International Conference
ICT for Language Learning

Information Structure of Contemporary Popular Scientific and Technical Text

Larisa Ilinska, Marina Platonova, Tatjana Smirnova
Riga Technical University (Latvia)
Larisa.ilinska@rtu.lv, marina.platonova@rtu.lv, tatjana.smirnova@rtu.lv,

Abstract

Investigations on the information structure of scientific and technical texts have become particularly topical with the introduction of new methods of text analysis using corpora and text processing software.
The concept 'information' is closely related to such notions as knowledge, meaning, comprehension, constraint, perception, representation, and communication. Following Shannon, Weaver (cf. [9]) proposed analyzing information considering (1) technical problems associated with the quantification of information; (2) semantic problems relating to meaning; (3) problems concerning the impact and effectiveness of information on human behavior.
Considering the general advancement of information technologies in any field of human activities, text processing tools should be able to perform multiple functions, including classifying texts according to genres and functions, distinguishing intra- and cross-disciplinary polysemic items, decoding different models of meaning extension. The computer software can manipulate long texts and/or separate sentences with the purpose to obtain relevant information in a user-friendly way. The challenges with processing of information and its extraction from the text are rooted in the fact that even the most advanced statistical methods are incapable to perform many tasks unless they are combined with the methods of cognitive analysis. The combination of both approaches is aimed at fast and efficient extraction of value from volume (cf. [7]).
The issues addressed in the present article include the information structure of popular scientific and technical texts, their hierarchical organization, and the problems of decoding of meaning at different levels in the process of information processing and extraction.
It should be kept in mind that in linguistics "…the term 'information' is not meant to be restricted to cognitive knowledge, but includes any possible item which is somehow present in the mental world of individuals, including their preconceptions and prejudices" [1]. Therefore, in order to establish the theoretical framework of the research, the semantic, pragmatic, cognitive and textual analyses of the texts on Telecommunications, Architecture, and Civil Engineering have been performed.

1. Introduction

Originally, information structure was studied within the structuralist approach to analyze topic-comment articulation within a sentence. Starting from the 1960s, a number of linguists became interested in what was determined as information structure or information packaging. The latter was defined by Lambrecht [5] as "the formal expression of the pragmatic structuring of a proposition in discourse".
Schwabe and Winkler argue that "the term Information Structure refers since Halliday to the linguistic encoding of notions such as focus versus background and topic versus comment, which are used to describe the information flow with respect to discourse-givenness and states of activation" [8].
Later, the scope of the concept expanded to include other dichotomies characterizing relations between the new and given information. Nowadays in order to investigate the way how background knowledge influences the creation of new meanings in the course of text interpretation, information structure is analyzed in terms of three basic dichotomies, namely, *focus vs. background*, *topic vs. comment*, and *given vs. new*.
Focus is represented by the highlighted elements, whereas background is made by complement notions. According to Steube [10], background constituents express familiar information, whereas focus constituents express non-familiar information which, being new information for the reader, i.e., it has not been verbalized before in the communicative situation, is highly dependent on the context.
Dik [1] defines pragmatic information as the full body of knowledge, beliefs, assumptions, opinions, and feelings available to an individual at any point in the interaction. Three main components of pragmatic information include general information (world knowledge), situational information (experience based), and contextual information.

Comprehension and interpretation of popular scientific and technical texts on the semantic and pragmatic levels require that the readers have certain cognitive potential, i.e. are familiar with the conceptual framework. Thus, they cannot only recognize, decode and apply terms, but also understand the conventions of professional communication in the given field and be able to understand implicit meanings.

All three dimensions of the information structure are related to the concept of intertextuality, because they represent the interaction between the known and the unknown information. However, it is the third dichotomy, namely, given/new, where this relation is most evident, because this dimension characterizes how the given text is related to the preceding texts and the preceding knowledge.

In the contemporary scientific and technical text, new information is often brought into focus using various foregrounding techniques such as application of metaphoric terms, allusions, proverbs, idioms, and terms belonging to different fields of knowledge. The application of stylistically marked vocabulary within the scientific and technical text allows focusing the attention of the readers on a particular information cluster, ensuring that the new information is not disregarded or missed.

The challenges associated with processing of information and its extraction from the text are rooted in the fact that even the most advanced computer-aided text processing methods are incapable of performing many tasks unless they are combined with the methods of cognitive analysis. Modern text processing tools should be able to perform multiple tasks, including classifying texts according to genres and functions, distinguishing intra-disciplinary and cross-disciplinary polysemic terms/words, decoding different models of meaning extension, and culture-specific items. Investigations on the information structure of scientific and technical texts have become particularly topical with the introduction of new methods of text analysis using corpora and text processing software.

2. Information Processing and Extraction

The majority of texts belonging to the technical discourse are organized according to definite conventions and possess a relatively rigid structure that facilitates the transmission and interpretation of information. Loose structure may block comprehension and hinder the performance of the primary informative function of the text. Contemporary scientific and technical texts are characterized by a growing degree of hybridity, both in terms of its contents and style.

The higher is the degree of formalization of the language, the more controlled is the process of vocabulary creation within it. On the one hand, it facilitates the process of data mining and information extraction, because with the clear information architecture (certain order of data representation) and pre-defined set of representative features (stated tokens) it is considerably easier to use “a cascade of transducers or modules that, at each step, add structure to the document and, sometimes, filter relevant information by means of applying rules” [3].

On the other hand, it becomes more difficult to adjust the text processing software to any changes in the order of the given information, which does not make the data mining and information extraction systems “as portable as possible to new situations” [11], because “new extraction scenarios could imply new concepts to be dealt with, which are beyond IE system’s capabilities” (ibid).

As a result, the computer is neither able to compensate the loss of stylistic coloring when aligning a metaphoric term into a more formal language, nor it is able to generate a metaphoric term based on the associative mining function to be used in the less formal language. In other words, the computer software is not advanced to that extent that it is able to assign meaning to linguistic expression taking into account the existing information, i.e. micro- and macro-context (cf. [4]). Therefore, for the natural languages with variable degree of controlled vocabulary creation, the processes of data mining and information extraction (IE) become sophisticated and highly dependent on the human intervention.

If computer-aided text processing tools are used in the process of translation, the deficiencies of the existing software become more apparent. Not only should the concept systems of the given domain coincide in two languages to map the term, the tolerance towards the unconventional term-formation process and degree of natural language formalization should also be comparable. At this point it is important to specify that the formalization of the language can be seen both as a means of controlled language development and as a process, which allows making certain meaningful transformations computationally, considering language as an isolated phenomenon separate from cognitive processes (cf. [2]).

The ability of computer software to search for the inquired information is closely linked to the ability of the human to set the criteria, frame the content, define the expected scenario, and predict the results. In other words, for the computer to successfully perform information mining function, “the type of content to be extracted must be defined a priori” [11]. It would block or considerably reduce the likelihood of sorting unexpected results, which, for the computer, are equal to irrelevant information. Therefore, computer-aided information extraction used to identify a definite set of concepts in a

particular field disregarding the irrelevant information is not always sufficient. According to Piskorski and Jangarber [6], the process of information extraction "involves identification of certain small-scale structures like noun phrases denoting a person or a person group, geographical references and numerical expressions, and finding semantic relations between them".
Even with the defined parameters for the associative mining, the computer will be able to trace only certain taxonomic associations limited by the domain, but it will not be able to end up with unexpected, although very relevant, data and establish associations beyond the initial domain of search.
It means that for the texts to be used by the computers, they should be highly structured, so that information processing and extraction can be organized in a straightforward manner (cf. [11]). Changes in the order of information representation lead to the results beyond the expected scenario, which would basically mean that computer fails to perform the task unless assisted by the human. These challenges are more evident in case of interlingual information extraction setting, i.e. translation of the text or alignment of terms across the languages.

3. CAT Tools in Alignment of Terms

Many principles of IE are similar to the methods applied to the analysis of a text to be translated into another language. The translator should possess both linguistic and technical domain competence to interpret and render the precise meaning of polysemic and/or not fully equivalent terms.
At present translators use resources and opportunities offered by computer-aided translation (CAT) tools and computer corpora and databases, but they also use strategies and methods of text analysis and message interpretation that cannot yet be used by machines. Modern computers are not capable of distinguishing and interpreting shades of meaning in the context, or establishing exact lexical relations between the items in the text.
One of the challenges characteristic of popular scientific and technical language is the tendency for uncontrolled metaphoric meaning extension of the existing lexical items. The process of extension of meaning by metaphorization is one of the main reasons for appearance of polysemic terms. For example, the entire ontology of the meanings of the term "body" as used in the field of chemistry may be presented as follows: *consistency, saturation, coverage capacity, strength, proof, viscosity, density, thickness, extractivity, intensity, glutinosity*. The exact meaning becomes explicit only when considering syntagmatic relationships the lexemes enter within a sentence and/or the wider context.
Not only polysemic terms but also occasionalisms and professionalisms can pose alignment problems. Such items of professional vocabulary are not fully lexicalized and, as a result, are not always recognized by CAT tools, for example, *tiger tail* (Civil Engineering), *ant colony optimization* (Telecommunications), or *Matilda bond* (Economics).
Precise interpretation of the meaning of polysemic terms is still beyond the competence of the computer. Besides formal knowledge, they should be able to take into account pragmatic aspects of the text environment, that is, social, field-specific, situational, cultural and individual contexts. Due to background knowledge, analytical skills and intuition human translators can interpret the meaning of words; establish relations and interconnections between them and external context.

4. Conclusions

The information structure of the contemporary popular scientific and technical text is characterized by the distinct hierarchical organization, growing information density, and the increased degree of intertextuality, i.e. interaction between the given and new information.
Much information communicated by popular scientific and technical texts is presented implicitly, thus sender's implicatures and presuppositions are not always adequately decoded even by human recipients, and are not fully accounted for by text processing and, subsequently, by information extraction tools.
Natural language is characterized by uncontrolled creative use of language resources resulting in the infinite number of meaning combinations. Therefore, information extraction from English popular scientific and technical texts is complicated due to the presence of terms based on metaphoric meaning extension, proper names based on metonymy, intra-disciplinary and cross-disciplinary polysemy, and culture-specific items. The challenges associated with decoding of meaning of foregrounded elements are most apparent when these elements should be communicated across the languages and recorded in multilingual databases.

References

[1] Dik, S.: The Theory of Functional Grammar. Part 1: The Structure of the Clause. Mouton de Gruyter, Berlin (1997), cited p.10
[2] Govindarajulu N. S., Bringsjord, S., Licato, J.: On Deep Computational Formalization of Natural Language. Presented at Formalizing Mechanisms for Artificial General Intelligence and Cognition, FORMAL MAGIC 2013. Beijing, China.
[3] Hobbs, J.: The Generic Information Extraction System. In: Proceedings of the 5th Message Understanding Conference (MUC-5) (1993).
[4] Kamp, H., Reyle, U.: From Discourse to Logic. Dordrecht, Kluwer (1993).
[5] Lambrecht, K.: Information Structure and Sentence Form: Topic, Focus, and the Mental Representation of Discourse Referents. CUP, UK (1994), cited p.5.
[6] Piskorski, J., Jangarber R.: Information Extraction: Past, Present and Future. In: Poibeau, T. Saggion, H., Piskorsi, J., Jangarber R. (eds.) Multi-source, Multilingual Information Extraction and Summarization, pp. 23-49. Springer-Verlag, Berlin (2013).
[7] Scarfe, R.T., Shortland, R.J.: Data mining applications in BT, Knowledge Discovery in Database, [IEE Colloquium on], pp.5/1—5/4 (1995).
[8] Schwabe, K., Winkler, S.: On Information Structure, Meaning and Form: Generalization Across Languages. John Benjamins Publishing, Amsterdam (2007), p.1.
[9] Shannon, C. E., Weaver, W.: The Mathematical Theory of Communication. Foreword by Richard E. Blahut and Bruce Hajek. University of Illinois Press, Urbana (reprinted in 1998).
[10] Steube, A.: Information Structure: Theoretical and Empirical Aspects. Walter de Gruyter, Berlin (2004), cited pp.15-16.
[11] Turmo, J., Ageno, A., Catala, N.: Adaptive Information Extraction. ACM Computing Surveys, Vol. 38, 2, Article 4 (2006), cited pp.1, 2, 12.