Computer Science Journal of Moldova, vol.**24**, no.2(71), 2016

Estimation of Morphological Tables Using Text Analysis Results*

Illia Savchenko

Abstract

This paper proposes methods for obtaining input data, necessary for the modified morphological analysis method, from the text sources of data using text analysis tools. Several methods are described that are suitable for calculating initial estimates of alternatives and cross-consistency matrix values based on processing text fragments by rule-based categorization and sentiment analysis tools. A practical implementation of this tool set for assessing statements in news regarding Ukraine is considered.

Keywords: Morphological analysis method, text analysis, morphological table, foresight.

1 Introduction

The modified morphological analysis method (MMAM) is a powerful quality analysis tool for problems, where the objects are characterized by uncertainty, incompleteness, inaccuracy, fuzziness of information and thus have a large multitude of possible alternative configurations [1]. These objects can be described and studied by MMAM, allowing to make decisions regarding such objects.

Work with objects in MMAM [2], [3] requires an initial estimation of alternatives and the cross-consistency matrix values, which is usually done by expert evaluation. This approach is the most exact, however, it requires a lot of time and financial resources. Besides, the number

*This research was sponsored by the NATO Science for Peace and Security Programme under grant 984877

of questions for experts even for relatively small morphological tables may amount to hundreds or thousands, which is inadmissible.

As the object descriptions in the foresight process are commonly verbal, it is proposed to involve text analysis tools for obtaining input values. Such analysis tool may include rule-based categorization and sentiment analysis tools [4], [5].

The MMAM in foresight problems is the most efficient one for two types of tasks [2]:

1. The description of objects (processes, phenomena), that emerge multiple times in different configurations, to study and make decisions against the possible multitude of such objects as a whole.

2. The description of a state of a given complex system, that has uncertain characteristics, e.g. a future state of this system.

The application of text analysis tools for estimating morphological tables is possible in both cases. In the first case the input data can be acquired from news, messages, claims regarding the objects of the studied type. In the second case information comes in the form of predictions, statements, comments of experts regarding the given system.

2 Statement of the problem

We will assume that we already have a collection of texts regarding the object, that is a target for morphological study.

The MMAM needs two types of input estimates: the initial probabilities of the alternatives, and the values of cross-consistency matrix for pairs of alternatives. To have the capacity to obtain this data from the unstructured information fragments, first we have to introduce the rules for text analysis [5], which allow relating the text with a certain degree of confidence to categories that correspond to the alternatives of the morphological table parameters.

Let's designate $R_j^{(i)}(g_k)$ as the rule for calculating the degree of relation for the text fragment g_k to a category that corresponds to the alternative $a_j^{(i)}$ of the morphological table. This set of rules is constructed for each alternative $a_j^{(i)}$ of each parameter F_i of the morphological table. These rules determine that the text fragment mentions an object of study with the characteristic, specified by parameter F_i, has a value that corresponds to alternative $a_j^{(i)}$.

Then, the following statement of problem can be written:
Given:

- a morphological table with N parameters $F_i, i \in \overline{1,N}$, each having a set of alternatives $a_j^{(i)}, j \in \overline{1,n_i}$;
- a collection of N_{text} text fragments $g_k, k \in \overline{1,N_{text}}$. It is considered to be already determined that each of the text fragments mentions an object of study;
- $R_j^{(i)}(g_k), i \in \overline{1,N}, j \in \overline{1,n_i}, k \in \overline{1,N_{text}}$ – the text analysis rules for calculating the degree of relation of the text fragment g_k to the category, that corresponds to the alternative $a_j^{(i)}$ of the morphological table.

Required:

- to calculate the initial estimates $p`_j^{(i)}$ for each alternative $a_j^{(i)}$;
- to calculate the cross-consistency matrix values $c_{i_1 j_1 i_2 j_2}$ for each pair of alternatives $a_{j_1}^{(i_1)}, a_{j_2}^{(i_2)}$.

3 Evaluating initial alternative values

Analyzing a large enough collection of texts, we can make conclusions regarding the distribution of probabilities between the alternatives for any morphological table parameter. But it is necessary to mention that text analysis tools give only a hypothesis regarding the relation of a text fragment to certain alternative $a_j^{(i)}$, with a confidence of $R_j^{(i)}(g_k)$. Thus a situation is possible, when a single text fragment is related to several alternatives simultaneously with different degrees of confidence.

Taking this into account, we proposed two methods of evaluating initial alternative values:
Method 1: taking a ratio of total degree of relation to category that corresponds to $a_j^{(i)}$, to the total degree of relation for all categories that correspond to alternatives of parameter F_i, for all text fragments:

$$p`_j^{(i)} = \frac{\sum_{k=1}^{N_{text}} R_j^{(i)}(g_k)}{\sum_{k=1}^{N_{text}} \sum_{j^*=1}^{n_i} R_{j^*}^{(i)}(g_k)}.$$

Method 2: taking a ratio of a number of text fragments, where the degree of relation to $a_j^{(i)}$ is maximal, to the total number of text fragments, where the degree of relation to at least one alternative of the parameter F_i is larger than zero.

$$p`_j^{(i)} = \frac{\left|\left\{g_k | R_j^{(i)}(g_k) = \max_{j^* \in \overline{1,n_i}} (R_{j^*}^{(i)}(g_k)), R_j^{(i)}(g_k) > 0\right\}\right|}{\left|g_k | \exists j^* \in \overline{1,n_i} : R_{j^*}^{(i)}(g_k) > 0\right|},$$

where $|A|$ is the power of set A, i.e. the number of elements in it.

The choice of method depends on the specifics of the problem. If the alternatives and their corresponding text analysis rules are defined in such a way that most text fragments relate to a single alternative, then the first method is more reliable. If most text fragments relate to multiple alternatives with positive degrees, then the second method may be more correct.

4 Evaluating cross-consistency matrix values

The cross-consistency matrix establishes the dependencies between the alternatives of different parameters, i.e. the type of influence of choosing one alternative in a pair on the probability of choosing the other one. For practical purposes this value can be regarded as the correlation between the choice of alternatives in a pair for the object configuration. Thus the first method of evaluating cross-consistency matrix values is formed.

Method 1. Using correlation between the degrees of confidence of relating the text fragments to alternatives from a pair, that is connected by the corresponding cross-consistency matrix value.

$$c_{i_1 j_1 i_2 j_2} = r(R_{j_1}^{(i_1)}, R_{j_2}^{(i_2)}) =$$

$$= \frac{\sum_{k=1}^{N_{text}} \big(R_{j_1}^{(i_1)}(g_k) - \overline{R_{j_1}^{(i_1)}}\big)\big(R_{j_2}^{(i_2)}(g_k) - \overline{R_{j_2}^{(i_2)}}\big)}{\sqrt{\sum_{k=1}^{N_{text}} \big(R_{j_1}^{(i_1)}(g_k) - \overline{R_{j_1}^{(i_1)}}\big)^2 \sum_{k=1}^{N_{text}} \big(R_{j_2}^{(i_2)}(g_k) - \overline{R_{j_2}^{(i_2)}}\big)^2}},$$

where $\overline{R_{j_1}^{(i_1)}} = \big(\sum_{k=1}^{N_{text}} R_{j_1}^{(i_1)}(g_k)\big)/N_{text}$ is the mean degree $R_{j_1}^{(i_1)}(g_k)$ for all text fragments. In this method only the text fragments with non-zero degrees of confidence for at least one alternative of each parameter F_{i_1}, F_{i_2} are considered:

$$g_k \in \big\{g_k | \exists j_1^* : R_{j_1^*}^{(i_1)}(g_k) > 0 \wedge \exists j_2^* : R_{j_2^*}^{(i_2)}(g_k) > 0\big\}.$$

The second method is based on the fact that the pair of alternatives, that have higher values in the cross-consistency matrix, also has the higher probability of mention in a text fragment.

Method 2. Calculating a ratio between the number of text fragments, where both of the alternatives are mentioned and the number of text fragments, where at least one alternative of the pair is mentioned:

$$c_{i_1 j_1 i_2 j_2} = 1 - 2\frac{\big|g_k | R_{j_1}^{(i_1)}(g_k) > 0 \wedge R_{j_2}^{(i_2)}(g_k) > 0\big|}{\big|g_k | R_{j_1}^{(i_1)}(g_k) > 0 \vee R_{j_2}^{(i_2)}(g_k) > 0\big|}.$$

For evaluating cross-consistency matrix the second method is more reliable in cases where most text fragments have a well-defined relation to a specific alternative. On the contrary, if the text fragments have non-zero degrees of relation to several alternatives of one parameter, then the first method is preferable.

5 Practical application

To test the developed tool set, it was employed to estimate the alternatives of a morphological table for international statements regarding Ukraine in 2013. The morphological table is presented in Table 1.

Table 1. Morphological table for statements regarding Ukraine

Statements regarding Ukraine		
1. Speaker	2. Subject	3. Tone
EU	Elections	Positive
NATO	Eurointegration	Neutral
Russia	Economics	Negative
USA		

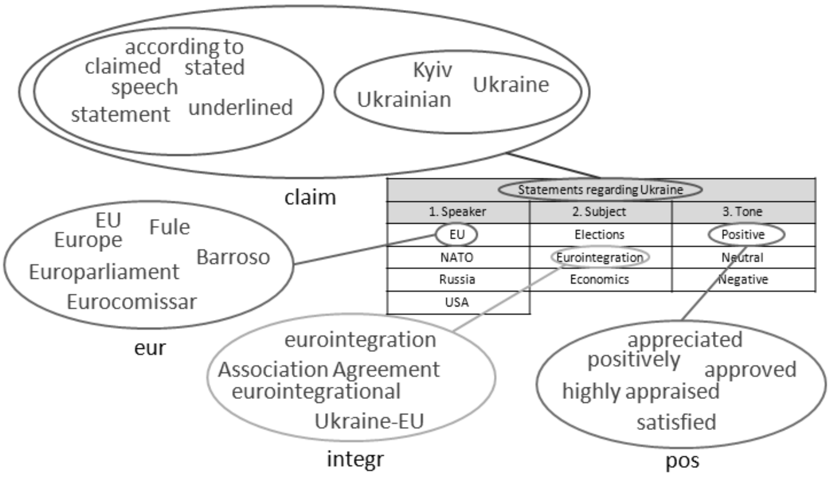
To conduct evaluation, the SAS Content Categorization Studio software was applied, using a collection of 3805 text fragments (news in 2013). Elements of the building blocks for SAS text analysis rules are presented in Figure 1.

Using the proposed tool set, the initial values for the alternatives of the morphological table were obtained (Table 2):

Table 2. The initial values for the alternatives, obtained by the analysis of text fragments

Statements regarding Ukraine		
1. Speaker	2. Subject	3. Tone
0.569	0.216	0.051
0.064	0.386	0.903
0.307	0.398	0.046
0.06		

Similarly the cross-consistency matrix was evaluated (Table 3).



Statements regarding Ukraine		
1. Speaker	2. Subject	3. Tone
EU	Elections	Positive
NATO	Eurointegration	Neutral
Russia	Economics	Negative
USA		

Figure 1. Samples of elements for text analysis rules

Table 3. The cross-consistency matrix values, obtained by the analysis of text fragments

		1. Speaker				2. Subject		
		$a_1^{(1)}$	$a_2^{(1)}$	$a_3^{(1)}$	$a_4^{(1)}$	$a_1^{(2)}$	$a_2^{(2)}$	$a_3^{(2)}$
2.Subject	$a_1^{(2)}$	–0.29	–0.27	–0.43	–0.13			
	$a_2^{(2)}$	0.25	–0.41	–0.39	–0.06			
	$a_3^{(2)}$	–0.32	–0.44	0.35	0.31			
3. Tone	$a_1^{(3)}$	–0.52	–0.71	–0.82	–0.78	–0.56	–0.16	–0.35
	$a_2^{(3)}$	0.59	0.51	0.76	0.38	0.87	0.87	0.9
	$a_3^{(3)}$	–0.52	–0.8	–0.88	–0.6	–0.51	–0.18	–0.39

6 Conclusions

The proposed above methods allow one to process large morphological tables without creating excess strain for experts. They also allow using in estimation the unstructured data, which is hard to account otherwise, to gather the statistics for the entities, where it would be inaccessible by other means.

The limitation of this method is the necessity of having a large enough volume of text information in the field of study. The input data must be processed as text fragments, each being a separate description of the object of study. The fragments also shouldn't duplicate the same description of the object, i.e. be independent of each other. A variety of thoughts and sources is encouraged to exclude one-sidedness and prejudice in estimates. Also the utilization of these methods require skills for creating and tuning text analysis tools.

Thus, the result of evaluation is highly dependent on the quality of text analysis rules and the collection of texts itself. These methods are rarely suitable as the only source of input data, however, they can be deemed as a starting point for further improvement, or as a thought of a single expert, when used in parallel with classic expert estimation.

References

[1] N. D. Pankratova, P. I. Bidyuk, Y. M. Selin, I. O. Savchenko, L. Y. Malafeeva, M. P. Makukha, V. V. Savastiyanov, "Foresight and Forecast for Prevention, Mitigation and Recovering after Social, Technical and Environmental Disasters," in *Improving Disaster Resilience and Mitigation – IT Means and Tools* (Part I), H.-N. Teodorescu, A. Kirschenbaum, S. Cojocaru and C. Bruderlein, Eds. Netherlands: Springer, 2014, pp. 119–134. Available: DOI: 10.1007/978-94-017-9136-6_8.

[2] N.D. Pankratova, I.O. Savchenko, *Morphological Analysis. Problems, Theory, Applications,* Kyiv, Ukraine: Naukova Dumka, 2015, 245 p. (in Ukranian)

[3] I.O. Savchenko, *Methodological and Mathematical Support for Solving Foresight Problems Using Modified Morphological Analysis Method.* Innovative Development of Socio-Economic Systems Based on Foresight and Cognitive Modeling Methodologies, Kyiv, Naukova Dumka, 2015, pp. 427–441.

I. O. Savchenko, “Methodological and Mathematical Support for Solving Foresight Problems Using Modified Morphological Analysis Method,” in *Innovative Development of Socio-Economic Systems Based on Foresight and Cognitive Modeling Methodologies*, G. V. Gorelova and N. D. Pankratova, Eds. Kyiv, Ukraine: Naukova Dumka, 2015, pp. 427–441. (In Russian)

[4] B. Liu, *Sentiment Analysis and Opinion Mining,* Morgan & Claypool Publishers, 2012, 180 p. ISBN-13: 978-1608458844.

[5] H. Reckman, Ch. Baird, J. Crawford, R. Crowell, L. Micciulla, S. Sethi, F. Veress, “Rule-based detection of sentiment phrases using SAS Sentiment Analysis,” in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics,* Atlanta, Georgia, 2013, pp. 513–519.

Illia Savchenko

Received February 20, 2016

Educational-Scientific Complex ”Institute for Applied System Analyses”,
National Technical University of Ukraine ”Kyiv Polytechnic Institute”
prosp. Peremohy, 37, bd. 35, Kyiv, Ukraine.
Phone: +38 050 3871688
E–mail: savil@inbox.ru