CALCULATION ALGORITHM OF THE SEMANTIC DISTANCE BETWEEN TEXTS

Akif Suleymanov

Azerbaijan Technical University, Baku, Azerbaijan
akif@inbox.ru

The problem of the information search and its classification became especially keen of late according to the huge rise of information amount. Most of the available sources are the texts in natural language processing of which becomes nontrivial task for a computer.

There are a great number of approaches to analysis of the text with the purpose of getting information from it. Most of them are based on uniting of syntactic analyzer with a certain syntactic machine which determines and stores information about the semantic closeness of word in the text. Such kind of methods gives a good result with a great amount of the same type of information.

But, unfortunately, they are less suitable for building so-called question-answer system which has to give maximum exact answer to the question using minimum quantity of the initial texts. That is, in contrast to the standard searching Internet system giving hundreds, but sometimes thousands texts in which one can see key words from the inquiry, the question-answer system has to give the answer formulated in the natural language.

Apart from its inner expression forms (lexicon), the lexical structure of the language an outer expression in the form of the different kinds of dictionaries and texts. Such kind of expression are also realized in the space where language units follow one-another in the definite order, and that's why there exist definite space between them [2]. But these spaces correspond to that way of the formation of the language units which is taken for their outer expression. As a rule the dictionary units are formed in the alphabet order that provide a formality of their search but don't reflect a real formation of words in the semantic space. It is neatly the same as a list of the unknown geographic names which are nonsense without attaching to a map. In the given case "the attachment" has been realized in the inner structure of the lexical form of language which appears as some analogues of the semantic space [1].

Connection between the word shaving the relation ship system existing between the element lies on the base of the semantic space. A word can be connected not only with one word but also with a lot of other words. With that such connection may be much stronger in one case but less strong in other one. One can consider that stronger connection exists between those words the meanings of which are similar in some case. In this case they usually speak about the closeness of the meaning and consequently about the words designating them. The notion of the word closeness or their remoteness is connected with the idea of existing some distances between them. With that the measure of these distances turn out to be dependent on the inner formation of the semantic space, i.e. on ties and relations existing between the language units on the semantic level [3]. In connection with it the main task for defining of semantic distances is that, the words given in the form of lists, i.e. in the form of the dictionary should be represented on the map of the semantic space. It makes possible to spread the words among their inner formation. In this case defining of the quantitative indices of the closeness and remoteness level of words would be a matter of technique. But the solution of the given task in such way is related with considerable difficulties so far as there are no necessary data for description of the semantic space so brimful and adequate at present. Conceptions of it are contradictory and fragmentary.

In such way the given task exists only as a theoretical possibility. But practically it is solved in other way. For defining the distances among the words we use some of their formal properties, on the base of which their certain regulation is realized. For all that, it's supposed, that such regulation at least partially corresponds to the organization of words on the semantic level and someway even intersects with it. So, here we solve so-called retroactive problem/ if in the first case we consider that internal organization of language semantic space is a means of defining semantic distances among words and consequently their regulation, in the second case – distances, which can be defined on the base of words formal properties and serve as a means of performance of exploiting performance of semantic space. This problem is usually solved by means of mathematical methods and all that explains the necessity of bearing to the formal properties of language units.

Language has two basic forms of external performance: in the form of list or vocabulary, reflecting its lexical structure, and in the form of linear succession of language units in speech (particularly in the text). The formal properties of language units are performed in full strength in the second case, because here they are placed relatively to each other not by chance but on the base of definite language laws, some of which reflects formal aspects of language. That is why mathematical methods are applied, as a rule, for analysis of such regulated successful language units existing in the text. For this purpose we usually use statistic, statistic-combinatorial and other types of mathematical analysis. They take into account first of all such formal properties of language units as frequency of their application, the frequency of their joint finding in taken separately text or in a number of texts in some specific fields. It may be considered that, in fact, the term "semantic distances" appears in diagnostic just in the connection of application of such mathematical methods, because beside of space here we can see the presence of another idea, the idea of measurability of degree, similar to those among words on some basis, in possibility of quantitative expression of the result of these measurements.

The most formal and simple procedure is a counting of frequency of finding separate words in massive texts selected beforehand. With the help of such procedure, as a rule, created so-called frequency vocabularies, in which words are placed according to decreasing of their absolute frequency indices. Vocabularies of such type are widely adopted. Though the aim of such vocabularies is not the defining of semantic distances, in non-obvious aspect such vocabulary allow us to divide words into some groups according, for example, to high, medium, low frequency. These groups in their turn may have more fractional gradation. For all that, words including in one group turned out to be closed to each other than words including into other groups. For example, as a rule, the group of high-frequency words includes preposition and some other link-words. Rather high frequency characterizes words related to current lexicon. Low frequency is a characteristic feature of specific subjects or certain text. Frequency homogeneity of words within such groups may be interpreted as "distances" among them, detected in the level of their frequency characteristics. These "distances" may be essentially specified when, parallel to frequency, other formal word characteristics are used. One of widely used characteristic is a distribution of words in analyzed texts. Necessity of accounting of such parameter is connected with the fact that words, which have got the same frequency, may be presented in the text in different ways. Some of them may have got more number of entering (including) in a small number of texts, while o theirs may be found rather seldom in each taken separately text, but in their great number [1].

Let's examine the mathematical model of algorithm of word finding by frequency method in detail. While solving the problem of classification words into such groups as sublanguages (ρ_i), semantic fields we introduce the term "relative entrance" (α) and "middle entrance" (β), on the base of which we can do all the calculations which are necessary for this purpose. It defines α as:

$$\alpha = \frac{\varphi}{\omega}$$

φ is a number of standard extracts in which this word has been observed;

ω is a number of all examined extracts;

$\alpha = 0 \ldots 1$.

The definition of β is based on the following laws. In texts, related to any $\rho_{i,}$ frequency of using of special words increases in several times, while frequency of general used (current) lexicon essentially decreases. In spite of this, quantity α displays more stability. If we add "weights" (λ_i) to sublanguages we can calculate middle weighted entrance (γ). At equal λ_i, different ρ_i $\gamma = \beta$. This gives an opportunity to distinguish general vocabulary from special one objectively.

The words in which a $\alpha \approx \beta$ in each of ρ_i belong to general language vocabulary. Those words who's $\alpha >> \beta$ in one of ρ_i have, therefore, clear inclination to this ρ_i stipulated by their semantics.

Here the specificity of the word (δ_i) is deduced which is defined as:

$\delta_i = \dfrac{\gamma}{\rho_i}$, the quantity of which must be >2.

$\delta_i > 0$, when α > 2.

$\delta_i < 0$, when ratio is inverse.

The introduced notions having the quantitative expression are used then as an instrument for the successive elimination of the specific words from the general vocabulary on the different levels. The specific vocabulary chosen on the one level forms ρ_i of the given level or rank. The totality of ρ_i chosen in this way forms the hierarchy.

To chose the semantic fields in the given approach the special coordinate net in the semantic space in used. This net is placed in accordance to each ρ_i on which the meanings δ_i typical for the word in the corresponding ρ_i are saved. Here the positive δ_i is designated +1, the negative -1, neutral 0. The obtained totality of the triple numbers (-1, 0, 1) is considered a vector of specificity δ_i of the described word. This gives a possibility to define the semantic fields as a sub plurality of words having $\delta_i > 0$ for the same ρ_i. It follows from the definition that the semantic fields can have as many ranks as they exist in the hierarchy ρ_i. The semantic field of the largest rank that is, of the lowest level in the hierarchy is called "phratria". It is known that the words belonging to the same phratria have maximum terminology closeness and are associated with each other most closely in the consciousness of the native speaker.

The result of the application of all these procedures is that numerous words of certain texts are subdivided into some groups which are in hierarchy relations with each others. They are characterized by certain uniformity and the bigger is rank of the chosen group the larger is the uniformity. You may consider them being semantic in contrast to the groups chosen only on the base of a private principle. But this articulation is big enough and in this connection the uniformity of words is relative. There for we can talk about semantic distances here with definite conditional degree. Probably in connection with this for determination semantic distances (s) which is equated with measure of proximity meaning applies another formal procedure of the heart of lies submission about semantic (σ) and paradigmatic (τ) remoteness of words. It consists in.

Accepted admission that if two words which is important to determine the distance between them is found near each other (for example, "radioactive" and "element") then positional (s) equals 1. In the case if these two words are within one sentence but separated with other words, then positional (s) equals to amount separated them words. After determination positional (s) in each phrase can be computed average out this distance s_{av}. Then frequency of occurrence of 2 words is determined within the limits of one phrase (ψ). For this result s_{av} must be increased to ratio of phrase amounts, containing any one from considered words (k) and to general amount of phrases containing even one of these words (k_{gen}).

$$\psi = \sigma = s_{av} \cdot \left(\frac{k}{k_{gen}} \right).$$

After that it is determined τ. With this purpose for analyzing copied out the words which are from the right. Then make division of total amount of words ($r_{totright}$) which are from the right of only one of them to general amount of words ($r_{genright}$) which are from the right of even one of them.

$$r_{right} = \frac{r_{totright}}{r_{genright}}$$

Similar procedure realized with the words which are from the left of analyzing words

$$r_{left} = \frac{r_{totleft}}{r_{genleft}},$$

which gives the possibility to account half-sum to two fractions. This half-sum is modulus of paradigmatic positional remoteness:

$$\tau = \frac{\sqrt{r_{right} + r_{left}}}{2}.$$

As a result of squaring the sizes, corresponding syntagmatic and paradigmatic item remoteness, multiplication of these squares to some leveling factors, additions of the received results and extraction from the sum of a square root, the 2-componental vector of semantic remoteness of considered words or semantic distance (d) between them turns out:

$$d = (a\sigma^2 + b\tau^2),$$

here a, b - leveling factors.

Definition of semantic fields is carried out in some stages. First of them is an establishment of connections between words on the basis of their joint occurrence and reflection of set of these connections as the column in which edges not only reflect presence of connection, but also and her intensity as the certain quantity indicators. The further task consists in allocating on it to the column separate fields, i.e. groupings more words close among themselves. The special algorithm which begins the work with association of the closest words is applied for this purpose, passing then to farther.

It is considered to be, that the semantic link between elements is carried out when in them there is something the common. In this case it means, that values of the words connected among them selves are among themselves concerning crossing. The size of crossing, i.e. the common, repeating contents, defines also force of connection. The more the area of crossing, is more and it. If it so such connection means as well that words taking place in these relations in something are similar under the contents, are close, similar each other. On the basis of such similarity of a word are united in the groups named lexico-semantic. These groups, in turn, can be connected among themselves as some independent formations. The system of relations both inside lexico-semantic groups, and between them allows counting all this that makes structure of semantic space.

References

1. Nekrestyanov I., Panteleyeva N. The system of test searching for web // Programming. M., 2002, N4.
2. Suleymanov A.Sh. System of language searching / Reports of international conference “Informational tools and technologies”. Moscow, 2003, pp. 58-61.
3. Сулейманов А.Ш. Метод определение контекстных слов при анализе текста Информационные технологии, теоретический и прикладной научно-технический журнал, 7(155), М., 2009, стр. 46-49.