Automated Detection of Semantic Connections in the Text Subject Organization

Irina Nikolaeva
Department of Applied and Experimental Linguistics
Moscow State Linguistic University, Moscow, Russia
irenkarm@yandex.ru

Abstract
The given report gives the review of different semantic connections among words and terms, which are used as a tool for subject organization. Here we see the most important task in defining the semantic connections, which are used for connected text organization. We assume that the way of automated theme detection will increase the efficiency of processing.

1. Introduction
The following possibilities of text processing are realized in the contemporary systems of information processing: key words determination, abstract creation, hypertext structure formation, text navigation, text classification.

The basic circuit of text processing is invariant for all natural languages. The analysis goes through the same stages independently on the language. First two stages (slitting into sentences and words) are the same for the majority of natural languages. The only steps where specific characteristics of a given language may be revealed are the abbreviation processing and punctuation processing.

Semantic analysis is based on the result of the previous processing stages which are always specific for a given language. Consequently, the ways of result presenting may vary, influencing greatly the methods of semantic analysis. The results of analysis, carried out on early stages, may be multiple-valued: there may be several meanings for output parameters (for example several ways of one and the same treatment of a given word). In this case all the next stages should choose most probable meanings of the earlier results and carry out further analysis on their base.

2. The basic principles of connections
These are the main principles of conceptual connection [6] description from the point of lexical connectivity mechanisms investigation:

1) The conceptual connections are constantly checked whether they reflect real lexical connections in the text.
Border troops “Putina” exercise to control poaching.

The border troops “are not saber rattling” in Russian territorial waters in the Far East as the mass media, especially the Japanese mass media, are attempting to portray it. Servicemen have been legally granted the right to utilize all the tools at their disposal, including weapons, to put a stop to poaching. Russian Border Troops Commander-In-Chief Colonel-General Andrey Nikolaev stated that to an ITAR-TASS correspondent while stressing that his subordinates are conducting a strict policy to put a stop to the illegal activities of foreign boats. He noted that the President of Russia supports the position of the border troops for the full observance of the law in the country’s territorial waters.

Incidentally, Japanese fishermen (read poachers) have learned about the operation beforehand that is called upon to put pressure on them. This is certainly how we explain why they have recently stepped their activities. So, just from 26 March through 1 April and only in the Southern Kurile direction (Izmena Strait and Tanfilyev and Anuchin islands), 49 Japanese boats undertook attempts to poach. The schooners penetrated up to 55 cable length (one cable length is approximately 200 meters) into Russian territorial waters. But then again, the problem is much broader than just putting a stop to poaching in our territorial waters.

It is also having or not we, having established monitoring of fishing, will be able to conduct our own fishing for fish and crabs in these waters using our own men and equipment, in these waters that have been designated by and that are so familiar for Japanese fishermen? That is not an idle question. The proposals to sell fish to the Japanese that are being increasingly loudly stated today are certainly well thought out. How?

Quite legally: By increasing their quota to catch fish in Russian territorial waters for hard currency. In a word, even in this case the border guards will not be standing idle by. Along with fish conservation personnel, they could carry out monitoring and continue to defend Russia’s economic interests in the region.

Three basic notions are Russia, territorial waters and fish. The subthemes should reveal the connections among these three notions. Let us determine the words and word combinations which compound the subject units of the basic notions, that means that they are used as references to these notions in the text.

<table>
<thead>
<tr>
<th>Russia (Russian)</th>
<th>Territorial waters</th>
<th>Fish</th>
</tr>
</thead>
<tbody>
<tr>
<td>State (country)</td>
<td>Ocean</td>
<td>Fish resources</td>
</tr>
<tr>
<td>Far East</td>
<td>Water transport</td>
<td>Natural resources</td>
</tr>
<tr>
<td>Curile</td>
<td>Ship</td>
<td>Fishing (catch fish)</td>
</tr>
<tr>
<td>Southern</td>
<td>Schooner</td>
<td>Vessel</td>
</tr>
<tr>
<td>Curile</td>
<td>Island</td>
<td>Fisherman</td>
</tr>
<tr>
<td>President of</td>
<td>State</td>
<td>Illegal</td>
</tr>
<tr>
<td>Russia</td>
<td>Country</td>
<td>Pouching</td>
</tr>
<tr>
<td>Japan</td>
<td>Border</td>
<td>Catch</td>
</tr>
<tr>
<td>Japanese</td>
<td>Troops</td>
<td>Offence</td>
</tr>
<tr>
<td>Country</td>
<td>Guards</td>
<td>Activity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Violence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Illegal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fisherman</td>
</tr>
</tbody>
</table>

The arrangement of the words, corresponding to these subject units, create the effect of lexical chains. We see that the structural organization of the given text is built on hidden semantic connections [10] among the words.

Let us determine the basic types of relations, used for text organization.

1) Semantic connectivity is carried out on the base of one and the same concept [11]; it is realized by semantic repetition or synonyms: state – country.

2) Semantic connectivity is carried out on the base of the notions which are directly connected between each other: country – HIGHER – Japan, violence – ASSOCIATION – illegal.

3) The way from one notion to another includes a set of HIGHER and WHOLE connections. For example:

- troops - HIGHER - border - WHOLE - ocean - HIGHER - territorial water.
- Southern Curile - WHOLE – Curile - HIGHER – state - WHOLE - Russia.
Here we observe succession of properties of element s of the then inflection and a set of PART and LOWER connections. The way consists of two unidirectional fragments [1 2]: the first fragment is a set of HIGHER and WHOLE connections, then inflection and a set of PART and LOWER connections. Here we observe succession of properties of elements of the first branch in the second one. For example:

- *catch* - HIGHER - *fishing* - HIGHER - *fish*

This semantic connection obtaining is based on the transitivitiy of HIGHER-LOWER and PART-WHOLE connections.

The way consists of two unidirectional fragments [12]; the first fragment is a set of HIGHER and WHOLE connections, then inflection and a set of PART and LOWER connections. Here we observe succession of properties of elements of the first branch in the second one. For example:

- *Natural resources* - HIGHER - *fish resources* - PART - *fish*;

5) Another type of inflection is possible: firstly, LOWER and PART connections, then WHOLE and HIGHER connections [13], for example:

- *border* - LOWER - *troops* - HIGHER - *ocean* - WHOLE - *territorial waters*.

- *offence* - LOWER - *activity* - WHOLE - *fish*.

### 4. Subject junction detection

When analyzing the received subject junctions, it is obvious that subject junction detection, and consequently, the main notion detection is rather a difficult process. Here is the list of main analysis difficulties [14]:

1) the necessity of including the notions *Russia* and *territorial waters* into the list of key notions is not clear from the title (the key notion *fish* first appears in the middle of the text), but follows from the whole text, from the connections among different parts of the text;

2) the majority of text notions are connected by different types of connections at the same time. For example the notion *state* is at the same connected with the notions *Russia, Japan, country,* and the notion *troops* – with the notions *guards, territorial waters, border,* etc;

3) the notions of the main theme are connected by the ways of the same conceptual type, which were used for subject junctions creation;

4) the connections and the ways may vary: in one text the elements compound one subject junction, in another text the same notions include different junctions.

It means that the right junction detection and notional detection demands multiaspect analysis [15]. Without reading the whole text even a human is not capable of detecting the whole set of semantic connections important for the text organization when analyzing sentence after sentence. It means that the automated system should firstly reconstruct the notional net and detect the connections, which may be used in the subject junctions, and after that analyze their distribution in sentences.

After the subject junctions have been created, the text connections of each subject junction are summed up and the connections between the subject joints are determined. The principal subject junctions are those which are all connected with each other and their connection frequency sum is the highest possible. Hence, the mentioned subject “threads” [16] run all through the whole text.

#### 5. Joint occurrence in the sentences

In order to find the main notions in the text, we need to find “the strongest” semantic chains in the text. But only the basic text theme determines notional net splitting into chains.

As notions united into sets (macronotions) [17] describe the general subject of the text, we assume that the connected text describes connections among notions. That is why the general contents of the majority of subsubjects is the description of connection between the elements of different subject junctions (subject junctions around notions). This means that the pairs of terms, belonging to different subject junctions, must appear in the text more often than the terms of subject junctions built around other notions.

Consequently, we approve that the text is devoted not to notional discussion, but to interword connection description. That is why not only basic notions and elements of their subject junctions are the most frequent, but also the pairs of different basic notions and the elements of their subject junctions, which should occur in the text more frequently than, for example, the elements of one and the same subject junction [18]. It means that if two notions, connected in thesaurus by one of the five conceptual ways, occur frequently in the same sentences, then none of these notions includes in the subject junction of another one.

Really, let us consider how the pairs of notions are distributed in the sentences of the text: 1) *Russia (state, country) – territorial waters (ocean)*, 2) *Russia (state country) – fish (fishing, catch)*, 3) *territorial waters (ocean) – fish (fishing, catch)*. We see that the first pairs occur in the same sentences 5 times, the second pairs – 3 times, the third pairs twice.

#### 6. Conclusions

Everything mentioned above allows us to draw the following conclusions.

1) The notions discussed in the text are connected by different semantic and encyclopedic connections, the majority of which may be used for thematic structure organization of this or that text;

2) When sliding through the text according to sentence order, it is impossible to find word sets corresponding to the main theme elements, because the exact set of used connections and the splitting of all words into sets of “close” words strongly depends on the whole text composition, on the main theme.

3) In order to detect automatically the main text notions it is necessary:

- to recover the net of notions;

- to split the received net into the set of subject junctions through the detected ways of notional connections within the junctions, based on mutual occurrence of the notions, connected in thesaurus in the same sentences;
to detect the main subject junctions among the totality.

The described systems keep being developed within the research work carried out under the supervision of Doctor of Linguistics, Professor, International Academy of Informatization Academician – Rodmonga Potapova at the Department of Applied Linguistic at Moscow State Linguistics University.

7. References

1. Potapova R.K. Novie informacionnie tehnologii i lingvistika - Moscow: Editorial URSS, 2005
7. Loukachevitch N. Text Summarization Based on Thematic Representation of Texts