# EVALUATING EFFECTIVENESS OF LINGUISTIC TECHNOLOGIES OF KNOWLEDGE IDENTIFICATION IN TEXT COLLECTIONS

**Nina Khairova, Gennady Shepelyov, Svetlana Petrasova**

***Abstract**: The possibility of using integral coefficients of recall and precision to evaluate effectiveness of linguistic technologies of knowledge identification in texts is analyzed in the paper. An approach is based on the method of test collections, which is used for experimental validation of received effectiveness coefficients, and on methods of mathematical statistics. The problem of maximizing the reliability of sample results in their propagation on the general population of the tested text collection is studied. The method for determining the confidence interval for the attribute proportion, which is based on Wilson's formula, and the method for determining the required size of the relevant sample under specified relative error and confidence probability, are considered.*

***Keywords**: recall, precision, relevance, confidence interval, sample size.*

***ACM Classification Keywords:** I.2.7. Natural Language Processing, G.3. Probability and Statistics –Statistical Computing.*

## Introduction

Nowadays linguistic technologies have become not only tools for modelling language but also the production factor. Computer linguistics is getting now the most strongly developing direction of information technologies. In fact, every intelligent information system with a user interface, both text and web-content processing systems, uses linguistic technologies.

The effectiveness of such technologies depends on morphological, contextual, and syntactic analysis and synthesis as well as on solving the semantic analysis problems. The number of linguistic and information approaches to solve this problem is constantly growing. Therefore common metrics should be introduced for evaluation of the effectiveness of such technologies and their comparison. But currently, there are no standard benchmarks to measure effectiveness using mentioned technologies in text collections. We propose here some indicators and test their reliability.

Usually the method of test collections is used for estimating the effectiveness of linguistic technologies in different systems of text classification, information retrieval, text mining, opinion mining, web mining etc. [Cormack, 1998]. The essence of this method consists in comparing the results of the tested technology at predetermined texts with expert evaluation for the same texts.

However comparing results of the method with experts' opinions generates the two main problems:

- expert subjectivity;
- the need for the determination of the text collection size to make experimental results reliable.

Notion the reliability means here that experimental results, which were received, will be true under certain conditions also in the framework of a certain wider class of objects.

## Integral Effectiveness Coefficients of Knowledge Identification

Let's use the quantitative effectiveness coefficients of retrieval and classification approved by interstate standards for information, library science, and publishing [ISO 12620:2009]. These coefficients are precision, recall. All these coefficients are based on the subjectively determined concept of relevance. The concept of relevance is difficult to define and has a rather psychological nature. We use the definition of relevance [Mizzaro, 1997], in which relevance depends on four concepts of Relevance (IR, IN, C, T) Here IR is an information resource, which is presented by a set of collection texts for processing, IN are information needs, C is context and T is time.

Relevance is defined by experts on the scale of relevant/irrelevant/undefined and shows the correspondence or discrepancy of a text to a certain knowledge domain.

To calculate the coefficients of system recall and system precision for each domain of expert's knowledge, it is necessary to determine the following parameters:

- $n_{yy}$ - a number of elements identified by the system as relevant, which are relevant to the local domain knowledge from an expert's viewpoint too,
- $n_{yn}$ - a number of elements identified by the system as relevant, which are irrelevant to the local domain knowledge from an expert's viewpoint,
- $n_{ny}$ - a number of elements that the system has not identified as relevant, which are relevant to the local domain knowledge from an expert's viewpoint.

Using these parameters coefficients of system precision and system recall are determined by the following formulas:

$$\text{precision} = \frac{n_{yy}}{n_{yy} + n_{yn}}, \tag{1}$$

$$\text{recall} = \frac{n_{yy}}{n_{yy} + n_{ny}}. \tag{2}$$

## Sampling Method for Text Collections

As the determination of effectiveness indicators is based on the notion of relevance that expert defines subjectively, the reliability of recall, precision, and other effectiveness indicators of the linguistic technologies requires experimental verification on text collection.

Since used text collections are huge it makes sense to study only a part of the objects from the experimental collection, that is to execute the so-called sampling research of the population and make valid conclusions about the properties of the whole population.

As the general population of any model of natural language texts processing is tending to infinite size, the ratio of the sample size to the general population size is much less than 5 - 10%, therefore the mathematical apparatus of the sampling with replacement theory can be used how it was shown by Chetyrkin [Chetyrkin, 1982]. Furthermore, in some cases, using the necessary correction coefficient, the results for a sample with replacement can be transferred to the corresponding results for a sample without replacement.

Within our issue let's evaluate an attribute proportion in the general population on a basis of the corresponding attribute proportion in the sample. Let's consider the share of relevant texts in the collection *R* as the attribute proportion that shows the ratio of the number of relevant texts to the total number of collection texts. The sample estimate $R_S$ of the proportion *R* is $R_S = M / N$, where *N* is the size of an experimental return sample, and *M* is the number of identified relevant texts in a sample using the identification method. It can be shown that the evaluation satisfies all of the requirements to statistical estimates (consistency, unbiasedness, sufficiency and effectiveness) [Chetyrkin, 1982].

Since the sample estimate $R_S$ is a point estimate of the attribute proportion, the interval estimation $R_S$ should be used in order to find the sampling error. Since sampling errors are random variables with the same probability distribution, we can define interval estimate within which the attribute proportion of the population will be found with a certain confidence probability P.

Usually, this approach leads to three issue types:

- determination of the confidence probability for a given confidence interval and sample size;
- determination of the confidence interval for a given confidence probability and sample size;
- determination of the necessary sample size for a given confidence probability and error limit.

The determination of the confidence interval and the necessary size of the sample are the most important in our problem.

## Determination of the Confidence Interval for a Given Confidence Probability and Sample Size

The determination of the confidence interval of the attribute proportion is based on the binomial distribution law [Clopper, 1934]. However, starting from samples that are more than 20 in size, the binomial distribution is symmetrized and is well approximated by normal distribution with parameters: average $<R_S> = R$, variance $D(R_S) = R(1 - R)/N$, standard deviation $\sigma(R_S) = [D(R_S)]^{1/2}$. In this case the confidence interval can be calculated using the formula:

$$P(|R - R_S| < E_\alpha) = 2\Phi(Z_\alpha) = 1 - \alpha, \tag{7}$$

where $\Phi(Z_\alpha)$ is the Laplace function. The margin error of the sample is found from the equation:

$$E_\alpha = Z_\alpha \sigma(R_S). \tag{8}$$

Let's choose the value of 0.95 usually used as the value of confidence probability, then the significance level α is 0.05. At that $Z_{0.05} = 1.96$. Then we can get expressions for right and left limits of the confidence interval $R$ from the relation:

$$|R - R_S| < Z_\alpha[R(1 - R)/N]^{1/2}. \tag{9}$$

To do so we should solve the corresponding quadratic equation for R [Wilson, 1927]. The adequacy of using this approach to estimate the confidence intervals of the attribute proportion for small samples was proved by L.D. Brown and M. A. García-Pérez [Brown, 2001; Garcia-Perez, 2005]. Using the results of the paper [Agresti, 1998], we can get values of confidence limits in simpler way:

$$Z_\alpha = |R - R_S|/[R(1 - R)/N]^{1/2}, \tag{10}$$

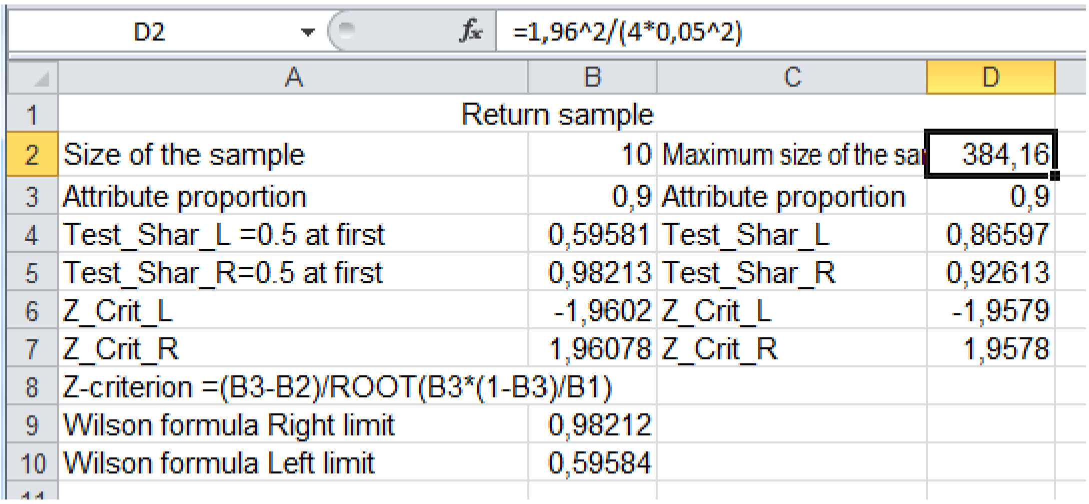
## Determination of the Necessary Sample Size

To determine the size of the necessary sample for given confidence probability and the margin of error, we replace $|R - R_S|$ in (10) with E and determine N. We can see that:

$$N = [Z^2 R_S(1 - R_S)]/E^2. \tag{11}$$

The ratio (11) for size of the sample includes yet unknown sample proportion $R_S$. Since this proportion is unknown, it is reasonable to determine it so that the size of the sample $N$ is maximal. Then it will be acceptable for all feasible $R_S$. It is easy to see that the maximum $N$ as the function from $R_S$ is reached at $R_S = ½$, that is $N_{MAX} = Z^2/4E^2$. Certainly, if there are a priori assumptions about the value of the attribute proportion during the research (by analogy, from the experience), this value should be used in the ratio (11). The necessary size of the sample is less than maximum one.

Quite often the value of the margin of error $E = 0.05$ is used for determining the proportions of attributes. Using MS-Excel, let's consider the following illustrative example given in Figure 1.



D2 =1,96^2/(4*0,05^2)

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Return sample | | | |
| 2 | Size of the sample | 10 | Maximum size of the sa | 384,16 |
| 3 | Attribute proportion | 0,9 | Attribute proportion | 0,9 |
| 4 | Test_Shar_L =0.5 at first | 0,59581 | Test_Shar_L | 0,86597 |
| 5 | Test_Shar_R=0.5 at first | 0,98213 | Test_Shar_R | 0,92613 |
| 6 | Z_Crit_L | -1,9602 | Z_Crit_L | -1,9579 |
| 7 | Z_Crit_R | 1,96078 | Z_Crit_R | 1,9578 |
| 8 | Z-criterion =(B3-B2)/ROOT(B3*(1-B3)/B1) | | | |
| 9 | Wilson formula Right limit | 0,98212 | | |
| 10 | Wilson formula Left limit | 0,59584 | | |

*Figure 1. Example of the Evaluation of the Necessary Size of Sample*

Let us have a sample with replacement of size $N = 10$ objects and the attribute proportion of $R_S = 0.9$. Using *Goal Seek* procedure, we get values of right and left confidence limits for the attribute proportion in the population $R$ for 0.95 confidence probability: $0.59 < R < 0.98$. The obtained confidence interval is too wide. Let's find the maximum size of the sample $N_{MAX}$ for the same confidence probability 0.95 (then $Z = 1.96$) and the margin error $E = 0.05$ (recall that $N_{MAX} = Z^2/4E^2$). Rounding the computed necessary size of the sample up to an integer, we have: $N_{MAX} = 385$. Using *Goal Seek* once again, we get a new narrower confidence interval: $0.87 < R < 0.93$. It has been achieved at the cost of considerable increase of the necessary size of the sample.

## Conclusion

Thus this paper substantiates the usage of recall and precision to evaluate effectiveness of knowledge identification by means of linguistic technologies in text collections. The methods of mathematical statistics are used for determining the evaluation error of chosen coefficients. We considered the problem of the evaluation of the results obtained from the sample and evaluated an attribute proportion in the general population on a basis of the corresponding attribute proportion in the sample. The attribute proportion is considered as a share of relevant texts in the collection. It shows the ratio of the number of relevant texts to the total number of collection texts. The confidence interval for the attribute proportion was computed and the necessary size of the relevant sample was determined in given confidence probability. The experimentally determined values of recall and precision coefficients for the sample size correspond to the values of the same coefficients for the complete text collection for given confidence probability equal to 0.95 and the error limit equal to 0.05.

## Bibliography

[Agresti, 1998] Agresti A., B. Coull A. Approximate is better than exact for interval estimation of binomial proportions // American statistician. – 1998. – N 52. – C. 119–126.

[Brown, 2001]. L. D. Brown, T. T. Cai, A. Dasgupta. D. Interval estimation for a binomial proportion // Statistical science. – 2001. – N 2. – P. 101–133.

[Chetyrkin, 1982] Chetyrkin Ye. M., Kalichman I. L. Probability and Statistics. M.: Finance and Statistics, 1982 (in Russian).

[Clopper, 1934] Clopper C. J., E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial // Biometrika. – 1934. – N 26. – P. 404–413.

[Cormack, 1998] Cormack G.V. A Efficient construction of large test collections / G. V. Cormack , C. R. Palmer, C. L. Clarke // Proc. of the SIGIR'98 — P. 282—289.

[Garcia-Perez, 2005] Garcia-Perez M. A. On the confidence interval for the binomial parameter // Quality and quantity. – 2005. – N 39. – P. 467–481.

[ISO 12620:2009] "ISO 12620:2009 - Terminology and other language and content resources -- Specification of data categories and management of a Data Category Registry for language resources". Retrieved 9 November 2011.

[Mizzaro, 1997] Mizzaro S. Relevance: The whole history. Journal of American Society for Information Science. — 1997. — V.48. — № 9 — P. 810-832.

[Wilson, 1927] Wilson E. B. Probable inference, the law of succession, and statistical inference //Journal of American Statistical Association. – 1927. – N 22. – P. 209–212.

## Author's Information

***Nina Khairova*** *– Doctor of Computer Linguistics, Professor of Intelligent Computer Systems Department of National Technical University “Kharkiv Polytechnic Institute”,*

*21, Frunze str., Kharkiv, Ukraine, 61002*

*E-mail: nina_khajrova@yahoo.com*

*Major Fields of Scientific Research: artificial intelligence, knowledge identification in texts, text mining, opinion mining, web mining, natural language processing.*

***Gennady Shepelyov*** *– Head of Laboratory "Computer systems based on knowledge" of Institute for systems studies of RAS,*

*Prospect 60-letiya Oktyabrya, 9 Moscow 117312 Russia*

*e-mail: gis@isa.ru*

*Major Fields of Scientific Research: mathematical modelling, probabilistic methods, interval analysis, generalized interval estimations, comparing interval alternatives.*

***Svetlana Petrasova*** *– Postgraduate of Intelligent Computer Systems Department of National Technical University “Kharkiv Polytechnic Institute”,*

*21, Frunze str., Kharkiv, Ukraine, 61002*

*E-mail: svetapetrasova@gmail.com*

*Major Fields of Scientific Research: artificial intelligence, knowledge engineering, intelligent systems of knowledge representation, computer linguistics, natural language processing.*