Proceedings of the 2013 Federated Conference on
Computer Science and Information Systems pp. 77–82

Semantic Tagging of Heterogeneous Data: Labeling Fire&Rescue Incidents with Threats

Adam Krasuski* and Andrzej Janusz†
*Chair of Computer Science, The Main School of Fire Service
Słowackiego 52/54, 01-629 Warsaw, Poland
†Institute of Mathematics, University of Warsaw
ul. Banacha 2, 02-097 Warsaw, Poland
krasuski@inf.sgsp.edu.pl, janusza@mimuw.edu.pl

***Abstract*—In the article we present a comparison of the classification algorithms focused on labeling Fire&Rescue incidents with threats appearing at the emergency scene. Each of the incidents is reported in a database and characterized by a set of quantitative attributes and by natural language descriptions of the cause, the scene and the course of actions undergone by firefighters. The training set for our experiments was manually labeled by the Fire Service commanders after deeper analysis of the emergency description. We also introduce a modified version of Explicit Semantic Analysis method and demonstrate how it can be employed for automatic labeling of the incident reports. The task we are trying to accomplish belongs to the multi-label classification problems. Its practical purpose is to support the commanders at a emergency scene and improve the analytics on the data collected by Polish State Fire Service.**

***Keywords*-Domain Knowledge, Multi-label Classification, Explicit Semantic Analysis, Fire Services**

I. Introduction

THE MAIN goal of Fire Services activity at the fire ground is elimination or neutralisation of arisen threats. Therefore, the core of the Fire&Rescue (F&R) action is to adequately recognize possible dangers for the involved people and properties. A specific emergency generates specific threats. It implies that if we posses a description of the emergency we could predict dangers, or more precisely threats, related to the actual emergency.

The recognition of the threats at the fire ground is only a part of activities that should be performed at the beginning of a F&R action. No less important is the recognition of threatened individuals or objects. Together, those two tasks play a pivotal role in planning of the further actions at the emergency scene.

The task of recognition and categorization of threats is formalized in the tactic of German Fire Service [1]. After arriving at a fire ground or an emergency scene German commanders have to evaluate and recognise the appearing threats. In order to do this systematically and not to miss any of the threats, they have to fulfill the Threats Matrix (in German – Gefahrenmatrix) [1]. The Threats Matrix helps to identify the threats emerging at the scene and the threatened objects. The columns of the matrix represent threats, and the rows represent objects which can be threatened. The Table I depicts the Threats Matrix.

TABLE I
THE THREATS MATRIX USED BY GERMAN COMMANDERS. LEGEND: A1 – FEAR, A2 – TOXIC SMOKE, A3 – RADIATION, A4 – FIRE SPREADING, C – CHEMICAL SUBSTANCES, E1 – COLLAPSE, E2 – ELECTRICITY, E3 – DISEASE OR INJURY, E4 – EXPLOSION

Threat/object	A1	A2	A3	A4	C	E1	E2	E3	E4
People (ME)									
Animals (T)									
Environment (U)	–					–	–	–	
Property (S)	–	–						–	
Rescuers (MA)									
Equipment (G)	–	–						–	

In German language, column names are chosen so that they can be easily remembered. In order to help to memorize all threats by commanders, German threats' names were taken to form the following pattern: AAAA-C-EEEE *Angstreaktion, Atemgifte, Atomare Strahlung, Ausbreitung, Chemische Stoffe, Einsturz, Elektrizität, Erkrankung, Explosion*. The sign '–' in the table indicates, that this threat in general does not apply to this object. At the background of filled Threats Matrix, German commanders define the Threat Focus and organize their actions accordingly.

If we could create a computer system which can recognize threats at the fire ground, we would effectively support the commanders in doing their duty. Moreover, our previous research [2], [3], [4] show that analytics performed in abstract spaces, such as Threats Matrix allows to reduce significantly the number of dimensions without loosing the information about complexity of the real phenomena. Unfortunately, the Polish Fire Services do not use the method of filling the Threats Matrix at the incident scene. Therefore Polish Incident Data Reporting System called EWID lacks of this information. In the previous work [4] we labeled incidents with threats manually. The reports from EWID database were analysed and labeled by extramural students of The Main School of Fire Service with commanding experience. We selected only commanders having at least seven years experience in commanding. They were involved as *experts – practitioners* in labeling real action reports from the EWID system.

We created a special system to support manual labeling the reports. The labeling process consists of two main phases:

tutorial phase and *labeling phase*. The tutorial phase was focused on introducing the Threats Matrix and the layout of EWID incident reports to the experts. It was divided into three consecutive parts. In the first part, experts were introduced to the format and purpose of the Threats Matrix. In the second part, some examples of filled Threats Matrices were presented and discussed with the experts. In the third part, experts received an exemplary EWID report together with a Threats Matrix describing this report. The labeling phase consisted of many evaluation stages. At every stage the experts were provided with a single EWID report. On the ground of the information about the incident described in the report, they were asked to evaluate threats which appeared during the reported incident and to complete its Threats Matrix. Every expert was asked to label at least 100 EWID reports. Every report description was labeled by only one expert. In total, we collected 406 labeled incident descriptions.

The presented method has very serious shortcomings – it is not scalable. If we need more labeled reports we need more commanders. Up to now approximately 7 million reports have been collected In the EWID database. Moreover, every day 1 500 new reports are submitted into the system. It is obvious that such a number of incidents is not manually manageable by people.

This article is devoted to prediction methods from Machine Learning which can be used to label the incidents automatically. We use the 406 labeled incidents as a data set to train and evaluate our classification algorithms. We analyse different data representation and different multi-label classification methods in order to find the best one.

The remaining of the paper is structured as follows. In Section II we describe our data set which was used as an input in our experiments. In Section III we present our method of the analysis focused on determining the best classification algorithm and data representation. Section IV contains the results of the conducted experiments. The article is concluded with the interpretation and a summary of the research results, as well as a discussion on perspectives for the future research.

II. Description of the Data

Our data set consists of 291 683 F&R reports. They contain information about the incidents responded by Fire Service, from the years 1992 to 2011. The data concerns the incidents which happened in Warsaw City and its surroundings. In this data set 136 856 reports represent fires, 123 139 local threats and 31 688 where false alarms.

Each of the reports consists of an attribute section and a natural language description part. The attribute section contains 506 attributes describing all types of incidents. However, depending on category of the incident, the number of attributes that take values different than zero varies from 120 to 180 for a report. Most of the attributes are boolean (True/False) type but there are also numerical values (i.e. fire area, amount of water used).

The natural language description (NL) part is an extension to the attribute part. It was designed to store information, which can not be represented in a form of a set of attributes. Unfortunately there is no clear regulation what should be written in the NL part. Therefore, in this part a full spectrum of information, from detailed information including time coordinates, to the very general and brief descriptions can be found. The simple statistics reveal that NL part contains approximately three sentences that describe the situation at the fire ground, actions undertaken and weather conditions. Figure II depicts the idea of a report representation in EWID database.

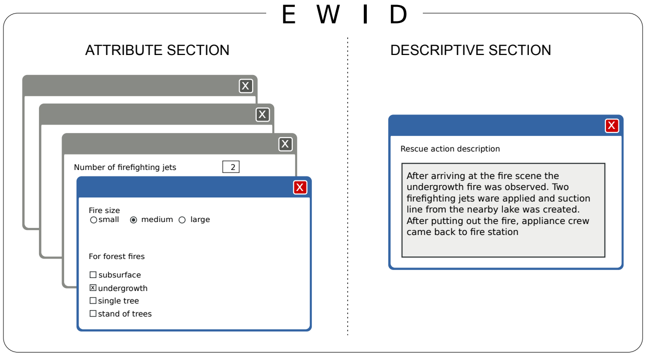

Fig. 1. Representation of a report in EWID database.

In factual aspects, the data stored in the EWID contain information about persons, objects involved in the incident and methods used to eliminate the arisen threats.

In our experiments we used a subset of this data set. For the process of labeling (assigning threats) the incidents by domain experts, we selected only the reports representing fires of residential buildings. This subset of the data consisted of 31 556 reports. From this set 406 reports were labeled by the experts. We used these reports in our experiments described in Section III.

III. Method

The labeling methodology was briefly presented in Section I and is broadly discussed in [4]. In this section we present several approaches to automated labeling of the reports.

In this research we pay a special attention to two aspects of the task: finding an appropriate classification algorithm and selecting a good input data representation. The first approach can be divided into two groups: classifiers which operate on incident features and classifiers which operate on features of the threats. The set of possible representations for the second task consists of: structured part only (SP), NL part only (NLP), structured part plus bag-of-words of descriptions of object (SP-OD), structured part bag-of-words and NL part transformed to the LSA space (SP-OD-LSA). In the next subsections we describe the utilized methods in detail. All the performance evaluation experiments were conducted on a training set of 285 incidents and a test set consisting of 122 incidents.

A. Classification on Structured Part Only

As was mentioned in section II the structured of part EWID database is represented by 506 attributes. However, for our subset of 406 incidents many of them have zero values for each of the incidents. Therefore, we removed those attributes from our subset. We also removed semantically irrelevant attributes, such as ID of a fire station. As a result we obtained an information system with 208 attributes, from which 24 were numeric and 184 were of a boolean type. The prediction targets were sets of combinations of threats and threatened objects (the risks). The set of the possible risks was created as a Cartesian product of threats and objects from the Threats Matrix. Such a representation constituted an input for our classifiers.

Our first experiment was focused on determining the best classifier for a given representation. In this experiment we used the whole set (406 cases) and 5 folds cross-validation technique to evaluate the efficiency of different classifiers. We tested: Naive-Bayes (NB), Classification Tree, Support Vectors Machine (SVM), Clark Niblett induction algorithm (CN2) and Random Forrest. We used Matthew correlation coefficient (MCC)[1] to evaluate the efficiency of the selected classifiers. Table II depicts the comparison of the results.

TABLE II
COMPARISON OF THE CLASSIFIERS. AUC – AREA UNDER CURVE, MCC – MATTHEW CORRELATION COEFFICIENT, SVM – SUPPORT VECTORS MACHINE, CN2 – CLARK NIBLETT INDUCTION ALGORITHM.

Method	AUC	F1-score	MCC
Naive Bayes	0.76	0.61	0.33
Classification Tree	0.71	0.53	0.26
SVM	0.75	0.46	0.27
CN2 rules	0.76	0.50	0.31
Random Forest	0.71	0.45	0.20

According to the criteria (MCC measure) presented in Table II the Naive-Bayes classifier obtained the best results and was selected as a representative for the rest of the experiments. Next, we used training and test methodology to compare the current methods with other approaches. The classifier was trained and tested separately for each of the decisions classes. Then we calculated some performance measures i.e. precision, recall, F1-score in two way: for each of the incidents and for each of the decision classes.

B. Classification on Structured and Object Description Parts

We repeated the experiment described in Section III-A extending the incident representation by the *object description* attribute. The attribute is an extension to the *object type* attribute and contains information such as: storey or room of the building where fire occurred, trash localization (inside, outside in the case of a trash fire), etc. This attribute stores the information in natural language form. In order to use this part as a feature vector we transformed it into the bag-of-words representation. We created Term-Document-Matrix (TDM) to transform the NL attribute into a feature vector. However, in order to reduce the number of dimensions, we firstly lemmatized the descriptions using the Morfologik software [5]. As a result of lemmatization we obtained 1 029 unigrams in bag-of-words representation. Then we repeated the experiments described in Section III-A and calculated the performance measures.

[1]http://en.wikipedia.org/wiki/Matthews_correlation_coefficient

C. Classification on Structured Object Description and NL Part

In this experiment we extended the incidents representation by the entire NL part. The difference between this approach and the one presented in Section III-B is that, we are not limited to the object description attribute only. We used the whole description of the incidents stored in the NL part.

As in the case of object description we lemmatized the textual data and created Term Document Matrix. The TDM revealed that the number of unique words equals 1 277. It our opinion the direct representation of NL part throughout term vector is to exhaustive due to the fact that TDM is a sparse matrix. In order to reduce the number of dimensions and increase the separation of the incidents we use the Latent Semantic Analysis method [6]. After conversion of the TDM into LSA space we obtain 85 dimensions. We calculate the number of dimensions finding the first position in the descending sequence of singular values where their sum, divided by the sum of all values, meets or exceeds the 0.5 share value. Next in order to obtain categorical arguments we discretized the attributes of incidents represented in the LSA space. After discretization, each of the LSA attributes had tree values: -1, 0 or 1. Then, we repeated the experiments described in Section III-A.

D. Classification Based on Features of the Threats

In this experiment we change our approach to the labeling task. Instead of training a classifier to recognize which attribute values should the incident have in order to be labeled with a given threat, we try to learn the features of the possible risks. In other words, we learn which features of the incidents most adequately represent the given risk.

The risks are defined as a Cartesian product of threats and objects from the Threats Matrix and are represented by a concatenation of a threat and an abbreviation of an object name (i.e. E1_ME). We decided to utilize the Explicit Semantic Analysis method [7] to devise the new representation of the risks in order to facilitate the labeling.

Explicit Semantic Analysis (ESA) proposed in [7] is a method for automatic tagging of textual data with predefined concepts. It utilizes natural language definitions of concepts from an external knowledge base, such as an encyclopedia or an ontology, which are matched against documents to find the best associations. The definitions of concepts are regarded as a regular collection of texts, with each description treated as a separate document. The model structure imposed by ESA can be interpreted as a one layer neural network [8] with L input nodes corresponding to terms and K output nodes corresponding to concepts. The associations between terms and concepts have numerical weights. Figure 2 depicts the idea of ESA in a form of a neural network.

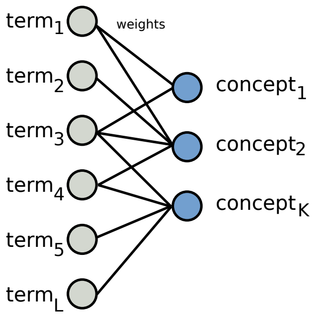

Fig. 2. Representation of the ESA as a neural network.

The implementations of ESA described in [7], [9] use external resources (e.g., an ontology, Wikipedia) which contain the definitions of concepts. In our case there are no external sources of knowledge which can serve as definitions of the risks, therefore we modified the primary idea of ESA and we created *self-defined ESA*. We aggregated the NL parts of incident descriptions from the training set by the assigned risks. In particular, all descriptions of incidents which were labeled with the same risk were concatenated into one document. We repeated this operation for consecutive threats obtaining as many documents as there were threats. Next, we created a Term-Document-Matrix where columns represent the risks and rows represent the terms. The intersection of a column and a row represents an association between a term and a risk. Then, for each incident description in the test set we iterate through the terms, obtain and sum the values of associations to the consecutive risks. The risks for which the sum of the associations is higher than zero constitute a bag-of-risks for a given incident.

For the sake of clarity, we formalize this approach. Let $\mathcal{T}$ be a set consisting of M incidents from the training set, $\mathcal{T} = \{I_1, \ldots, I_i, \ldots, I_M\}$ and $\mathcal{T}'$ is a set consisting of J incidents from test set $\mathcal{T}' = \{I'_1, \ldots, I'_i, \ldots, I'_J\}$. Let $\mathcal{R} = \{r_1, \ldots, r_k, \ldots, r_K\}$ be a set of risks at a fire ground defined as $\mathcal{R} \subseteq \mathcal{H} \times \mathcal{O}$, where $\mathcal{H}$ is a set of the threats from the Threats Matrix and $\mathcal{O}$ is a set of the objects from the Threat Matrix (see Table I). Moreover, let us assume that there were identified L unigrams (e.g. words, stems) $W = \langle w_1, \ldots, w_i, \ldots, w_L\rangle$ from the descriptions of incidents in the training set (NL part of the EWID database). Any incident I_i in the set $\mathcal{T}$ can be represented by a vector of features $I_i = \langle E_i, U_i\rangle$, where $E_i = \langle e_1, \ldots, e_j, \ldots, e_L\rangle \in \mathbb{R}^L$ is a bag-of-words representation of the incident I_i and $U_i = \langle u_1, \ldots, u_k, \ldots, u_K\rangle \in \{0,1\}^K$ represents risks assigned by the experts to the I_i. Each coordinate e_i expresses a value of some relatedness measure for i-th term in vocabulary (w_i), relative to the given E_i. The coordinate $u_k = 1$ if the risk r_k was assigned to the incident I_i by the experts. Respectively, any incident I'_i from the test set is represented by a vector of features $I'_i = \langle E'_i, U'_i\rangle$. However each coordinate of U'_i equals 0 because the incidents from the test set were not labeled by the experts.

We can now define the description of a risk r_k. The description D_{r_k} for risk the r_k is a sum of all bag-of-words representations of incidents in $\mathcal{T}$ where $u_k = 1$

$$D_{r_k} = \sum_{i:u_k=1} E_i \tag{1}$$

Next, all the D_{r_k} from set $\mathcal{D}$ (set of all designations) were converted to the Term-Document-Matrix (TDM), where columns are labeled by descriptions, rows by terms from vocabulary W and the coordinates d_i represent relatedness measure for i-th term in vocabulary (w_i), relative to the given D_{r_k}. The measure used by us to calculate d_i is the *tf-idf* (term frequency-inverse document frequency) index (see [10]) defined as:

$$d_i = tf_{i,k} \times idf_i = \frac{n_{k,i}}{\sum_{j=1}^{L} n_{k,j}} \times \log \frac{|D|}{|\{D_{r_k} : w_i \in D_{r_k}\}|}, \tag{2}$$

where $n_{k,i}$ is the number of occurrences of the term w_i in the description D_{r_k}, $|D|$ is the cardinality of $\mathcal{D}$ which equals K, and $|\{D_{r_k} : w_i \in D_{r_k}\}|$ is the number of the descriptions where the term w_i appears.

Next, the TDM representation of the risks descriptions can be used as an inverted index that maps terms into lists of risks. Each row of TDM expresses an association of the given term from W to K risk descriptions. The inverted index is utilized as a semantic interpreter to assign the risk into incidents from the set $\mathcal{T}'$. Given an incident description E'_i, it iterates over terms from the description, retrieves the corresponding entries and merges them into a weighted vector of risks that represent the given incident.

Let $E'_i = \langle e'_1, \ldots, e'_j, \ldots, e'_L\rangle$ be a bag-of-words representation of the description of incident I'_i from the set $\mathcal{T}'$. Let $inv_{i,k}$ be an inverted index entry for e'_i. It quantifies the strength of association of the term w_i to a risk r_k. For convenience, all the weights $inv_{i,k}$ can be arranged in a sparse matrix structure with L rows and K columns, denoted by INV, such that $INV[i,k] = inv_{i,k}$ for any pair (i,k).

Next we can create a new representation U_i^{INV} of incident I'_i from the test set as a sum of values from the TDM of terms which appear in E'_i:

$$U_k = \sum_{i:e'_i \neq 0} e'_i \times inv_{i,k} = W_i * INV[\cdot, k]. \tag{3}$$

In the above equation $*$ is the standard dot product and $INV[\cdot,k]$ indicates k-th column of the sparse matrix INV. This new representation will be called a *bag-of-risks* of an incident I_i.

The new representation of incident I'_i can be now definded as $I''_i = \langle E'_i, U''_i\rangle$ where U''_i is created as follows $U''_i = \langle u''_1, \ldots, u''_k, \ldots, u''_K\rangle \in \{0,1\}$ $u_k^{INV} > 0 \Rightarrow u''_k = 1$.

For practical reasons it may also be useful to represent the incidents only by the most relevant risks. In such a case, the association weights can be used to rank the risks and to select only the top risks from the ranked list. Therefore, we changed the rule $u_k^{INV} > 0 \Rightarrow u_k'' = 1$ into $u_k^{INV} > var \Rightarrow u_k'' = 1$ where var is some threshold.

IV. Results of the Experiments

In Table III we summarized the results obtained in the experiments. We calculated the performance measures separately for each of the consecutive incidents from the test set, and then we calculated the average for each of the methods. The number of assigned risks for the ESA was set to five with the highest score.

TABLE III
The performance comparison for the classifications methods.

Method	Precision	Recall	F1-score
SP	0.68	0.64	0.61
SP-OD	0.45	0.50	0.43
SP-OD-LSA	0.43	0.51	0.43
ESA NLP	0.48	0.70	0.54

The second summarization (see Table IV) compares the performance of different classification methods, according to the risks from Threats Matrix. In this table we also presented the number of the incidents in the training ($\#\mathcal{T}$) and test ($\#\mathcal{T}'$) sets, which were labeled by a given threat.

Figure 3 outlines the comparison of versatility of the methods. We compare the classification methods according to the number of different risks which were at least once properly assigned to an incident.

The practical usefulness and the importance of the results presented in the Tables and Figure 3 is more broadly discussed in Section V.

TABLE IV
The performance comparison (F1-score) of the classifications methods relative to the risks. $\#\mathcal{T}$ – number of incidents in training set labeled by the given risk, $\#\mathcal{T}'$ – number of incidents in test set labeled by the given risk. The risks are defined as a Cartesian product of threats and objects from the Threats Matrix and are represented by a concatenation of an abbreviation of a threat and an object name (i.e. E1_ME: callapse_people).

Risk	$\#\mathcal{T}$	$\#\mathcal{T}'$	SP	SP-OD	SP-OD-LSA	ESA NL
A1_MA	99	46	0.38	0.49	0.54	0.45
A1_ME	245	107	0.86	0.71	0.69	0.82
A1_T	27	5	–	0.10	0.06	0.07
A2_MA	242	108	0.81	0.64	0.65	0.84
A2_ME	251	103	0.83	0.64	0.70	0.84
A2_S	8	6	0.29	–	0.11	0.22
A2_T	36	8	0.05	0.14	0.06	0.14
A2_U	57	28	0.39	0.37	0.37	0.30
A4_G	3	4	–	–	–	0.08
A4_MA	15	7	0.30	0.32	0.14	0.22
A4_ME	18	9	0.27	0.27	0.15	0.17
A4_S	20	13	–	0.38	0.05	–
A4_T	1	1	–	–	–	0.13
E1_MA	6	3	–	–	0.50	0.11
E1_ME	3	1	–	–	–	–
E2_MA	29	13	0.11	–	0.11	0.31
E2_ME	14	6	–	–	0.13	0.24
E2_S	9	1	–	–	–	0.15
E3_G	3	1	–	–	–	0.12
E3_MA	9	13	–	–	0.10	0.50
E3_ME	2	7	–	–	–	0.12
E4_MA	3	4	–	–	–	–
E4_ME	2	1	–	–	–	–
E4_S	5	2	–	–	–	–
Average F1-score			0.179	0.169	0.181	0.243

V. Discussion of the Results

The obtained results revealed that for the maximization of F1-score for a given document we should choose the method which is based on the attribute section only and the classification algorithm (SP). This method achieved very good performance – F1-score reach the value 0.61 (see Table III). However the value of recall is lower than value obtained by ESA NLP approach. That means the best scoring method does not detect the full spectrum of risks.

The intuition is confirmed by the results from Table IV. We observed that if we calculate the F1-score according to the risks, the SP method is classified as the second best. Table IV also reveals the reason for this situation. The SP approach achieves a very good performance for the risks which are assigned very often to the incidents. As an example may serve the risks: A1_ME (86% of incidents labeled in the training set and 88% in the test set), A2_MA (85% and 89%, respectively), A2_ME (88% and 84% of incidents). For these risks the SP method achieves scores 0.86, 0.81 and 0.83, respectively. However, for the risks which were rarely assigned to the incidents the SP methods fails to achieve good performance.

Figure 3 compares the versatility of the methods. It depicts the spectrum of risks used by different methods. It illustrates that for the method SP and SP-OD only 10 out of 24 risks could be successfully assigned. The extension of the information by the NL part of the EWID database increases the spectrum of the utilized risks. The SP-OD-LSA method successfully assigned 15 out of 24 risks. However, the most versatile method is ESA NL which is able to properly assign 19 out of 24 risks. We may conclude that the attribute section lacks information related to very rare risks. Only the extension by the NL part allows labeling the incidents with these risks.

The conducted experiments proved that there is a potential in ESA method even for short texts and even in a situation when there are no descriptions available for the concepts in a form of external knowledge base (compare the experiments with long text and an external ontology [7], [9], [11]). However, it should be stated that the descriptions stored in the NL part of EWID database are very specific. In the future work the method should be tested against some more general texts like blogs or news.

The future work should also concentrate on methods for improving ESA. Results of our preliminary experiments suggest that a properly adjusted weights in the inverted index

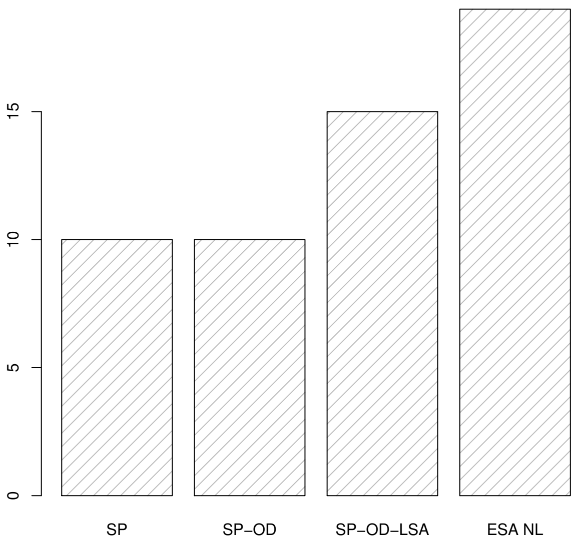

Fig. 3. The comparison of versatility of the methods. Y-axis represents the number of different risks which were at least once properly assigned to an incident.

used by ESA can increase the average F1-score by more than 100%. Therefore, in our research we will focus on finding an appropriate algorithm for updating ESA on for this set.

The experiments also reveal that different risks have different best scoring methods. Therefore, we also consider utilization an ensemble approach in order to assemble a multi-classifier algorithm [12].

ACKNOWLEDGMENTS

The research was supported by the Polish National Centre for Research and Development (NCBiR) - Grant No. O ROB/0010/03/001 in the frame of 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 the buildings" and also by the grants N N516 077837 from the Ministry of Science and Higher Education of the Republic of Poland, and 2011/01/B/ST6/03867 from the Polish National Science Centre.

REFERENCES

[1] A. Graeger, U. Cimolino, H. de Vries, and J. Sümersen, *Einsatz- und Abschnittsleitung: Das Einsatz-Führungs-System (EFS)*. Ecomed Sicherheit, 2009.

[2] A. Krasuski, K. Kreński, and S. Łazowy, "A Method for Estimating the Efficiency of Commanding in the State Fire Service of Poland," *Fire Technology*, vol. 48, no. 4, pp. 795–805, 2012.

[3] A. Krasuski, D. Ślęzak, K. Kreński, and S. Łazowy, "Granular Knowledge Discovery Framework," *New Trends in Databases and Information Systems*, pp. 109–118, 2013.

[4] A. Krasuski and P. Wasilewski, "The Detection of Outlying Fire Service's Reports. The FCA Driven Analytics," in *Processings of the 11-th International Conference on Formal Concept Analysis*, 2013, pp. 35–50.

[5] "Morfologik – About the project," http://morfologik.blogspot.com/2006/05/about-project.html.

[6] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[7] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proc. of The 20th Int. Joint Conf. on Artificial Intelligence*, Hyderabad, India, 2007, pp. 1606–1611. [Online]. Available: http://www.cs.technion.ac.il/~shaulm/papers/pdf/Gabrilovich-Markovitch-ijcai2007.pdf

[8] T. M. Mitchell, *Machine Learning*, ser. McGraw Hill series in computer science. McGraw-Hill, 1997.

[9] A. Janusz, W. Świeboda, A. Krasuski, and H. Nguyen, "Interactive Document Indexing Method Based on Explicit Semantic Analysis," *Rough Sets and Current Trends in Computing*, pp. 156–165, 2012.

[10] C. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval. 2008," *Online edition*, 2007.

[11] M. Szczuka and A. Janusz, "Semantic clustering of scientific articles using explicit semantic analysis," in *Transactions on Rough Sets XVI*. Springer, 2013, pp. 83–102.

[12] K. Kurach, K. Pawłowski, Ł. Romaszko, M. Tatjewski, A. Janusz, and H. S. Nguyen, "An ensemble approach to multi-label classification of textual data," in *Advanced Data Mining and Applications*. Springer, 2012, pp. 306–317.