Bernadette Sharp, Wiesław Lubaszewski
and Rodolfo Delmonte (eds)

Natural Language Processing and Cognitive Science

Proceedings 2015

CAFO
SCAR
INA_

Editors

Bernadette Sharp, Staffordshire University, U.K.
Wiesław Lubaszewski, Jagiellonian University, Poland
Rodolfo Delmonte, Ca' Foscari University, Italy

Natural Language Processing and Cognitive Science
Bernadette Sharp, Wiesław Lubaszewski and Rodolfo Delmonte (eds)

ISBN 978-88-7543-394-9

Libreria Editrice Cafoscarina
Dorsoduro 3259, 30123 Venezia
www.cafoscarina.it

Prima edizione luglio 2015

Contents

Hierarchies of Terms on the Euromaidan Events: Networks and Respondents' Perception

D. Lande, A. Snarskii, E. Yagunova, E. Pronoza and S. Volskaya

Institute for Information Recording NAS of Ukraine, Kiev, Ukraine
NTUU "Kiev Polytechnic Institute", Kiev, Ukraine
{dwlande, asnarskii}@gmail.com
Saint-Petersburg State University, Saint-Petersburg, Russian Federation
{iagounova.elena, katpronoza, svetlana.volskaya}@gmail.com

Abstract. In this paper we describe the construction methodology of a network of natural terms hierarchy on the base of the subject arrays of news texts. The proposed method is illustrated using automatic processing of the full texts of the articles about the Euromaidan events in Kiev.

1 Introduction

Constructing a large domain-specific ontology is a challenging problem. The ontology development process includes such a task as terms learning, but the problem of effective unsupervised terms learning is unsolved, and the problem of the links identification and automatic network construction is also still open.

Another important task is the formal estimation of the number of new topics in data streams. Appearance of new topics naturally causes appearance of the series of terms marking new themes. A linguist dealing with news texts has to know the specifics of different segments of media data streams. Particularly, sometimes one can correlate separate news topics with the subjects of whole data flows using lexical features.

In this paper we propose an approach to the construction of a terminological basis for interrelated events, which are described in the messages of electronic media, and for separate subjects of data flows for a certain time period. We also consider some principles of making a language network on the base of the selected terms. Correlation of unit message terminology with general subject terminology can be considered as a formal criterion of event relevance to the considered subject area (sequence of events).

The problems of events modeling and analyzing their perception by the informants have been an object of many recent studies [0]. Unsupervised terms extraction task is also widely addressed by the researchers. Terms extraction methods are either statistics-based (e.g., clustering [0]), or use fine-grained linguistic analysis (e.g., dependency parsing [0]). Some researchers also employ external sources of knowledge like Wikipedia or Wordnet [0]. Our method is statistics-based. It is fast and language independent and does not demand any linguistic resources.

2 Data

The data for our research consists of news reports about the confrontation in Kiev in 2013-2014, which was caused by so-called Euromaidan. We collected more than 200 thousand of news reports from RuNet web sites during the period from November 2013 till March 2014.

First of all, it is necessary to choose a text corpus for the further analysis. To collect the data for our research, we use “InfoStream” – a system of content monitoring. To retrieve the news reports which are relevant to the subject area we make the following request:

(maydan|euromaidan)&(beat|dispersal|storm|
berkut|molotov|titushk|was killed) & lang.RUS.

The collected corpus consists of more than 200 thousand of news reports. On the base of the corpus the dynamics of subject reports should be identified. The mode «Dynamics of events» in the system of content monitoring «InfoStream» allows getting information about the number of published articles which are relevant to the request for a certain time period. This information is presented in the form of a plot (see Fig. 1).

Fig. 1. Dynamics of the number of publications which are relevant to the request

The time dynamics data is normalized for each day, and time series is built. Each relative frequency value in this series equals the ratio of the number of subject reports per day to the number of all the reports per day. It allows us to ignore weekly periodicity in the number of subject reports.

After we get the information about publications dynamics, the critical points should be identified. These critical points are the local maxima of time series in the dynamics of publications [3].

On the base of these results three dates were chosen (2013.11.30, 2014.01.22, 2014.02.19) as critical points for the sequence of events under consideration.

After the critical points are selected, it is necessary to extract the main sequences of subject reports which are relevant to the request for the necessary dates (see Fig. 2). It is also done via the system of content monitoring.

Fig. 2. Main subject concatenations for necessary dates

3 Construction method of Network of Natural Term Hierarchy

3.1 Extraction of Terms for Ontology

For the further analysis we build three corpora from the reports. Each corpus corresponds to one of the three found critical points; lexical features of each corpus are the objects of monitoring.

Preprocessing of these corpora includes division of text into fragments (separate reports, paragraphs, sentences, words, bigrams, and trigrams), deletion of analphabetical symbols and cutting off inflections – stemming (this is an option).

Then each term from the text (unigram, bigram or trigram) receives an estimation of its «discriminant power», represented by TF-IDF. The preliminary technique description was published in [12].

3.2 Construction of Terms Hierarchy

The process of network constructing is based on using semantically important text elements. To identify these elements in the text one can use methods described in [0], [0] and [0]. An advantage of a network built on the base of important text elements, pivot words and words combinations is that such a network embraces separate knowledge domains.

Extracting of the terms for a network is done using the feature based on the discriminant power of words. Nevertheless one should remember that this feature cannot guarantee high quality of ontology. Most frequent words from the chosen subject area, which have low discriminant power (for example, the words “Ukraine”, “Maidan", "Protest" in the news corpus about Euromaidan in Kiev) could be the most important ones for the network construction.

The content of the corpora is the base of the future network. In this work we consider a natural network. We call the network natural due to the fact that its construction does not include any special methods of semantic analysis (including part of speech tagging). All the relations in this network are determined by the positions of the words and word combinations, which are extracted from the texts of statistically significant size. Terms hierarchy which is built completely automatically is the base for the further automatic ontology construction with experts.

In our work we propose a method of constructing terms hierarchy which includes the construction of a compactified horizontal visibility graph (CHVG) and terms weights recalculation (for unigrams, bigrams and trigrams) [0].

Language network is built in three stages using the CHVG algorithm.

1. In the first stage nodes sequence is marked on the horizontal axis. Each node corresponds to the word in the order it appears in the text. On the vertical axis TF-IDF weights are put. Vertical lines are drawn between these TF-IDF values and their projections on the x-axis.

2. In the second stage a traditional horizontal visibility graph is built [0]. An edge is drawn between every two nodes if these nodes are in "direct visibility". "Direct visibility" of the nodes means that they can be connected by a horizontal line which does not intersect any vertical line in the plot.

3. In the third stage we merge the nodes with the same words. The edges of such nodes are also merged. Such procedure is called graph compactification. Node weights are recalculated. TF-IDF values are replaced with the corresponding node degrees in CHVG. Finally, the terms are sorted according to their new CHVG weights in descending order. Stop words are excluded from further analysis. In this paper a list of stop words is formed using following web-resources:
https://code.google.com/p/stop-words/downloads/list,
http://www.ranks.nl/stopwords/, http://www.textfixer.com/resources/common-english-words.txt.

Experts estimate the size of the network (let us denote it by N). Then N unigrams, N bigrams and N trigrams with the largest CHVG weights are selected. The network is constructed using the obtained terms. In this network nodes identify terms and links represent part-whole relations between the terms. Fig. 3 presents an example of the terms hierarchy construction. Different geometric figures denote different words in Fig. 3. Unigrams are grouped in the first column, while bigrams and trigrams are in the second and third columns respectively. If a unigram belongs to some bigram, or a bigram is a part of some trigram, an arrow is drawn between them (denoting a part-whole relation). The set of terms together with the links between them forms a three-level Natural Network of Terms Hierarchy [0], [12].

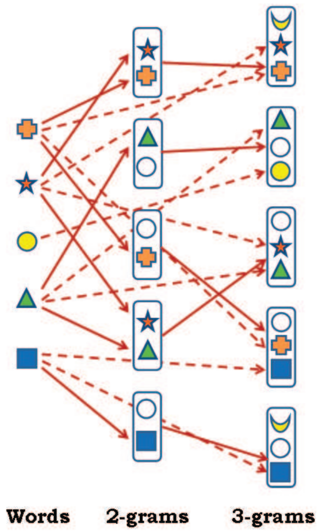


Fig. 3. Relations construction in a three-level hierarchy

3.3 Visualization of Network of Natural Term Hierarchy

We select top-20 Euromaidan terms (unigrams, bigrams and trigrams) with the largest CHVG weights to visualize the network we build. These terms are presented in Table 1.

Table 1. Top-20 Euromaidan terms with the largest CHVG weights

Unigram	Bigram	Trigram
Украина /Ukraine/	Виктор Янукович /Viktor Yanukovych/	президент Виктор Янукович /President Viktor Yanukovych/
Киев /Kiev/	центр Киева /Centre of Kiev/	сотрудники правоохранительных органов / law enforcement officials /
власть /Power/	верховная Рада /Verkhovna Rada/	введение чрезвычайного положения /Introduction of state of emergency/
страна /State/	улица Грушевского /Grushevskogo Street/	батькивщина Арсений Яценюк /Batkivshina Arseniy Yatsenyuk/
Янукович /Yanukovych/	президент Украины /President of Ukraine/	Олимпийские игры Сочи /Olympic Games Sochi/
Майдан /Maidan/	Майдан Независимости /Maidan Nezavisimosti (Independence Square) /	Глава Администрации Президента /Head of presidential administration/
люди /People/	партия регионов /The party of regions/	фракция партии регионов /The party of regions fraction/
милиция /Police/	пресс-служба /Press centre/	штаб национального сопротивления /National resistance

Unigram	Bigram	Trigram
		headquarters/
Беркут /Berkut/	Арсений Яценюк /Arseniy Yatsenyuk/	действие благодати Пресвятой /Holy Grace effect/
оппозиция /Opposition/	Михайловская Площадь /Mikhailovskaya Square/	Майдан Незалежности Киев /Maidan Nezalezhnosti Kiev /
президент /President/	лидеры оппозиции /Opposition leaders/	страницы социальных сетей /Social network pages/
Яценюк / Yatsenyuk/	разгон Евромайдана /Euromaidan dispersal/	УДАР Виталий Кличко /UDAR Vitali Klitschko/
украинский /Ukrainian/	объявление перемирия /Armistice announcement/	Германия Франция Великобритания /Germany France UK/
Евромайдан /Euromaidan/	Виталий Кличко /Vitali Klitschko/	улица Грушевского Киев /Grushevskogo Street Kiev/
штурм /Attack/	Майдан Незалежности /Maidan Nezalezhnosti /	офис партии регионов /Office of the party of regions/
акция /Act/	акция протеста /Act of protest/	михайловская площадь киев /Mikhailovskaya Square Kiev/
здание /Building/	правый сектор /Right Sector/	силовой разгон евромайдана /Military dispersal of Euromaidan/
активист /Activist/	огнестрельное оружие /Firearms/	беркут внутренние войска /Berkut the internal troops/
МВД /Ministry of Internal Affairs/	правоохранительные органы /Law machinery/	премьер николай азаров /Premiere Mykola Azarov/
площадь /Square/	штурм зачистка /Attack cleanup/	мирная акция протеста /Peaceful protest act/
улица /Street/	штурм майдана /Attack of Maidan/	здание верховной рады /Verkhovna Rada building/
Грушевского /Grushevskogo/	внутренние войска /The internal troops/	законная власть Украины /Ukraine's legitimate government/
лидер /Leader/	применение силы /Use of force/	лидер партии УДАР /Leader of UDAR party/

Finally when the terms hierarchy network is constructed, we visualize it using Gephi tool (https://gephi.org). To load the network into a database we represent it by an incidence matrix in ".csv" format.

4 Results

To illustrate the final network we present a small fragment of 20 terms (20+20+20 in total) in Fig. 4.

It can be noticed that the words in large print (Киев / Kiev, президент / president, Майдан / Maidan) in Fig. 4 are the topmost terms from Table 1. These words represent the nodes with the highest weights. Unigram nodes are connected with bigram and trigram nodes, and bigram nodes are connected with trigram ones. Arc thickness is proportional to the joint frequency of the terms (i.e., n-grams) it unites.

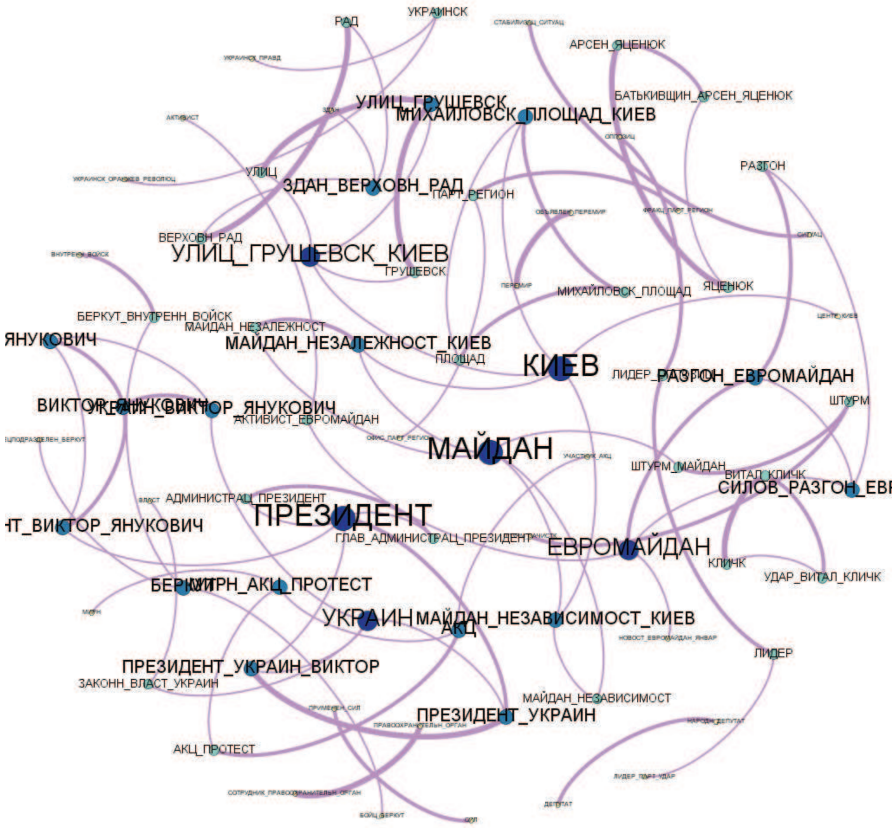


Fig. 4. Euromaidan Natural Network of Terms Hierarchy example (20+20+20)

We should also consider a larger Euromaidan network fragment (200+200+200), which is presented in Fig. 5. In Fig. 5 it can be seen that in spite of the large density of this fragment the terms “Киев” / Kiev and “Майдан” / Maidan remain in large print. Meanwhile the unigram “Президент” / President is replaced by the term “Беркут” / Berkut. It can be explained by the fact that the unigram “Беркут” / Berkut has higher weight than “Президент” / President.

After experimenting with networks of different sizes we deduced that node degree distribution (for outgoing links only) follows power law ($p(k) = Ck^{\alpha}$). It means that such networks are scale-free (see Fig. 4). Power coefficient α varies from 2.1 to 2.3 for networks of different sizes (e.g., from 20+20+20 to 500+500+500) that in general complies with Language Networks structure [0].

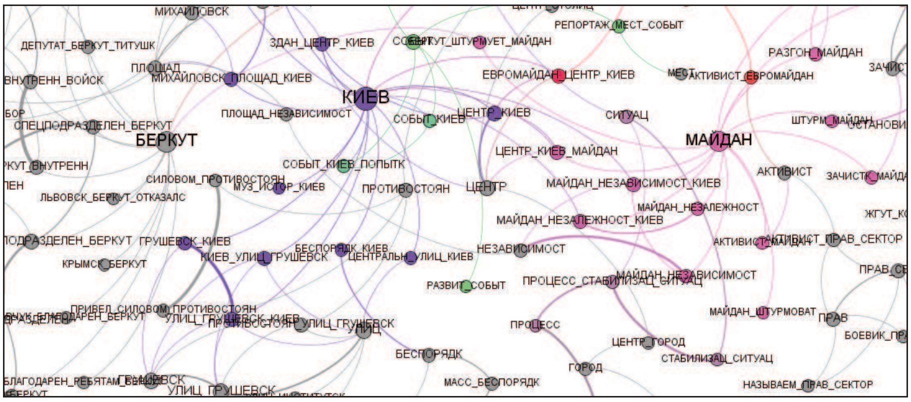


Fig. 5. Larger network fragment (200+200+200) visualized using Gephi

It also turned out that according to the proposed algorithm one node can have 5 ingoing links at most (for the network in our example, see Fig. 4). Single words (unigrams) have 0 ingoing links, bigrams – 2 ingoing links at most and trigrams – 5 ingoing links at most (with 3 links inherited from each word of a trigram and other 2 inherited from the two bigrams a trigram consists of). Top-20 nodes with the largest ingoing degrees for the 200+200+200 network of natural terms hierarchy are presented in Table 2.

Table 2. Top-20 nodes with the largest ingoing degree

Outgoing degree	Node
5	участники акции протеста /Protest act participants/
5	улица Грушевского Киев /Grushevskogo Street Kiev/
5	(президент) Украины Виктор Янукович /(President of) Ukraine Viktor Yanukovych/
5	силовой разгон Евромайдана /Military dispersal of Euromaidan/
5	мирная акция протеста /Peaceful protest act/
5	глава администрации президента /Head of presidential administration/
5	фракция партии регионов /The party of regions fraction/
5	бойцы спецподразделения Беркут /Berkut special units/
5	Батькивщина Арсений Яценюк /Batkivshina Arseniy Yatsenyuk/
4	администрация президента Украины /Ukrainian Presidential administration/
4	здание Верховной Рады /Verkhovna Rada building/
4	здания центра Киева /Buildings of the centre of Kiev/
4	Верховная Рада Украины /Verkhovna Rada of Ukraine/
4	УДАР Виталий Кличко /UDAR Vitali Klitschko/
4	сотрудники спецподразделения Беркут /Berkut officers/
4	сотрудники правоохранительных органов /Law machinery officers/
4	силовой разгон митингующих /Military dispersal of meeting participants/
4	политический кризис Украина /Political crisis Ukraine/
4	применение силы сторонами /Use of force by the parties/
4	пресс-служба МВД /Press centre of Ministry of Internal Affairs/

The nodes with the largest ingoing degree are also semantically the most important ones. They include the following word combinations: “участники акции протеста” /Protest act participants/; “улица Грушевского Киев” /Grushevskogo Street Kiev/; “силовой разгон Евромайдана” /Military dispersal of Euromaidan/; “мирная акция протеста” /Peaceful protest act/; “бойцы спецподразделения Беркут” /Berkut special units/.

CHVG values are calculated for single subjects as well, and the network is constructed for them. In Fig. 6 three network examples are shown. Their interrelation network is given in Fig. 7.

Our assumptions regarding the importance of the selected events for network constructing were confirmed during the experiments with informants. Each informant was given a standard instruction: “Remember the recent events in the world. Write down 10-15 words which are best to describe these events”. More than 40 informants were questioned [11].

In Table 3 the results of the experiments with informants are shown. In spite of the fact that the informants were not asked to describe the events in Ukraine, the majority of them still speak about the Euromaidan events.

Table 3. Significance of the selected events (results of public opinion poll)

%	Events. Informants under 30 ages	%	Events. Informants of 30 ages and older
39	The joining of Crimea to the Russian Federation	58	Winter Olympic Games in Sochi (Russia), the joining of Crimea to the Russian Federation
24	Disturbances in Maidan		
18	Olympic Games in Sochi	40	Disturbances in Maidan
14	Excellent results of the Russian team in the Winter Olympic Games, Referendum in Crimea, Sanctions against Russia, Civil war in Ukraine	33	Referendum in Crime
		23	Excellent results of the Russian team in Winter Olympic Games
		20	Civil war in Ukraine, Sanctions against Russia
12	Murders of civilian residents in Ukraine	13	Murders of civilian residents in Ukraine, War in the East of Ukraine
8	Beginning of combat operations in the Donetsk republic, Escape of Yanukovich from the country	8	Escape of Yanukovich from the country, Beginning of combat operations in the Donets republic, Revolution in Ukraine, Deceitful propaganda in Russian media, Little green men in Crimea, Crisis in Ukraine

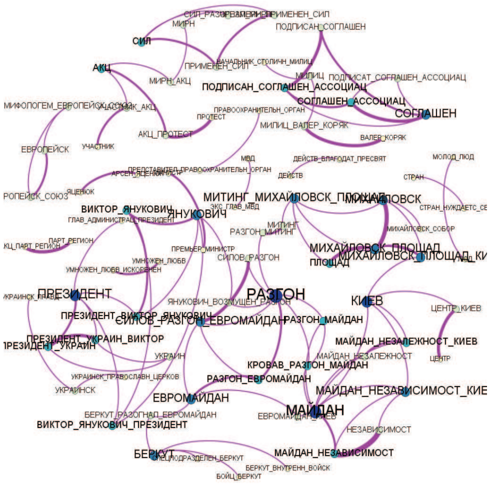
It is important to note that all the informants were divided into two groups according to their age. People of older age turn out to be quite critical while estimating the events of Euromaidan. On the other hand, one can find a large number of appraisal words in the answers of the younger group.

To confirm that the keywords and word combinations which we got as terms during the process of network construction are really semantically important for our theme we conducted another experiment with informants. We asked them to define the subject which these keywords can be connected with.

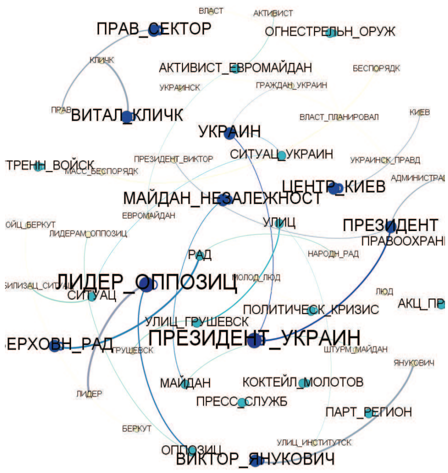
There were 7 informants in total and all of them were sure that these keywords were extracted from the texts about the events of Euromaidan in Kiev.

Such high level of agreement is caused by the high subject homogeneity of the corpus which we chose to analyze. In fact, all the texts within the corpus describe the same event.

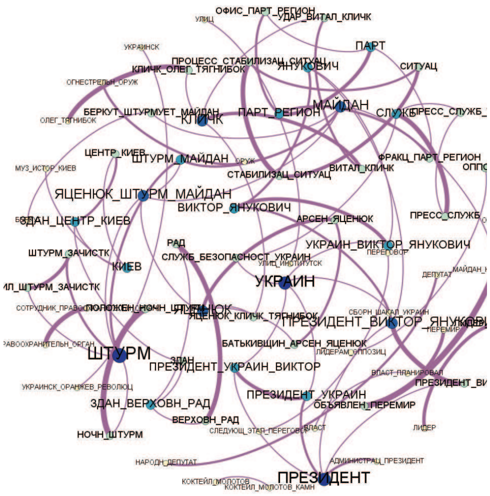
In Fig. 7 it is shown that a set of terms corresponds to each subject (a node identified with a date). The terms which take place on several different dates can be seen in the central part of the network while those which are more specific appear at the periphery.



a



b



c

Fig. 6. Euromaidan network (20+20+20) (a – 2013.11.30, b – 2014.01.22, c – 2014.02.19)

Central zone does not necessarily include all the terms from all the subjects – it is enough to include some portion of a subject's terms, e.g., one half. The more terms the central zone of a subject includes, the closer its content is to the main events trend, and the more relevant it is. In our example the "2014.01.22" node is the most relevant to the general events trend (see Fig. 7).

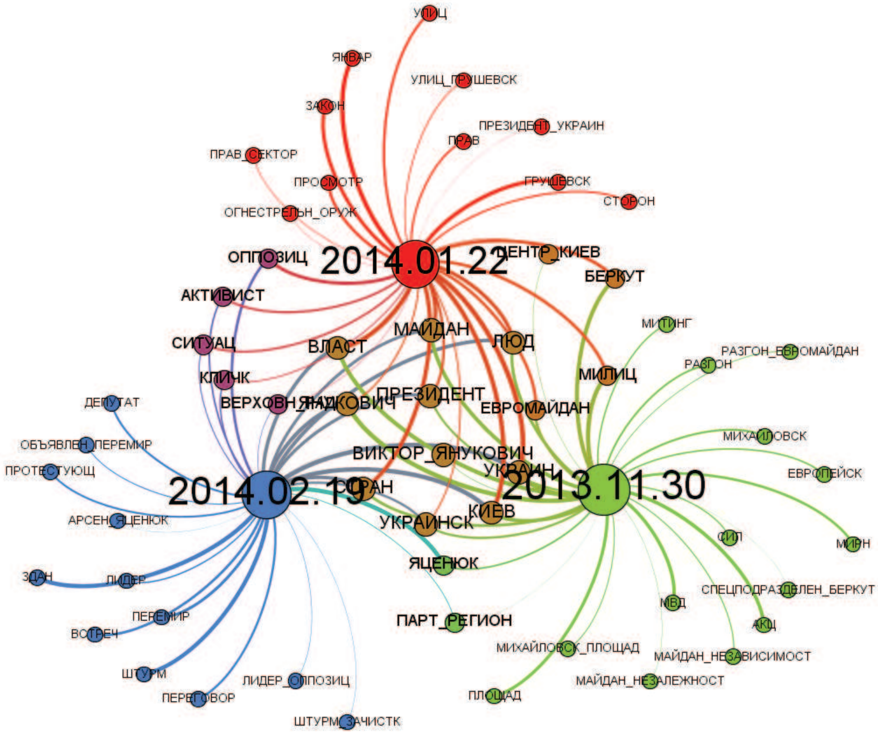


Fig. 7. Euromaidan terms interrelations for the three chosen dates

We also propose a linguistic criterion of subject relevance: the more terms of a paper are in the central zone of the term interrelation network the more relevant this subject is to the general events theme. In other words, subject relevance is proportional to the number of its terms in the central zone of the term interrelation network.

5 Conclusion

As a result of the research:

- an algorithm of constructing a network of natural terms hierarchy based on corpus analysis is proposed;
- the algorithm is illustrated with the examples of a Euromaidan-related network;
- network of natural terms hierarchy appears to be scale-free while considering outgoing links;

- programming tools for visualization of a network of natural terms hierarchy are introduced;
- the criterion of subject relevance to the event is proposed;
- the verification of this criterion according to the informants opinion is proposed.

Language network, constructed according to the proposed method, can be used as 1) a basis for ontology construction (e.g., for Ukrainian acts of protest theme), 2) a tool for database navigation and 3) a tool for organizing user prompts in information retrieval systems.

Our future work includes constructing networks on the base of a less homogeneous corpus. We are already working on the improvement of the estimation of our results, involving more informants with more complex stratification (country, region, profession and so on).

Acknowledgments

The authors acknowledge Saint-Petersburg State University for a research grant 30.38.305.2014. Special thanks to Vlad Kotov for sharing materials with informants.

References

Bolshakova, E.I., Klyshinsky, E.S., Lande, D.V., Noskov, A.A., Peskova, O.V., Yagunova, E.V. (2011) *Avtomaticheskaya Obrabotka Tekstov na Estestvennom Yazike i Kompuernaya Lingvistika*a. Moscow, MIEM, 272 p. Title in Russian: Автоматическая обработка текстов на естественном языке и компьютерная лингвистика.

Drymonas, E., Zervanou, K., Petrakis, Euripides G.M. (2010) Unsupervised Onthology Acquisition from Plain Texts: The OntoGain System. *Lecture Notes in Computer Science*, vol. 6177, pp. 277–287.

Lande, D. V., Snarskii, A. A. (2013) Compactified HVG for the Language Network. *Proceedings of the International Conference on Intelligent Information Systems*: The Conference is dedicated to the 50th anniversary of the Institute of Mathematics and Computer Science, 20–23 Aug. 2013, Chisinau, Moldova: Proceedings IIS / Institute of Mathematics and Computer Science, pp. 108–113.

Lande, D.V. (2014) Building of Networks of Natural Hierarchies of Terms Based on Analysis of Texts Corpora. E-preprint viXra 1404.0069.

Lande, D.V., Snarskii, A.A., Yagunova, E.V., Pronoza, E.V. (2013) The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text. *Proceedings of the 12th Mexican International Conference on Artificial Intelligence*, pp. 209–215.

Liu H., Salerno J., Young M. Berlin (eds) (2009) *Social Computing and Behavioral Modeling*. Springer, 280 p.

Luque, B., Lacasa, L., Ballesteros F., Luque, J. (2009) Horizontal Visibility Graphs: Exact Results for Random Time Series. *Phys. Review E*, pp. 046103-1– 046103-11.

Poon, H., Domingos, P. (2010) Unsupervised Ontology Induction from Text. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 296–305.

Wong, W., Liu, W. & Bennamoun, M. (2012) Ontology Learning from Text: A Look back and into the Future. *ACM Computing Surveys*, vol. 44 (4), pp. 20:1–20:36.

Yagunova, E., Lande, D. (2012) Dynamic Frequency Features as the Basis for the Structural Description of Diverse Linguistic Objects. *CEUR Workshop Proceedings*, vol. 934, Russian Conference on Digital Libraries, pp. 150–159.

Yagunova, E.V., Krilova, I.V, Makarova, O.E., Pivovarova, L.M. (2014) Snezhnaya revoluciya v Rossii: znachimie nominacii, sobitiya, ocenki (oceenka sobitiy informantami i dannie SMI). In: *“Mi ne nemi!”: tvorchesvo protestuyshey ulici*, Tartu. Title in Russian: “Снежная революция в России”: значимые номинации, события, оценки (оценка событий информантами и данные СМИ)

Lande D., Snarskii A., Jagunova E. Network of Natural Hierarchies of Terms of News Messages on the “Euromaydan” Events. *CEUR Workshop Proceedings*, vol. 1297, Russian Conference on Digital Libraries, pp. 55-74 (2014)