ECONTECHMOD. AN INTERNATIONAL QUARTERLY JOURNAL – 2013. Vol. 2. No. 4. 47–51

Definition of the semantic metrics on the basis of thesaurus of subject area

V. Lytvyn, O. Semotuyk, O. Moroz

Department of Information Systems and Networks, Department of Appliued Lingustics
Lviv Polytechnic National University, 79012, Lviv, Bandery st. 12, Ukraine
e-mail: vasyll@ukr.net, orest.semotiuk@gmail.com, olha.moroz.81@gmail.com

Received June 21. 2013 : accepted Oktober 10.2013

Abstract. The paper proposes an approach to construction of semantic metrics based on thesaurus of the domain of linguistics. The process of constructing a thesaurus is described. A way is proposed to use the built knowledge base to find potential partners who are engaged in similar research issues in the subject area for which thesaurus was constructed.

Key words: thesaurus, knowledge base, semantic metrics, relation, weight ratio.

INTRODUCTION

The language of science is structured scientific knowledge, sets a hierarchical multilayer formation, which allocated blocks: terminological, nomenclature, methods and rules for forming apparatus and conceptual terms.

Encyclopedias, dictionaries and terminology on which terminological system of the subject area is based tend to have a clear structure and consist of entries. It is therefore necessary to investigate their possible arrangements to recognize concepts and relations between them to build a thesaurus software.

In [1-3] the construction of a thesaurus is described in detail. This paper proposes to use a thesaurus of linguistic terms developed by the authors to find potential partners who are engaged in similar research problems in a given software. To solve this problem it is necessary to build a semantic metric.

METHODS FOR DETERMINING SEMANTIC METRICS

There are several ways to determine the semantic metrics.

Table 1 shows how to calculate the degree of similarity of text documents (TD) based on:

- word frequency in text documents,
- distance in the taxonomy of concepts,
- word frequency and distance in the taxonomy of concepts simultaneously.

Google Distance - a degree of semantic coherence, which is calculated based on the number of pages obtained by pursuing Google for a given set of keywords. The table shows the formula for calculating the normalized Google distance (NGD) for two terms: *x* i *y*, where *M* is the total number of web-pages indexed by Google; $f(x)$ i $f(y)$ – number of pages containing keywords *x* i *y*, respectively $f(x, y)$ – number of pages containing both *x*, and *y*. If *x* and *y* are found on all pages together, then we consider *NGD*=0, if they occur only separately , then we consider *NGD*=∞.

We select a class of metrics that compute similarity based on taxonomy data. These metrics are used to compute the similarity of concepts WordNet [6], GermaNet, Wikipedia [4].

In [13] a formula is proposed that takes into account both the depth in the hierarchy of concepts, and the depth of the *lcs* (least common subsumer):

$$wup(C_1, C_2) = \frac{lcs(C_1, C_2)}{depth(C_1) + depth(C_2)}.$$

Ryeznyk [8] proposed to consider that two words are the more similarly the more informative concept is, which relate to these two word, this means the lower in the taxonomy is a common top concept (synset in

Table 1. Semantic metrics classification

Formula/ description of the algorithm	Title
1. Word frequency in text document	
$NGD(x,y)=\frac{max(\log f(x),\log f(y))-\log f(x,y)}{\log M-min(\log f(x),\log f(y))}$	Normalized distance Google (*NGD*)
$jaccard(x,y)=\frac{Hits(x\wedge y)}{Hits(x)+Hits(y)-Hits(x\wedge y)}$	Jaccard [4]
2. Distances in the taxonomy of terms	
Distance corresponds to the number of edges shortest path between concepts	Metrics was used for the concepts of Roget's thesaurus [5]
$lch(C_1,C_2)=-\log\frac{length(C_1,C_2)}{2D}$	Leacock & Chodorov 1997, [6] pp. 265-283
$wup(C_1,C_2)=\frac{lcs(C_1,C_2)}{depth(C_1)+depth(C_2)}$	Wu & Palmer [7]
$res_{hypo}(C_1,C_2)=1-\frac{\log(hypo(lcs(C_1,C_2))+1)}{\log(C)}$	Metrics *res* [8], adapted to the taxonomy of the Wikipedia categories
3. Frequency words and distances in the taxonomy	
$res(C_1,C_2)=\max_{C\in S(C_1,C_2)}[-\log(P(C))]$	Distance *res* [9]
$lin(C_1,C_2)=\frac{2\cdot\log(P(C_0))}{\log(P(C_1))+\log(P(C_2))}$	Distance *lin* [10]
4. Text intersection	
Text intersection (based on WordNet)	Lesk [11]
extended gloss overlap – text crossing considering the neighboring concepts WordNet	Banerjee & Pedersen, 2003 [12]
$relate_{gloss/text}(T_1,T_2)=tanh\frac{overlap(T_1,T_2)}{length(T_1)+length(T_2)}$	Відстань *relate* [4]

wordNet). In constructing probabilistic functions $P(C)$, it is considered that the concept probability should not be changed while moving up the hierarchy: $res(C_1,C_2)=\max_{C\in S(C_1,C_2)}[-\log(P(C))]$. Then abstract concepts are less informative. Ryeznyk proposed to estimate the probability over frequency synonyms concept in a text document (TD) so: $P(C)=\frac{freq(C)}{N}$, $freq(C)=\sum_{n\in words(C)} count(n)$, where $words(C)$ –are nouns with the value *C*; *N* – total number of nouns in text document.

In the paper [9] Ryeznik's metric has been adapted to Wikipedia and informative category was calculated as a function of the hyponyms number (categories in Wikipedia), but not statistically:

$$res_{hypo}(C_1,C_2)=1-\frac{\log(hypo(lcs(C_1,C_2))+1)}{\log(C)},$$

where: *lcs* is the least common subsumer of concepts C_1 і C_2, *hypo* – number of Hyponyms of this subsummer, and *C* – total number of concepts in the hierarchy.

In [10] Lin determines the similarity of objects A and B as the ratio of the amount of information required to describe the similarity of A and B, to the amount of information that fully describes A and B. To measure the similarity between words *lin* takes into account the frequency distribution of words in the text (similar to the measure Reznik):

$$lin(C_1,C_2)=\frac{2\cdot\log(P(C_0))}{\log(P(C_1))+\log(P(C_2))},$$

where: C_0 – nearest common super class in the concept hierarchy for both conceps C_1 і C_2, *P* –probability of concept, calculated on the basisf of his frequency in the text document. It differs from the formula *res* by normalization method, correct computation $lin(x,x)$ (independent of the concept's position in the hierarchy), takes into account existence of common and distinctive properties in objects.

In the paper [4] similarity of the two texts T_1 і T_2 is calculated from the double normalization (the length of the text and using hyperbolic tangent) as:

$$relate_{gloss/text}(T_1,T_2)=\tanh\frac{overlap(T_1,T_2)}{length(T_1)+length(T_2)},$$

$$overlap(T_1,T_2)=\sum_n m^2,$$

where *n* phrases та *m* words overlap.

Thus the analysis showed that no semantic metric is not based on thesauri, only a few of them take into account the taxonomy of concepts.

To say clearly, is introduced the metric on the feature space. In this space is defined the point corresponding to the current problem, and in the frames of this metric is detecting the nearest point to it among the points, which represent the precedents. To each attribute is prescribed weight, considering its relative value. Completely the degree of proximity precedent by all parameters can be calculated by using of generalized formula, which looks like:

$$\sum_k w_k \cdot sim\left(x_{ki}, x_{kj}\right), \ \sum_k w_k = 1,$$

where: w_k – weight of k-feature , sim – function of similarity (metric), x_{ki} and x_{kj} – meaning of the feature x_k for the current problem i of the precedent – j. After the calculating the degrees of proximity, all precedents are ranking. The current situation is referring to the precedent with the highest rank.

Selecting a metric (or degree of proximity) is the central point from which will greatly depend on searching for the relevant precedents. In every particular problem this choice is in its own way, with including the main goals of the research, physical and statistical basis of information etc. As methods for solving such a problems use algorithms such as Lazy-Learning, for example – known algorithms of the nearest neighbor and of the nearest k-neighbors, neural networks, genetic algorithms, Bayesian networks, decision trees.

The main disadvantage of the paradigm of the neural network is the necessity to have a very big amount of training samples. Another significant disadvantage is that the scale of several hundred interneural connections, are not a subject of analysis and interpretation by a human.

The popularity of the decision trees is associated with clearness and clarity. But for them very actual is the problem of importance. The fact is that some nodes on every new-built tree level correspond to less and less number of data records – tree fractions data for a large number of individual cases, so it does not give statistically valid answers. How the practice shows, in the most of systems, which are using decisions trees, this problem can't find satisfactory solution. By the way, well-known, and it's easy to show, that the decision trees give useful results only in case of independent features. Otherwise they only create the illusion of the logical derivation (output).

Genetic algorithms also have several disadvantages. Selection criterion of chromosomes and used procedures are heuristic and don't guarantee to find "better" solution. Besides, efficiently formulate objectives, identify criteria for selection of chromosomes in strength only to the specialist. Because of these factors today genetic algorithms are in need to be treated more like a research tool than as a means of analyzing data for practical application. In our opinion, to get rid of the above disadvantages allow the ontology of the subject area and the ontology of the problems.

APPROACH TO THE CONSTRUCTION OF THE THESAURUS OF SUBJECT AREA

Thesaurus is a list of logical- semantic relations between linguistic terms. This thesaurus embraces not only set of the terms provided in the form of an alphabetical list of their definitions , but also contains the models which represent relationships between terms. Based on the achievements of modern linguistics in a compact and accessible form given interpretation of terminological units from terminological dictionaries and encyclopaedias. The thesaurus contains terms in main research areas of theoretical and applied linguistics: grammar, word formation, lexicology, semantics, lingvosemiotisc, computational linguistics, lexicoghraphy etc. We selected these terms from the abstracts of papers, published in the Ukrainian linguistic periodicals in the 2009-2011.

Building a thesaurus provides for the disclosure of the main types of relations between concepts, the main ones are correlation, synonymy, hiponymy/hyperonymy, holonymy/meronymy. Contents relations expanded so that you can reach the widest layer of terms , which linked the analyzed period as the registry .

Title ratio is double predicate R (A, B), which binds headword article (A) and put this predicate term (B) [14].

APPROACH TO CONSTRUCTION OF SEMANTIC METRICS ON THE BASIS OF THE THESAURUS

For the definition of the importance of the weight of concepts and relations, we are proposing to use the methods of the intellectual data analysis (IDA), such as decisions trees. Using IDA, we define the weight of some subset of concepts, which we are calling – basic. Then based on the ontology of the SA, we will develop the received weights for the whole ontology. This procedure we will make for every precedent. Then for searching the relevant precedent we will use the value of such N_i concepts, which for proper precedent have the biggest weight. As for the importance of the weight of the relations, we are offering to make them like it is shown on the table 2.

We consider, that the weight of the vertical relations (hierarchy, aggregation) is equal to 1, 2 (the more specific, the better). Relations by quantum are not examined, because the synonymy and the harmonization don't make any influence on the value of the attributes. At the same time this is believed to be one and the same attribute.

Table 2. The weights of the importance of relations

Group of relations	Relation	The value of the weights of the importance
Hierarchy	Genus↔species	*1,2*
	Attribute↔the value of the attribute	*1,2*
	Invariant↔variant	*1,2*
Aggregation	Integer↔part	*1,2*
	Object↔the realization space (localization) of the object	*1,2*
	Object↔property/attribute	*1,2*
	level↔one unit of the level	*1,2*
Semiotic	The term↔way of expression	*0,2*
	The term↔way of representation	*0,2*
	The term↔the main mark of the term	*0,2*
Functional	Object of the action↔action↔subject of the action	*1*
	Reason↔consequence	*0,9*
	Condition↔action	*0,9*
	Fact↔action	*0,9*
	State↔action	*0,9*
	Fact↔state	*0,9*
	Tool↔action	*0,9*
	Data↔action	*0,9*

The set of relations R we divide into types (correlation, hyperonymy - hyponymy, synonymy, holonymy-meronymy) - $R=\{R_1,R_2,...,R_k\}$. n_i indicates the number of relations of type R_i in the thesaurus. Then the total number of relations is $N=\sum_{i=1}^{k} n_i$. We consider that the weight of the ratio is more, when this type of relation is more frequent in the thesaurus. This weight of the ratio we define as $L_i=\frac{n_i}{N}$.

Let us weigh our semantic network that sets the thesaurus. For this purpose we define the weight of the relationship between thesaurus terms. The smaller the weight, the terms are more similar. Therefore, the weight oft he arcs of semantic network is defined as inversely proportional to the weight of such ratio that sets this arc: $l_i=\frac{K}{L_i}=\frac{K\cdot N}{n_i}$, where K is some constant that specifies the amount of weight measurement arcs semantic network [15-17].

We use the thus weighted semantic network to find potential partners who are engaged in similar research issues in the subject area for which the thesaurus was built.

To do this, we should define a set of key terms $C=\{C_1,C_2,...C_n\}$ from the thesaurus, which we believe best define specific research issues. Search Engine finds a set of documents, which contain terms from the thesaurus. For each such document T_s we will build a set with capacity m, which contain terms from the thesaurus that are frequently used in the document T_s: $\hat{C}^s=\{\hat{C}_1^s,\hat{C}_2^s,...,\hat{C}_m^s\}$. By the Floyd-Warshall or DEikstra Method [18] we find $n\times m$ of the shortest distance $d_{ij}^s=d\left(C_i,C_j^s\right)$ between terms from sets C and $\hat{C}^s$. Then we calculate the distance to the document found T_s according to the formula: $d^s=\sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij}^s$. We rank found documents according to increasing values d^s. The authors oft the document with the higher rank may be our potential partners [19-21].

CONCLUSIONS

This article contains the approach to construction of semantic metrics based on the thesaurus of linguistic terms. Detailed description of the process of constructing a thesaurus as semantic network is given. It was proposed to build a set of arcs of the network scales as inversely proportional to the number of relations of a certain type. We constructed a semantic metric based on the weighted semantic network. We consider that this metric can be used to find potential partners who are engaged in similar research issues in the subject area for which thesaurus was constructed.

REFERENCES

1. **Korshunov S.O. 2009.** Tezarazusnoje modelirovanije terminologii i sintaksisa: dis. kandidata filolog. nauk 10.02.19.— Irkutsk— 253.
2. **Tabakova V.D. 2001.** Ideograficheskoje opisanije nauchnoj terminologiji v specyalnyh slovariah. d-ra filolog. nauk 10.02.21. — Tiumen,— 282.
3. **Frolova N.G.** Metod tezarazusnogo modelirovanija kak sposob uporiaducheniya nauchnoj terminologiji. Rezym dostupa: www.cross-apk.ru/.../krt/.../Фролова% 20Н.Г..d...
4. **Strube M. 2006.** WikiRelate! Computing semantic relatedness using Wikipedia. In Proceedings of the 21st National Conference on Artificial Intelligence(AAAI 06).

Boston, Mass., July 16-20/ - http://www.eml-research.de/english/research/nlp/public

5. **Jarmasz M. 2003.** Roget's Thesaurus and semantic similarity / M.Jarmasz, S.Szpakowicz // In Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP 2003). – Borovets, Bulgaria, September.– 212-219.
6. **Fellbaum C. 1998.** WordNet: an electronic lexical database / C.Fellbaum. – MIT Press, Cambridge, Massachusetts– 423.
7. **Wu Z. 1994.** Verb semantics and lexical selection / Z.Wu, M.Palmer // In Proc. of ACL-94.–133-138.
8. **Resnik P. 1995.** Disambiguating noun groupings with respect to WordNet senses / In Proceedings of the 3rd Workshop on Very Large Corpora. MIT, June,– http://xxx.lanl.gov/abs/cmp-lg/9511006
9. **Resnik P. 1999.** Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language / P.Resnik // Journal of Artificial Intelligence Research (JAIR).– Vol. 11. – 95-130.
10. **Lin D. 1998.** An information-theoretic definition of similarity D.Lin // In Proceedings of International Conference on Machine Learning, Madison, Wisconsin, July– http://www.cs.ualberta.ca/~lindek/papers.htm
11. **Smirnov A.V. 2002.** Ontologii v sistemah uskustvennogo intelekta: sposoby postrojenija I organizacii. Novosti iskustvennogo intelekta. – M.: Uzdat. RAII.– № 2. – 3–9.
12. **Sovpel' Y.V. 2004.** Systema avtomatycheskoho yzvlechenyya znanyy yz teksta y ee prylozhenyya / Y.V. Sovpel' // Nauch.-teoret. zhurnal "Yskustvennыy yntelekt", IPShI "Nauka i osvita". — Vyp. 3 – 668–677.
13. **Wu Z. 1994.** Verb semantics and lexical selection / Z.Wu, M.Palmer // In Proc. of ACL-94.– 133-138.
14. **Nykytyna S. E. 1978.** Tezaurus po teoretycheskoy y prykladnoy lynhvystyke. – M.
15. **LytvynV., Shakhovska N., PasichnykV. and Dosyn D. 2012.** Searching the Relevant Precedents in Dataspaces Based on Adaptive Ontology. Computational Problems of Electrical Engineering. – Volume 2, Number 1. – Lviv, 75-81.
16. **Dosyn D. and Lytvyn V. 2012.** Planning of Intelligent Diagnostics Systems Based Domain Ontology The VIIIth International Conference Perspective Technologies and Methods in MEMS Design. - Polyana, Ukraine, 103.
17. **Lytvyn V., Dosyn D., Medykovskyj M. and Shakhovska N. 2011.** Intelligent agent on the basis of adaptive ontologies construction Signal Modelling Control. – Lodz.
18. **Svamy M. 1984.** Hrafy, sety i alhorytmy / M. Svamy, K. Tkhulasyraman. – M.: Nauka,– 256.
19. **Montes-y-Gómez M., Gelbukh A. and López-López A. 2000.** Comparison of Conceptual Graphs Lecture Notes in Artificial Intelligence Vol. 1793. – Springer-Verlag: http://ccc.inaoep.mx/~mmontesg/publicaciones/ 2000/ComparisonCG.
20. **Knappe R., Bulskov H. and Andreasen T. 2004.** Perspectives on Ontology-based Querying International Journal of Intelligent Systems: http://akira.ruc.dk/~ knappe/publications/ijis2004.pdf
21. **Lytvyn V. 2013.** Design of intelligent decision support systems using ontological approach / V.Lytvyn // An international quarterly journal on economics in technology, new technologies and modelling processes. – Vol. II. – No 1. – 31-38.