FORMAL REPRESENTATION OF THE TEXT CONTENT

Kravchenko Sergei

Department of Informatics and Applied Linguistics, Minsk State Linguistic University,
21 Zakharova st., 220062 Minsk, Belarus, e-mail: *SergeyKravchenko@scnsoft.com*

Abstract. Current article overviews questions of text theory, methods of text content modeling, basic formal models of content representation and their suitability for multilingual texts. The usage of an existing formal language (SEMSYNT) is suggested for overall text content description, syntactic and semantic levels of representation. The research is based on the material of English, French and Russian electronic documents with well-defined formal structure.

1. Introduction

The problem of formal text content representation is one the major problems of the modern computational linguistics. Text content formalization is required for a number of tasks such as information extraction, document annotation, summarization and classification, machine translation, etc. In our work we will describe an approach to the task of formal text content representation for the systems of multilingual text processing based on an existing formal language SEMSYNT.

2. Problem

In order to determine the boundaries of this problem and of our research, it is necessary to clearly identify and separate such basic notions as *text content* and *text sense*. Many researches on text and text structure has already been undertaken in the past. As it has been shown in [7] – text can be considered as a linguistic sign and described in terms of semiotic theory.

The detailed analysis of the text as a sign has been undertaken in the work [12]. As it is noted in the research, the origin of the text is a certain situation. In this context, *text sense* can be presented as a set of the most relevant elements of this situation. *Text content* is then considered as a number of linguistic means used to convey the text sense to the recipient.

According to the different points of view on the text sense it is obvious that there exist different approaches to the text content structure. Such approaches are formulated in different scientific disciplines, which deal with the phenomenon of text – psycholinguistics, literature, computational linguistics, semiotics, etc.

While analyzing text from the position of the psycholinguistics [12], the main focus of interest is turned towards the individual. The processes of the text generation and perception are considered as a result of the intellectual activity of the individual. The psycholinguistics treats the text from two aspects:

- As a copy or a model of the reality, existing in the consciousness of the individual, which was created in accordance to the goals of the individual's activity.
- As a specific projection the individual's linguistic system, conditioned by the current communicative situation.

According to the generalized scheme of the speech act production, the text content can be represented in different ways on each stage of the sentence production. At the first stage, the text is represented as "the image of the result" or "the future model". Then it is transformed into a hierarchical net of themes, subthemes and microthemes, which is in turn converted into a set of predicative statements.

In the study of literature, the text content is usually studied as a union of intra- and extralinguistic features, while the accent is, generally, made on the text expressivity [6]. The rhythm is considered as the first organizing element of the text content. The rhythm is itself treated as a union of two factors: uniformity, the repetition of some of the text elements and regular artistic and expressive divergences from this uniformity. The second factor of the text organization is a fable. Each writer has his/her own way of organizing the story line, events and characters, therefore, the individual writer's style is considered as one of the most important features of the text content.

From the point of view of the formal study of the text content, it is possible to define the two basic aspects of the text [12]:

- Static
- Dynamic

Such division reflects the basic structure of the natural language formations – semantical and syntactical aspects. In other words, "static" stands for vocabulary, "dynamic" - for syntax. As it has already been mentioned the text represents a certain situation. This situation presupposes the presence

of some *characters*, their *actions*, which take place in certain *conditions* and at a certain *time*. It is obvious, that all these elements of the situation have to be actualized in the form of concrete words. Such words form the basic text content, i.e. words, which are directly related to the text conception. The distribution of these words in the text represents its *static structure*. It is evident that the static structure alone is not sufficient to represent the overall text content. A number of relations between the elements of the basic text content form the *dynamic structure*, which is reflected in the text syntax.

Resuming everything mentioned above, we are able formulate the problem of our research. In order to create an overall text content representation, it is necessary to solve the following tasks:

1) Extract the basic text content elements.
2) Identify the relations between these elements.

3. Research

A number of methods and approaches have been developed to represent both static and dynamic aspects of the text content. We will try to review these methods from the point of their applicability to the multilingual text processing systems (such as machine translation, information retrieval, etc.).

Three basic methods of the **static text structure** acquisition can be mentioned here: statistic, pattern-matching and integrated parsing [3].

While providing very good results, both - pattern matching and integrated parsing approaches rely heavily on linguistic information and language structure data. Statistical approaches are limited to the word frequency and distribution data.

In pattern matching approach, human knowledge engineers create sets of "patterns", which can recognize words and phrases that identify and distinguish the basic text content. Documents are then automatically scanned to find instances of the patterns, and are classified accordingly to the results of the content extraction. While achieving good results in content extraction accuracy, this approach has some disadvantages - one of them is that it cannot be easily applied to multilingual texts since a significant human effort is required to develop the patterns.

Integrated parsing employs natural language processing in order to extract text content and to get rid of the syntactic and semantic ambiguity. Using the parsing results as a base for the content representation, this approach can provide very good results. But as it is noted in multiple researches, the lack of robustness is a serious problem for systems using integrated parsing for text content analysis. It has been noted that even with a large grammar, the parsing does not cover a significant amount of the previously unseen text. Moreover, text parsing is language dependent and, usually, it cannot be successfully applied to the languages of different structure.

Term frequency based approaches for basic text content extraction and representation have been developed and refined over a period of many years. These approaches use statistical methods to analyze documents based on the frequency distribution of "terms" in the documents. A large number of techniques for the analysis of both - unstructured documents and the ones with well-defined structure have been developed. It has been experimentally proven that this approach can be successfully applied to any type of document without any significant modifications.

One of the most advanced approaches to the statistic-based text content extraction has been described and tested in the work [10]. The procedure described in this article has been successfully applied to a large number of documents of different types and structure. We also believe that this technique can be used for text content description in different languages, and therefore – implemented in multilingual text processing systems.

As it has been noted earlier, the **dynamic text structure** should characterize the relations between the elements of the situation, described in the text. In other words, it should reflect the events (processes, actions, and states) which occur to the referents represented by the basic text content. Such events should be represented in the text as certain linguistic entities. However, there is no unanimity among the researchers on the structure of such entity – the theories state from a single word or a syntagmatic formation to the sentence and the entire text. As it has been stated and proved in the work [8], the text paragraph can be considered as such linguistic entity. The choice of the paragraph as the main component of the dynamic text structure is not eventual. Such choice is based on profound researches and experiments, which we will not examine here. The experiments prove, that in order to create an overall description of the text content, it is necessary to take into account not only the relations between words and word combinations within sentences, but it is also vital to register the relations between sentences and paragraphs. Based on this theoretical issue, we can suggest a corresponding method for the formalization of the dynamic structure of the text.

As our study shows, there are not many formal models, which take into consideration the relations between sentences within a paragraph, not to mention the relations between the paragraphs

[2,5,9]. The most suitable solution for the task of representation of the dynamic structure of the text we see in adaptation of the existing formal language of the text content description, which meets the needs stated above.

The semantico-syntactical language SEMSYNT takes into account the requirements necessary for the formal text content representation [13].

SEMSYNT includes the following components:

1) language alphabet;
2) means for registering of the semantical relations between the parts of a sentence;
3) means for registering of the syntactical relations between the parts of a sentence;
4) mechanism for registering of the logico-semantical relations between the sentences and the paragraphs;
5) mechanism for the registering of the text theme;

The SEMSYNT language is based on the case grammar theory developed by Ch. Fillmore [1]. The general semantic structure of the sentence is presented as one or more predicate groups and their arguments (argument groups). Semantic relations between predicates and argument groups are presented as semantic functions (e.g., “agent”, “receiver”, “object”, “location”, etc.).

As an experiment, we have encoded a number of parallel english-french texts using SEMSYNT language notation in order to compare their formal representation (syntactico-semantical formulae) and evaluate the language’s fitness for multilingual text processing systems. Here we have presented two fragments of an e-mail text in French and in English languages. Each natural language sentence is followed by its semantical formula, where:

Semantic classes of words:		**Semantic functions:**	
N*n*	NOUN	Agi	Inanimate agent (action initiator)
K*k*	PREPOSITION	Aga	Human agent (action initiator)
Z*z*	DETERMINER	AR	Receiver of action
V*v*	VERB	AT	Action time modifier
J*j*	ADJECTIVE	AM	Action manner modifier
M*m*	NUMERAL	AL	Action location modifier
P*p*	PRONOUN	AO	Object of action

FRENCH VERSION:

1. *Soumission d'un article*
   N54*K10*Z10*N10
2. *L'article doit être écrit en français ou anglais, avec résumés et mots clés dans les deux langues.*
   AGi<Z01*N10>+R2<V15*V02*V18>+AM<K02*N13|N15> , AM<K04*N16&N17*J09> AM<K03*Z01*M02*N12> .
3. *Sa présentation doit suivre les règles indiquées sur le site de la revue (instructions).*
   AGi<Z02*N30>+R3<V15*V19>+AO<Z01*N33*V22*AL<K05*Z01*N50*K10*Z03*N31>> (N35) .
4. *Les manuscrits soumis doivent être envoyés à l'adresse suivante : arima-office@inria.fr*
   AGi<Z01*N36*V37>+R4<V15*V02*V37>+AR<K05*Z01*N38*J10 : N39>
5. *Il est possible de mentionner, avec l'envoi, l'un des rédacteurs principaux comme destinataire.*
   R9<P07*V02*J11*K10*V40> , AM<K04*Z01*N51> , AO<Z01*M01*K10*Z01*N52*J12*AD<K11*N53>> .

ENGLISH VERSION:

1. *Submitting a paper*
   R3<V37>+AO<Z10*N10>
2. *The paper must be written in French or English, with abstracts and keywords in both languages.*
   AGi<Z01*N10>+R2<V15*V02*V18>+AM<K02*N13|N15> , AM<K04*N16&N18> AM<K03*M02*N12> .
3. *The guide to prepare the manuscripts site at the website of the journal (information).*
   AGi<Z01*N33*K13*V41>+R12<V42>+AL<K05*Z01*N50*K10*Z03*N31> (N35) .
4. *Manuscripts must be submitted under an electronic format at the email address : arima-office@inria.fr.*
   AGi<N36>+R4<V15*V02*V37>+AM<K12*J14*N56>+AR<K05*Z01*N38 : N39>
5. *In their mail, the authors may suggest to which principal editor they address their manuscript.*
   AM<K04*Z14*N51> , AGa<Z01*N57>+R5<V44*V40> AR<K13*Z16*J12*N52>+AGa<P07>+R7<V37>+AO<Z14*N36>

These fragments are not identical – some of their sentences (1,3,5) contain different syntactic constructions specific for English and French languages. However, the comparison results prove that even with different syntactic constructions, the sentences can be translated correctly into both languages as shown in Fig.1.

French sentence:
Sa présentation doit suivre les règles indiquées sur le site de la revue (instructions).

↓

Sentence representation:
AS1<Z02*N30>+R3<V15*V19>+AO<Z01*N33*V22*AL<K05*Z01*N50*K10*Z03*N31>> (N35)

↓

English sentence (based on French sentence representation):
Its presentation must follow the rules specified on the website of the journal (information).

Fig.1

4. Conclusion

In our research, we have attempted to find the most efficient way for the text content representation. As multiple previous studies show, the text content can be considered from different aspects and points of view. One of the most developed text theories affirms that the text is a multiplanar entity, which can be analyzed from different sides. We have suggested a number of techniques for formal representation of different text aspects resulting in overall text content representation. We suggest that these techniques can be used for such actual tasks as machine translation, information retrieval, automatic document classification, etc.

References.

[1]. Ch. Fillmore, "The case for case", New in foreign linguistics, Moscow: Progress, 1981.
[2]. K. Haenelt. "A Context-Based Approach Towards Content Processing of Electronic Documents", Germany (2000).
[3]. Carole D. Hafner, "A Linguistically Sound Approach to Content Analysis of Natural Language Text", Boson (1994).
[4]. S. E. Kuncevitch, "Formalization of the content of the english technical documents", Minsk (1997), 87-94.
[5]. I. A. Liokumovitch, "Sentence semantic and its formal representation", *Problems of computational linguistics*, Minsk (1997), 57-67.
[6]. V.A. Maslova, "Linguistic analysis of the fiction text", Minsk, Visheishaya shkola (1997).
[7]. R.G. Piotrovsky, "Linguistic apparatus and its speeking and congnitive basis", Minsk (1999).
[8]. I.P. Udinskaya, "Text and its main elements", *Problems of computational linguistics*, Minsk (1997), 51-56.
[9]. V.S. Yakovshin, "Language of the formal syntagmatic structures", *Computational linguistics and language learning*, Minsk (2000).
[10]. A. Zubov, "Formalization of the Procedure of Singling out of the Basic Text Contents". Minsk (2003).
[11]. A. Zubov, "Static aspect of the text content and its formal representation", Tartu (1986), 75-94.
[12]. A. Zubov, "Text generation problems. Part I. Theory and Algorithms", Minsk (1987).
[13]. A. Zubov, "Semantico-syntactical language for text representation in PC memory", Minsk: MSPIFL (1990), 110-116.