DOI: 10.15439/2014F253
ACSIS, Vol. 2

Extracting Semantic Prototypes and Factual Information from a Large Scale Corpus Using Variable Size Window Topic Modelling

Michał Korzycki, Wojciech Korczyński
AGH University of Science and Technology
in Kraków
ul. Mickiewicza 30, 30-962, Kraków, Poland
Email: {korzycki, wojciech.korczynski}@agh.edu.pl

***Abstract*—In this paper a model of textual events composed of a mixture of semantic stereotypes and factual information is proposed. A method is introduced that enables distinguishing automatically semantic prototypes of a general nature describing general categories of events from factual elements specific to a given event. Next, this paper presents the results of an experiment of unsupervised topic extraction performed on documents from a large-scale corpus with an additional temporal structure. This experiment was realized as a comparison of the nature of information provided by Latent Dirichlet Allocation and Vector Space modelling based on Log-Entropy weights. The impact of using different time windows of the corpus on the results of topic modelling is presented. Finally, a discussion is suggested on the issue if unsupervised topic modelling may reflect deeper semantic information, such as elements describing a given event or its causes and results, and discern it from pure factual data.**

I. Introduction

UNSUPERVISED probabilistic topic modelling is one of the most widely applied information retrieval techniques, in particular in researches on large-scale corpora. Its main assumption states that text documents are mixtures of topics which may be treated as a multinomial probability distribution over words. These distributions might be created with the use of a couple of methods [1], [2], [3].

In this paper we present the results of experiments in which we attempted to extract deeper semantic information from postprocessed results of unsupervised probabilistic topic modelling methods. It is worth emphasizing that unsupervised methods—despite their utility in information retrieval—cannot directly retrieve semantical relations from texts [4], [5]. Some introductory premises were presented by the authors in [6], such as the hypothesis, that although unsupervised topic modelling does not reflect directly semantic prototypes, those prototypes can be inferred from the extracted topics. This paper provides additional research in the discussed subject and presents new results that prove our hypothesis.

The size of the corpus we operated on was millions of documents. Moreover, it possessed an additional temporal structure.

The framework of the experiments presented in this paper is an integral part of a large scale project related to security and intelligence analysis. The suggested approach permits to find in an analyzed text elements related to the specific semantic prototypes (i.e. mental models) of the events described within and to discern them from pure factual information, given that a sufficiently large corpus is available. Discussed technique may be a first step towards creation of a method for automatic semantic prototype identification. Such methods are essential in a security and intelligence analysis as they can be applied in an automatic identification of objects and their properties in full-text sources.

We start our paper with a short overview of works related to the presented subject - it is included in Section II. Section III introduces and explains the concept of semantic prototypes, firstly described in [7]. Subsequent Sections focus on the experiments of unsupervised topic extraction performed in order to present a method of discerning semantic prototypes from factual information: Section IV introduces corpus used in our experiments, Section V describes the method itself. Results of the experiments are presented and discussed in Section VI. We conclude our paper in Section VII. At the end of the paper, in Section VIII, suggestions of future work are discoursed.

II. Related work

Latent Semantic Analysis (LSA) is an original word/document matrix rank reduction algorithm which extracts word co-occurrences in the frame of a text. As a result, each word in the corpus is related to all co-occurring words and to all texts in which it occurs. The LSA algorithm may be applied in various domains—from a text content comparison [8] to an analysis of human association norm [9]. Unfortunately, there is still little interest in studying the linguistic significance of LSA-made associations.

Latent Dirichlet Allocation (LDA), presented by David Blei, Andrew Ng, and Michael Jordan in 2003, is one of the best known generative model used for topic extraction. It assumes that a collection of documents may be represented by a mixture of latent topics, however words creating each topic are chosen according to a multinomial Dirichlet distribution of fixed dimensionality over the whole corpus. LDA is a technique based on the „bag of words” paradigm and it can infer

978-83-60810-58-3/$25.00 © 2014, IEEE

distributions of topics e.g. with the use of variational Bayes approximation [10], [11], Gibbs sampling [2] or expectation propagation [12].

Some recent research was focusing on finding if the relationships coming from the unsupervised topic extraction methods reflect semantic relationship reflected in human association norms. A comparison of human association norm and LSA-made association lists can be found in [4] and it should be the base of the study. Results of the other preliminary studies based on such a comparison: [5], [13], [14], show that the problem needs further investigation. It is worth noticing that all the types of research referred to, used a stimulus-response association strength to make a comparison. The results of the aforementioned research have shown that using unsupervised topic extraction methods one is able to create associations between words that are strongly divergent from the ones obtained by analysing the human generated associations.

As it has been already noticed, the methods mentioned above are not able to retrieve additional semantic information, however in this paper we introduce some postprocessing methods that may be useful in a semantic text classification.

III. Semantic prototypes

The notion of a semantic prototype comes from cognitive theory [7] where a notion is represented by its elements with their features. So, according to this model, a notion of a „bird" would be „composed" of such elements and features as „feathers", „beak" and „ability to fly". Semantic prototypes can also be discussed in the context of event descriptions that occur in texts. Prototype theory has also been applied in linguistics for mapping from phonological structure to semantics.

In a domain of natural language processing, this approach is reflected in so-called content theories. A content theory is used to determine a meaning of phrases and information they carry. One of the classic and most known elements of content theory is the Conceptual Dependency theory that was invented and developed by Robert Schank [15]. His main goal was to create a conceptual theory consistent in every natural language. The theory's main assumptions were: the generalization of representation language and inference about implicitly stated information. The first assumption means that two synonymous sentences should have identical conceptual representations. The second one states that all the implicit information has to be stated explicitly by inferring it from explicit information. Each sentence can be then represented in a form of conceptual dependency graph built of three types of elements: primitives, states and dependencies. Primitives are predicates that represent a type of an action, states specify the preconditions and results of actions and dependencies define conceptual relations between primitives, states and other objects [16]. Accordingly to the Conceptual Dependency theory we may represent the event of „tea drinking" by a sequence of events: „tea making", „cup operating", „tea sipping" and so on, that are composed of action, object etc.

The described event model proved itself to be very useful in many applications [17] and we found it very suitable to quantifiable comparisons to unsupervised topic modelling methods [18].

From that theory we deduce our model of an event - its prototype - a compound structure of actions, actors, states, and dependencies but also composed of preconditions and results, being events themselves. This event model reflects also very well the semantic structure of a text. If a document describes an event, it is almost always presented in the context of the causes of the event and the resulting consequences. This will be reflected in various topic models that will tend to reflect that most texts are represented as a linear combination of multiple topics. Determining from such combination which topic (event) can be classified as a cause and which is an effect would be very interesting, but that issue is beyond the scope of this paper.

On the other hand, the modelled text is composed not only of events, but also features of those events specific only to that instance of the event. As such, a text can be seen as having two aspects - the main event of the text intermingled with the elements of cause and result events and factual features that are specific to that single event. The latter aspect would relate to places, actors and contextual information. The former aspect would relate to generic elements that are common to similar events that occur in the corpus.

This paper focuses on unsupervised identification and retrieval of those two different event model aspects of texts.

IV. Corpus

The experiment was conducted on a subset of a 30-year worldwide news archive coming from the New York Times. That corpus has been chosen as it is interesting for a number of reasons:

- it is freely available,
- some interesting research results have been obtained based on it [19],
- it is quite comprehensive in terms of vocabulary and range of described event types,
- its relatively large size (approximately 90.000 documents per year, for a total of 2.092.828 documents spanning the years 1981-2012) gives an ample opportunity to experiment with various document time spans without impacting noticeably its event scope representiveness and lexicon balance.

After a set of trials with various time spans ranging from months to 15 years, the most pronounced effects of the experiment described below could be obtained by comparing two time spans - one covering 6 months and the other covering 4 years. Those sub-corpora contain over 45.000 and 350.000 documents, respectively.

V. Method

We based the data of our experiment on the term/document matrix populated with Log-Entropy weights [20]. More precisely, the value a_{ij} in the matrix corresponding to the i-th word and j-th document can be expressed as the usual ratio of a local and a global weight:

$$a_{ij} = e_i \log(t_{ij} + 1)$$

where:

$$e_i = 1 - \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$$

for n - the total number of documents, t_{ij} is the number of times the i-th word occurs in the j-th document and $p_{ij} = \frac{t_{ij}}{g_i}$, with g_i the total number of times the term i occurs in the whole corpus.

After building an LDA model (with d dimensions), we obtain a matrix v of size $n \times d$ composed of elements $v_j k$ describing how much the topic k impacts the document j. The v_{jk} matrix contains per design only non-negative values.

Each topic is in turn represented by a vector of probabilities describing how much a single word participates in this topic, in the form of a $N \times d$ matrix w (for N - the size of the lexicon).

As all the values in the matrices v and w are non-negative, their product contains the cumulative impact of each word for each document, summed over the range of all topics. For each document j and word i, what we will call the word model matrix m of size $n \times N$ composed of elements m_{ji}, is obtained by multiplying the document weights with the topic model $m = vw^{\top}$.

For a given document j we will now analyze the rank of words in the sorted vector m_{ji}, for each word i.

VI. Experiment

We based our experiment on a subset of the New York Times corpus described above. The two compared subsets were spanning 6 months and 4 years, respectively. We focused on the representation of the news item related to the accident of Kursk, the Russian submarine that sank on 12th August 2000. In order to affirm the results of our experiments, additional tests were performed, focusing on the information about the terrorist attacks launched upon New York City on the 11th September 2001. Their outcome is discussed in the second part of this Section.

The experiments were performed using 2 methods of topic modelling: LDA based on a term/document matrix populated with Log-Entropy weights and the pure Log-Entropy model based on the mentioned matrix. These models were computed basing on texts from a 4-year article span. Additionally, the Log-Entropy model was built on texts from a 6-month range in order to observe the changes resulting in considering different time windows. Finally, a ratio of Log-Entropy results from the 6-month and 4-year ranges was calculated so that we could better analyze changes that took place in models built in different time windows.

A. The accident of Kursk

Below is a fragment of the input text used for the primary experiments focusing on the accident of the Russian submarine Kursk:

> *A Russian submarine plunged to the seabed in the Barents Sea on Sunday during a naval exercise, possibly after an explosion on board, officials said today. They said the submarine was badly damaged, and was trapped at least 450 feet below the surface. They said they did not know how many of the more than 100 crew members on board were alive or how long they could survive. Tonight the navy began preparing a desperate attempt to rescue the crew. But navy officials said the odds of saving the men were slim. The submarine, called the Kursk, was not carrying nuclear weapons, the navy said, but was powered by two nuclear reactors, raising concerns about possible radioactive contamination. But Russian officials said the reactors had been turned off, and officials in Norway said a scientific vessel in the area had detected no signs of a radioactive leak. A flotilla of ships and rescue vessels was on the scene off Russia's northern coast in rough weather tonight, frantically searching for ways to reach the men. Navy officials said they intended to mount a rescue attempt on Tuesday. News reports said rescue workers had been trying to hook lines to the submarine to bring it air and fuel. If the sub lost power, the men could suffocate and the submarine's compartments could turn unbearably cold in the frigid waters. The navy's commander, Adm. Vladimir Kuroyedov, said, „Despite all the efforts being taken, the probability of a successful outcome from the situation with the Kursk is not very high." The White House spokesman, Joe Lockhart, said that President Clinton had been briefed about the accident, and that his national security adviser, Samuel R. Berger, had told Foreign Minister Igor S. Ivanov that the United States was willing to help.*

Results presented below are very similar to the ones described in [6], with the exception of unnecessary terms that carry no importance and that we were able to exclude.

In order to properly understand the results of LDA-based modelling, one has to look at the analyzed event - the sinking of the Kursk submarine - in a more general way as an accident of a naval vehicle that happened in Russia.

After analysing which words are the highest ranked in our model (30 words with the highest score are presented in Table I), it may be observed that LDA model distinguished words that may be somehow connected with:

- vehicles: ship (ranked 2nd with score 0.00418), vessel (7th, 0.00246), plane (8th, 0.00238), boat (11th, 0.00229)
- transport: port (1st, 0.00434), airline (3rd, 0.00304), flight (4th, 0.00276), airport (5th 0.00271), tunnel (15th, 0.00200), pilot (25th, 0.00169), passenger (26th, 0.00168)
- Russia: Russia (6th, 0.00252), Russian (9th, 0.00236), Moscow (28th, 0.00157)
- sea (except for the already mentioned port, ship, vessel, boat): navy (10th, 0.00232), sea (13th, 0.00209), harbor

TABLE I
TOP 30 WORDS BASED ON THE LDA MODEL (THE KURSK ACCIDENT)

no.	Word	LDA model score
1	port	0.00434231934161
2	ship	0.00417804388954
3	airline	0.00304061756597
4	flight	0.00275556563789
5	airport	0.00271146829417
6	Russia	0.00251820497727
7	vessel	0.00246363118477
8	plane	0.0023790659542
9	Russian	0.00236238563557
10	navy	0.00231835409854
11	boat	0.00228900596583
12	mile	0.0021881530589
13	sea	0.00208970459425
14	harbor	0.0020105874839
15	tunnel	0.00200078016362
16	minister	0.00196687109163
17	authority	0.00186827318326
18	profitability	0.00180866361081
19	crew	0.00180463741216
20	air	0.00178208129395
21	official	0.00175995025308
22	treaty	0.00171964506823
23	hart	0.00170921811033
24	united	0.00169356866222
25	pilot	0.00169325060507
26	passenger	0.00167712980293
27	naval	0.0016418101694
28	Moscow	0.00156616846396
29	shipping	0.00153479445129
30	state	0.00145645985742

TABLE II
TOP 30 WORDS BASED ON THE LOG-ENTROPY MODEL IN A 4-YEAR TIME WINDOW (THE KURSK ACCIDENT)

no.	Word	Log-Entropy model score
1	submarine	0.22720800350779063
2	Kursk	0.21964992269828088
3	minisub	0.2072304031428077
4	Barents	0.18764922042161675
5	navy	0.14100811315980294
6	reactor	0.1318971983154962
7	thresher	0.1257656918120723
8	Russian	0.12420286868127497
9	vessel	0.11552029612400631
10	rescue	0.11231821882564066
11	naval	0.11103796756237282
12	crew	0.10625421439440351
13	nuclear	0.10489886205812088
14	diving	0.1021749945646489
15	fleet	0.10093483510873145
16	accident	0.09974534960726877
17	hatche	0.09903148933253046
18	pressurized	0.0973239282607489
19	Russia	0.09658403552800936
20	ship	0.09603201115598564
21	baker	0.08955212916344289
22	Kuroyedov	0.08620535346215855
23	sank	0.0847013514975405
24	sea	0.08276943964465774
25	Lockhart	0.08162721527202152
26	radioactive	0.0787603000930047
27	Nilsen	0.07615887534697596
28	breathable	0.07448263140205012
29	Komsomolet	0.07436551614837245
30	stricken	0.07292194644563503

(14th, 0.00201), naval (27th, 0.00164), shipping (29th, 0.00153), water (49th, 0.00121), sailor (54th, 0.00119)

- accidents: besides many of the words already mentioned, the word crash (33rd, 0.00139) is significant.

These words are very general and are common terms used while describing some event. It has to be emphasized that there is no word specific for a given event. They were filtered out in accordance with the nature of LDA that rejects words characteristic for just a narrow set of documents and promotes words that are specific to extracted topics. Therefore, we cannot expect highly ranked terms that would be strictly connected with the accident of the Kursk submarine but rather words related generally to accidents or vehicles, transport, sea and Russia.

These words are very general and descriptive. Using them, it is not possible to state anything specific („factual”) about the nature of a given event, its causes or consequences.

Analysing the results of the Log-Entropy model calculations, we are able to see that the highest ranked words are more specific than in the case of the LDA model.

Table II presents the 30 highest ranked words according to the Log-Entropy model in 4-year time window. Among them there are ones that are related to the causes of the main event:

- *reactor* (6th, 0.13190), *nuclear* (13th, 0.10490), *radioactive* (26th, 0.07876): despite the fact that in case of Kursk accident reactors shutting down is rather a consequence, many news described also some previous submarine accidents caused by malfunction of nuclear reactors
- *accident* (16th, 0.09975): some „accident” as a reason of submarine sinking
- *pressurized* (18th, 0.09732): media reported that the lack of pressurized escape chambers was the reason why the crew was not able to get out of a submarine
- *sank* (23rd, 0.08470): „the submarine sank” as a the central event
- *stricken* (30th, 0.07292): „submarine was stricken” as a reason of the accident

Words that can also be found as related to the consequences of the discussed accident:

- *minisub* (ranked 3rd with score 0.20723): a minisub was sent with a rescue mission
- *Thresher* (7th, 0.12577): USS Thresher was a submarine, which sinking was frequently compared to the accident of Kursk
- *rescue* (10th, 0.11232) and *crew* (12th, 0.10625): rescue crew was sent in order to help sailors
- *Kuroyedov* (22nd, 0.08621): Fleet Admiral Vladimir Kuroyedov was in charge of navy when Kursk sank and therefore after the accident spoke with the media very often
- *Lockhart* (25th, 0.08163): Joe Lockhart was the White House spokesman that talked to the media after the accident of Kursk and informed about the American president's offer of help
- *Nilsen* (27th, 0.07616): Thomas Nilsen is a Norwegian researcher that wrote a report on Russian fleet. He was also interviewed by media after the accident
- *Komsomolet* (29th, 0.07437): K-278 Komsomolet was a Soviet nuclear-power submarine that was mentioned frequently in many reports on Soviet/Russian fleet after the accident of Kursk
- *stricken* (30th, 0.07292): „stricken submarine" as a consequence of the accident

These words are much more specific than in the case of those extracted by the LDA model. They strictly concern this event and describe its causes and consequences.

At first sight, the results of Log-Entropy model calculations in 6-month time window are very similar to the previous ones. We can see the same elements that we identified as the cause of the accident (*sank, stricken, reactor, nuclear*) and its consequences (eg.: *minisub, Thresher, rescue, crew, Kuroyedov*). However, the results yielded in these two time windows differ in scores. In order to analyze how the rank of particular words changed, we calculated a ratio of each word's score in two time span windows - spanning 4 years and 6 months. Having in mind that changing the time window practically does not change the local weight of a given term but changes its global weight, this ratio would emphasize these changes as a comparison of each word's global weights while similar local weights would become irrelevant.

As the Table III presents, it turned out that the words that could be used in describing causes and consequences of the Kursk sinking are now much more emphasized. Moreover, the most specific for this particular event words are stressed, while terms that could be used in descriptions of other, similar accidents (e.g. reactor, nuclear, radioactive, rescue, crew) have lower rank. Besides, new interesting words appeared when considering a ratio-based ranked list of words:

- *Kuroyedov* (ranked 3rd), *minisub* (ranked 4th), *Nilsen* (ranked 6th): they are still high ranked as the most specific words for this particular event
- *seabed* (ranked 5th): as a consequence of the accident, the submarine was plunged to the seabed
- *torpedo* (ranked 8th), *torpedoes* (ranked 22nd): an explosion of one of torpedoes that the Kursk was carrying, has been recognized as the main reason of the accident
- *Ivanov* (ranked 23rd): in time of the Kursk sinking Sergei Ivanov was the head of the Russian Security Council, therefore was highly involved in this case, so his name was often mentioned as a consequence of this accident
- *hull* (ranked 25th): after the accident, the rescue crew tried to get into the submarine through its hull
- *slim* (ranked 35th): day after day the chances of saving sailors were slimmer

TABLE III
Top 30 words based on the Log-Entropy models ratio (the Kursk accident)

no.	Word	Log-Entropy ratio
1	Kursk	1.24871629725
2	Barents	1.19631168918
3	Kuroyedov	1.17117438796
4	minisub	1.1623897656
5	seabed	1.08585777559
6	Nilsen	1.07438727067
7	Vladimir	1.0649817326
8	torpedo	1.06238346439
9	flotilla	1.06017391338
10	Thresher	1.05610904592
11	photo	1.05231801755
12	certified	1.05071293879
13	outcome	1.04815882266
14	avalon	1.0380988501
15	sailor	1.03431234091
16	periscope	1.02936384958
17	fuel	1.02728401954
18	site	1.02714550054
19	hatche	1.02449383024
20	submarine	1.023577913
21	underwater	1.02209583656
22	torpedoes	1.01904285092
23	Ivanov	1.01647340538
24	doubtful	1.01608381017
25	hull	1.01556031012
26	naval	1.01495656826
27	Joe	1.01339486585
28	Baker	1.0120830858
29	sunk	1.01187373966
30	sunken	1.01048072834

It seems very interesting how calculating of the ratio helped with finding new words describing causes and consequences and how it distinguished terms that are specific for a given event. It also emphasized changes that occurred in different time windows.

B. September 11 terrorist attacks

Some additional tests needed to be launched in order to affirm results obtained in the previously performed experiments.

A fragment below exemplifies the input text used in the subsequent experiments, focused on the September 11 terrorist attacks on New York City:

> *Hijackers rammed jetliners into each of New York's World Trade Center towers yesterday, toppling both in a hellish storm of ash, glass, smoke and leaping victims, while a third jetliner crashed into the Pentagon in Virginia. There was no official count, but President Bush said thousands had perished, and in the immediate aftermath the calamity was already being ranked the worst and most audacious terror attack in American history. The attacks seemed carefully coordinated. The hijacked planes were all en route to California, and therefore gorged with fuel, and their departures were spaced within an hour and 40 minutes. The first, American Airlines Flight 11, a Boeing 767 out of Boston for Los Angeles, crashed into the north tower at 8:48 a.m. Eighteen minutes later, United Airlines Flight 175, also headed from Boston to Los Angeles, plowed into the south tower. Then an American Airlines Boeing 757, Flight 77, left Washington's Dulles International Airport bound for Los Angeles, but instead hit the western part of the Pentagon, the military headquarters where 24,000 people work, at 9:40 a.m. Finally, United Airlines Flight 93, a Boeing 757 flying from Newark to San Francisco, crashed near Pittsburgh, raising the possibility that its hijackers had failed in whatever their mission was. There were indications that the hijackers on at least two of the planes were armed with knives. Attorney General John Ashcroft told reporters in the evening that the suspects on Flight 11 were armed that way.*

TABLE IV
TOP 30 WORDS BASED ON THE LDA MODEL (THE SEPTEMBER 11TH TERRORIST ATTACKS)

no.	Word	LDA model score
1	attack	0.00270258155503
2	Bush	0.00265822334759
3	anthrax	0.00202525476511
4	police	0.00196184349059
5	Afghanistan	0.00190457356009
6	American	0.00183689824487
7	Taliban	0.00182571967641
8	war	0.00176326965287
9	airline	0.00174704614254
10	bin	0.00171653103426
11	official	0.00171409002157
12	united	0.00165387405236
13	military	0.0016267486976
14	state	0.00160529481878
15	Laden	0.00148723727284
16	people	0.00147777652497
17	government	0.00145696458969
18	terrorist	0.00142711844383
19	city	0.00139962463902
20	security	0.00139337982458
21	world	0.00136066739187
22	president	0.00124062013179
23	plane	0.00122558021987
24	York	0.00120866113945
25	airport	0.00118486349022
26	nation	0.00115097052471
27	center	0.0011502315111
28	flight	0.00113663752416
29	force	0.00112785897603
30	time	0.00107041659097

Table IV presents top 30 results of LDA-based modelling performed in a 4-year time span window. As previously, we are able to perceive some groups of words, linked together by a certain topic:

- war: attack (ranked 1st with score 0.00270), war (8th, 0.00176), military (13th, 0.00163), force (29th, 0.00113)
- United Stated of America: Bush (2nd, 0.00266), American (6th, 0.00184), York (24th, 0.00121)
- terrorism (except for already mentioned attack, force): anthrax (3rd, 0.00203), terrorist (18th, 0.00143)
- public service: police (4th, 0.00196), security (20th, 0.00139), firefighter (38th, 0.00099)
- Afghanistan: Afghanistan (5th, 0.00190), Taliban (7th, 0.00183)
- aircraft: airline (9th, 0.00175), plane (23rd, 0.00123), airport (25th, 0.00118), flight (28th, 0.00114)
- society: state (14th, 0.00161), government (17th, 0.00146), president (22nd, 0.00124), nation (26th, 0.00115), administration (32nd, 0.00104), country (42nd, 0.00098)

As it was observed previously, LDA-based modelling rejects words that are specific for a given event. The highest ranked terms are general and cannot be linked with any factual information. As one can see, there are no words that are characteristic for the September 11 terrorist attacks but rather words that could be related to any document focused on the subject of war, terrorism, United States of America and so on.

These conclusions are very similar to the ones drawn in case of the previous experiments.

The results of Log-Entropy model calculations are also analogous to the case of documents related to Kursk accident, including the possibility of distinguishing causes and consequences of a given event, however we decided not to present them for the reason of shortening the paper.

Bigger expressiveness might be attributed to the ratio of each word's Log-Entropy model score in two time span windows - spanning, as previously, 4 years and 6 months.

Table V presents 30 words with the highest ratio of Log-Entropy model score in two forementioned time span windows. As it might be noticed, there are much more terms that are related to the particular event - September 11 terrorist

TABLE V
TOP 30 WORDS BASED ON THE LOG-ENTROPY MODELS RATIO (THE SEPTEMBER 11 TERRORIST ATTACKS)

no.	Word	Log-Entropy ratio
1	terrorist	1.57893538599
2	attack	1.42460445953
3	Afghanistan	1.42152193493
4	Osama	1.40325332485
5	Taliban	1.36639135448
6	Afghan	1.3476264106
7	bin	1.3345158298
8	hijacker	1.32412486848
9	Kabul	1.31277718096
10	Laden	1.31109812221
11	hijacked	1.26364453579
12	terror	1.25068637547
13	Pentagon	1.22267231541
14	hijacking	1.22261266319
15	trade	1.22143888629
16	Bush	1.19449256298
17	aftermath	1.19022203646
18	Islamic	1.18133136868
19	firefighter	1.16997652186
20	tower	1.16270222181
21	Ashcroft	1.16229315062
22	jetliner	1.14621366889
23	rubble	1.13927570543
24	inhalation	1.13484053252
25	plane	1.12086607749
26	disaster	1.09775233687
27	twin	1.08233678639
28	airline	1.08152641944
29	Vesey	1.08150236466
30	militant	1.07469496007

attacks. In this case, more general words that could be used in any other description of attack, war, etc. are less stressed.

Moreover, we are able to distinguish words that could be considered as causes and consequences of the given event:

- *Osama* (ranked 4th with score 1.40325), *bin* (7th, 1.33451), *Laden* (10th, 1.31110): Osama bin Laden was the founder of terrorist organisation al-Qaeda which was responsible for launching the attacks
- *Afghanistan* (3rd, 1.42152), *Afghan* (6th, 1.34763), Kabul (9th, 1.31278): the war in Afghanistan (with the capital in Kabul) was one of the consequences of terrorist attacks on 11th of September 2001
- *hijacker* (8th, 1.32412), *hijacked* (11th, 1.26364453579), *hijacking* (14th, 1.22261): the planes were used as destructive weapons, because they were hijacked
- *aftermath* (17th, 1.19022): the usage of this word indicates an introduction of a given event's consequences
- *Ashcroft* (21st, 1.16229): on 11th of September 2001 John Ashcroft was an Attorney General who, in consequence of the terrorist attacks, was a supporter of passage of one of the main antiterrorism acts (USA Patriot Act)
- *rubble* (23rd, 1.13928), *disaster* (26th, 1.09775), *crashed* (31st, 1.07208), *perished* (36th, 1.06022): these are some words used to describe the consequences of an attack
- *rescue* (33rd, 1.07091), *rescuer* (35th, 1.06506), *evacuated* (39th, 1.05352): in a consequence of the terrorist attack, rescue crews tried to help people and evacuate then from the World Trade Center

Again, it proved that Log-Entropy model discerns more factual data than LDA-based model. Moreover, calculating of the ratio of score obtained in two time span windows helped us to find new interesting terms and stress the changes of results in different time windows.

VII. CONCLUSION

In this paper we extended our work introduced in [6] where we introduced the concept of the text being a structure consisting of a mixture of event descriptions and factual information. Additional experiments performed on the second subset of the large-scale corpus proved again that some methods of postprocessing the results of unsupervised methods could help model an event in a semantically meaningful way, reflecting its semantic structure. Moreover, by comparing the results of Latent Dirichlet Allocation (LDA) and Vector Space Model methods we were able once more to observe how the former distinguished descriptive and general information, while the latter emphasized more specific terms. This specific information could be useful in description of event's causes and consequences.

However, it has to be stressed that the method of discerning causes and consequences of a given event is not a subject of our work and would be an interesting topic of future work. In our paper we tried to distinguish causes and consequences more or less accurately without any advanced method.

VIII. FUTURE WORK

This paper presents the new experiments and affirms the authors' hypothesis discussed already in [6] that by comparing the results of topic modelling and vector modelling coming from different subsets of a corpus, varying by time scope and size, the obtained information can be additionally graded by the level of its generality or specificity. That in turn can show us a way to create a method for discerning semantic prototypes (general description of events) from factual information (specific to events).

However, as seen in the preliminary results above, this hypothesis is supported by manually verified examples that do not scale to a more generic case. Thus, the current ongoing research focuses on creating a metric being able to assess automatically the level of generality or the amount of facts in a specific result. Such a metric takes into consideration factors related to the relative amount of Named Entities in the results, the distance from various text clusters obtained via topic modelling etc. By defining such a metric, crucial parameters for a correct fact versus semantic prototype extraction method can be automatically determined. Some of the parameters currently considered are: the chosen time window relative and absolute sizes (the analyzed corpus covers over 30 years of press notes), the time shift of the window time frame relative to the analyzed event (preceding, succinct or just surrounding), the topic modelling methods settings.

The long term goal of this research is the creation of a method for automatic semantic prototype identification. Pure unsupervised methods, as those presented in this paper, are not the only venue of approach considered. A parallel research is conducted, based on human based association networks, as presented in [4]. We expect to obtain valuable results coming from the convergence of both approaches.

Acknowledgments

The research reported in the paper was partially supported by grants "Advanced IT techniques supporting data processing in criminal analysis" (No. 0008/R/ID1/2011/01) and "Information management and decision support system for the Government Protection Bureau" (No. DOBR-BIO4/060/13423/2013) from the Polish National Centre for Research and Development.

References

[1] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2005. [Online]. Available: http://psiexp.ss.uci.edu/research/papers/SteyversGriffithsLSABookFormatted.pdf

[2] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. Suppl. 1, pp. 5228–5235, April 2004. doi: 10.1073/pnas.0307752101. [Online]. Available: http://dx.doi.org/10.1073/pnas.0307752101

[3] J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, and D. Blei, "Reading tea leaves: How humans interpret topic models," in *Neural Information Processing Systems (NIPS)*, 2009.

[4] I. Gatkowska, M. Korzycki, and W. Lubaszewski, "Can human association norm evaluate latent semantic analysis?" in *Proceedings of the 10th NLPCS Workshop*, 2013, pp. 92–104.

[5] T. Wandmacher, "How semantic is latent semantic analysis?" in *Proceedings of TALN/RECITAL*, 2005.

[6] M. Korzycki and W. Korczyński, "Does topic modelling reflect semantic prototypes?" in *New Research in Multimedia and Internet Systems*, ser. Advances in Intelligent Systems and Computing, A. Zgrzywa, K. Choroś, and A. Siemiński, Eds. Springer International Publishing, 2015, vol. 314, pp. 113–122. ISBN 978-3-319-10382-2. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10383-9_11

[7] E. Rosch, "Principles of categorization," in *Cognition and categorization*, E. Rosch and B. Lloyd, Eds. Hillsdale, New Jersey: Erlbaum, 1978, pp. 27–48.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. 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. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

[9] D. Ortega-Pacheco, N. Arias-Trejo, and J. B. B. Martinez, "Latent semantic analysis model as a representation of free-association word norms." in *MICAI (Special Sessions)*. IEEE, 2012. doi: 10.1109/MICAI.2012.13. ISBN 978-1-4673-4731-0 pp. 21–25. [Online]. Available: http://dx.doi.org/10.1109/MICAI.2012.13

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[11] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[12] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002. ISBN 1-55860-897-4 pp. 352–359. [Online]. Available: http://dl.acm.org/citation.cfm?id=2073876.2073918

[13] M. Wettler, R. Rapp, and P. Sedlmeier, "Free word associations correspond to contiguities between words in texts." *Journal of Quantitative Linguistics*, vol. 12, no. 2-3, pp. 111–122, 2005. doi: 10.1080/09296170500172403. [Online]. Available: http://dx.doi.org/10.1080/09296170500172403

[14] T. Wandmacher, E. Ovchinnikova, and T. Alexandrov, "Does latent semantic analysis reflect human associations?" in *Proceedings of the Lexical Semantics workshop at ESSLLI'08*, 2008.

[15] R. C. Schank, "Conceptual dependency: A theory of natural language understanding," *Cognitive Psychology*, vol. 3, no. 4, pp. pages 532–631, 1972. doi: 10.1016/0010-0285(72)90022-9. [Online]. Available: http://dx.doi.org/10.1016/0010-0285(72)90022-9

[16] S. L. Lytinen, "Conceptual dependency and its descendants." *Computers and Mathematics with Applications*, vol. 23, pp. 51–73, 1992. doi: 10.1016/0898-1221(92)90136-6. [Online]. Available: http://dx.doi.org/10.1016/0898-1221(92)90136-6

[17] W. Lubaszewski, K. Dorosz, and M. Korzycki, "System for web information monitoring," in *Computer Applications Technology (ICCAT), 2013 International Conference on*, Jan 2013. doi: 10.1109/ICCAT.2013.6522053 pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/ICCAT.2013.6522053

[18] K. Dorosz and M. Korzycki, "Latent semantic analysis evaluation of conceptual dependency driven focused crawling," in *Multimedia Communications, Services and Security*, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2012, vol. 287, pp. 77–84. ISBN 978-3-642-30720-1. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-30721-8_8

[19] K. Leetaru, "Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space," *First Monday*, vol. 16, no. 9, 2011.

[20] T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, Eds., *Handbook of Latent Semantic Analysis*, ser. University of Colorado Institute of Cognitive Science Series. Mahwah, New Jersey, USA: Lawrence Erlbaum Associates, 2007. ISBN 9780805854183