with default settings to be unsatisfactory. In order to improve the results we resorted to cluster analysis. The reports were transformed to LSA representation (report-concept matrix) and then clustered with using the Partitioning Around Medoids (PAM) algorithm [8]. In order to obtain the number of cluster in PAM clustering, we used the silhouette index [9] as a primary measure, complemented with our final performance measures (recall, precision, F-measure). After several repetitions of experiments we have established the desired number of clusters to be 10.

In order to assess the performance of the approach was proceeded as follows. First, we inserted the terms which define our information need, for example "mini-bus". Then, we obtained the names of the clusters which contain such a term. Next, reviewing the reports in these clusters we retained only the clusters

in which the concept (mini-bus) appear in the desired context. In the mini-bus example the desired context would be "road accident with mini-bus". This allows us to eliminate clusters that contain the concept of mini-bus which is not taking part in a road accidents. For example, mini-buses that were burned in parking fire. The similar situation was in the case of carbon monoxide. The clustering helped us in finding the "CO" in the proper context. With carbon monoxide searched as "CO" the situation is tricky because of homonymy. The abbreviation "CO" (uppercase[4]) in Polish is commonly used denote a concept of "central heating" (*Centralne Ogrzewanie*).

4 Results

In Table 1 we show a summary of results obtained from experiments. For each of the methods we calculated the values of three measures: recall precision, and F-measure. The measures were calculated with use of manually labeled reports as the reference.

Table 1. Comparison of search methods.

Method	Measure	Carbon monoxide	Mini-buses	Hymenoptera
Regular expressions	recall	0.451	0.898	0.402
	precision	0.069	0.877	**0.987**
	F-measure	0.120	0.888	0.571
Set of terms	recall	0.671	0.892	**0.990**
	precision	**0.671**	**0.914**	0.981
	F-measure	**0.671**	**0.903**	**0.985**
LSA	recall	0.021	0.347	0.014
	precision	0.011	0.397	0.016
	F-measure	0.012	0.371	0.019
LSA with clustering	recall	**0.768**	**0.928**	0.974
	precision	0.173	0.330	0.557
	F-measure	0.282	0.487	0.709

According to Table 1 the best results were achieved by the approach which used predefined search terms. For each of the classes it obtained the best value of F-measure. Reasonably good results were obtained using the representation of reports in LSA space coupled with clustering.

[4] The situation is even more complicated with a lowercase word "co". It is a common stop word in Polish roughly equivalent – depending of context – to English "what" or "which/that".

5 Discussion and Conclusions

Even though the traditional, term-based search in EWID records returns reasonable results, it is not prefect. In order to get the desired outcome the user of the system must possess some (expert) knowledge of topics from F&R and associated domains. For example, obtaining satisfactory result of search for incidents that involved Hymenoptera requires setting of several filter conditions. Establishing such filtering conditions may be complicated and inconvenient. Similar complications were also symptomatic for other types of searches.

The problems posed by existence of synonyms, homonyms and various elements of specialized jargon had to be overcame. The first attempt was based on analysis of hidden semantic groups derived with use of LSA. Two separate experiments were made. These experiments differ by the operations that were used to transform the (matrix) LSA representation. First of these attempts was made using the *cosine* measure. It measures the angle between the vectors in LSA representation. In this particular case we were interested in measuring the angle between vectors that represent the query and the documents (EWID descriptions). During the experimental evaluation we have determined that this method is inefficient. For most types of queries finding the proper threshold for the value of cosine was problematic. This threshold is used to decide whether a document answers the query or not. Only in the case of querying for incidents involving mini-bus the results were reasonable. In this case we have obtained value of F-measure at 0.37, but the value of precision was merely 0.35. Moreover, the retrieved set of records contained quite high number of incidents involving insects. This may be a result of the two categories (mini-bus-related and insect-related) being identified as semantically close. This semantical closeness is most likely a result of existence of several reports involving both kinds of incidents.

As the results obtained with simple cosine approach were far from satisfactory we had to look for improvements by changing the way the LSA representation was used. In the next attempt we divided the corpus of texts into a pre-set number of clusters. After several experiments we have determined that the best number of clusters in this case is ten. Each cluster was meant to contain reports that share similar context. It was indeed possible to perform clustering in such a way that the clusters were semantically consistent. The best cluster was associated with incidents involving mini-buses and contained 93% of relevant reports. This was, in fact, the best single result we have obtained with any of the methods used. Unfortunately, results obtained with use of this cluster were inferior to manual retrieval attempts since the precision value for this record is only 0.33. This means that a large number of incidents involving types vehicles other than mini-bus was also present in this cluster. The clustering helped to increase the quality of searches involving Hymenoptera. In this case, union of four most prominent clusters contained 97% of all relevant incident reports. The problem with relatively low precision of this approach still remains. For Hymenoptera queries the clusters provided precision at the level of 0.56, which yields the value of F-measure at 0.709. To put this in context, the manual filtering on lemmatized corpus had value of F-measure at 0.985. Clustering approach provided also some

results for queries involving carbon monoxide. One of clusters that were found for this case had the recall 0.768, which is the highest value that we have achieved in all of our experiments. Nevertheless, this cluster is still only marginally useful, as it has very low precision, resulting in value of F-measure that is only 0.282.

To sum up, the best overall results w.r.t. precision, recall, and F-measure were obtained using the semi-manual retrieval with pre-defined set of search terms. This can be further improved by lemmatization which, combined with removal of additional stop words (using Morfologik library), significantly reduces data size and hence the computational effort. The downside of this approach is the necessity of defining and manually entering several search terms. Experiments show that entering search terms one-by-one is ineffective. The have to be several of them in the filter so that they cover a broad range of possible combinations that may occur in incident descriptions. This requires the user to have a good overview of the data corpus and some domain knowledge about incidents stored in EWID.

The problem with requirements for extended users' expertise can be to some reasonable extent – as our initial experiments show – addressed with use of LSA. Conversion of EWID description to LSA representation makes it possible to group (cluster) similar reports. In order to make the demonstrated approach usable we would have to prepare tools that allow user to navigate through the cluster in an intuitive and efficient manner. In particular, the user can be presented with clusters that are represented by a selection of frequent and relevant terms that occur in descriptions that belong to such clusters. Yet another possible extension of our approach could make use of hierarchical clustering. Last, but not the least, we are considering building a search engine that would make it possible for user to perform a faceted search (see [10]) for relevant reports with use of clustering. Facets such as the number and relevance of search terms in aa given cluster may be then used to discriminate between the valuable information and noise.

References

1. Krasuski, A., Kreński, K., Łazowy, S.: A Method for Estimating the Efficiency of Commanding in the State Fire Service of Poland. Fire Technology **48**(4) (2012) 795–805
2. Aha, D.W., Breslow, L.A., Muñoz Avila, H.: Conversational case-based reasoning. Applied Intelligence **14**(1) (2001) 9–32
3. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press (2008)
4. Morfologik: About the project. `http://morfologik.blogspot.com/2006/05/about-project.html`
5. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science **41**(6) (1990) 391–407
6. Landauer, T., Foltz, P., Laham, D.: An introduction to Latent Semantic Analysis. Discourse Processes **25**(2) (1998) 259–284

7. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2008)
8. Reynolds, A., Richards, G., De La Iglesia, B., Rayward-Smith, V.: Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. Journal of Mathematical Modelling and Algorithms **5**(4) (2006) 475–504
9. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20** (1987) 53–65
10. Sacco, G.M., Tzitzikas, Y., eds.: Dynamic Taxonomies and Faceted Search. Volume 25 of The Information Retrieval Series. Springer Berlin Heidelberg (2009)