# Aspects of Summarization in the Discourse of Electrical Engineering in Contemporary English and French – Towards a Unitary Account

**Miloš D. Đurić**, *Senior Lecturer in English Language and Literature, Faculty of Electrical Engineering, University of Belgrade*

***Abstract* — In the past two decades or so, we have witnessed an increasingly growing interest in the study of discourse parsing and discourse summarization from both computational and linguistic perspectives. This paper reconsiders certain theoretical and empirical aspects of discourse summarization. By way of illustration it provides a unitary account of plausible aspects of automatic and/or manual summarization of the discourse of electrical engineering in contemporary English and French.**

***Keywords* — Discourse of Electrical Engineering, English, French, Parsing, Semantics, Summarization.**

## I. Introduction

According to [1] the basic aim of automation in the field of computational linguistics may be presenting a computer with a request for information on a certain specific topic and then being provided with the published material available. It seems to me that this assertion is still valid today as it was valid at the time of the citation. However, I should add that this "browsing" process would be enhanced by means of improving the possibilities of using automatic summarization tools.

This paper reconsiders the question of automatic summarization; taking as a starting point claims, found in the literature [2], that coherent texts possess internal structure, this structure being conveniently characterized by discourse and/or rhetorical relations. Semantic-syntactic interface might be viewed as input which linguistic competence hands over to cognitive performance. Briefly, the linguistic parser delivers linguistically determined semantic representations as output over which the inferential mechanisms perform certain computations to yield conceptual representations of the writer's/speaker's communicative intention. Sometimes computing meaning might be a time consuming activity, which might be, therefore, constrained by considerations of loss and gain.

Unfortunately, certain aspects pertaining to this area have remained unexplored. Therefore, taking all the above said into consideration one might be inclined to provide a unitary account of the phenomena pertaining to discourse summarization. This paper is an attempt at analyzing the linguistic contribution to automatic summarization.

But, before I proceed I should like to define the concept of *summary*. According to [3], a 'summary' may be loosely defined as "[…] a text that is produced from one or more texts that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that." It seems to me that the quoted reference uses the term *text* rather loosely. However, in my paper the terms 'text' and 'discourse' refer only to the discourse of electrical engineering.

As [4] noted, the lexical item 'summary' has been associated with a variety of meanings and has been used in a variety of contexts. Nevertheless, differences in classification were merely a symptom of a more serious problem that carried over into the subsequent summarization approaches, and which essentially evolved around two points of departure: first, the application of discourse extraction methods and fact extraction methods (or, rather template-filling methods), respectively, and, second, the organization of such methods. Thus, there are different types of summaries. As [4] highlighted, this variety of meanings resulted in a proliferation of different taxonomies, which were aimed at capturing the essence of summarization, and the notion of summary itself.

Bearing [4] classification in mind, we might suggest that four distinctions have emerged as fundamental in answering the question of how many types of summaries there actually are. One is a distinction about the way the input is handled – this is the distinction between single-document and multiple-document summaries (i.e. single-document/multiple-document axis). If we focus on the output, the following two oppositions appear: extract-like and abstract-like summaries (i.e. extract-like summary/abstract-like summary axis). These distinctions usually go hand in hand with the third distinction in terms of usage, according to which the opposition '*indicative summary* vs. *informative summary*' is proposed. And, finally, in terms of purpose, summaries can be *generic* and *query-oriented*.

Certain observations have been made concerning the idea that summarization process is primarily connected with cognitive and linguistic mechanisms which have nothing to do with the surface manifestation of the text. In

Author is with the Faculty of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (phone: +381-11-3281366; e-mail: djuric@etf.rs).

addition to this, generally speaking, many existing summarization tools are based on detecting the clear surface signals within the text, focusing particularly on cohesive devices. For that reason, summarization process is notoriously hard to pin down in purely linguistic or purely computational terms.

The question I would like to pose is the following: Is it plausible to generate a discourse-based summarizer? In order to provide an answer to this question, I should like to describe only some current state-of-the-art summarization tools. However, it should be noted that recent summarization approaches utilize fairly sophisticated techniques for deciding which linguistic units to extract. These sophisticated techniques often rely solely on machine learning. Nevertheless, it seems to me that this practice is not sufficient enough for a proper linguistic analysis. At the same time, this practice is linguistically vague and indefinite.

## II. Discourse Structure of English and French Electrical Engineering Discourses

I shall now confine my discussion to discourse structure of English and French electrical engineering discourse, using the term 'discourse' without any prescriptive intention. The focus is only on written discourse, although, probably investigating spoken discourse summarization might also represent an interesting challenge.

Strictly speaking, both English and French discourses of electrical engineering have very similar text structure (at least, the electronic corpus I analyzed). Namely, this specific type of discourse within my corpus almost always has an introduction, which is followed by details and illustrated by examples, facts, and then a secondary point is introduced. Broadly speaking, the introduction is often a good short summary of the electrical engineering discourse chunk itself, while for other sub-genres of this discourse this may not be the case. In other words, apart from fairly clear cases of electrical engineering discourse, there may also be found borderline cases pertaining to certain sub-genres.

While trying to establish a model-theoretic account of valid electrical engineering discourse structures in English and French, I shall use the terms 'multinuclear' and 'mononuclear', respectively. The former refers to paratactic relations, while the latter refers to hypotactic relations existing in my corpus. I have adopted Marcu's formalization according to which two adjacent spans may be joined in a larger span by a given rhetorical relation only if one of the following has been met: 1. Either a similar relation holds between at least two of the most pertinent units of these spans, or, 2. An extended rhetorical relation holds between the two adjacent spans.

In this sketchy analysis rhetorical relations in the electrical engineering discourse are treated as such relations existing between two non-overlapping discourse spans called *nucleus* (*N*) and *satellite* (*S*), respectively.

Let us now see an example of English discourse of electrical engineering:

[The defining postulates of both Einstein's theory of relativity and quantum theory are indisputably supported by rigorous and repeated empirical evidence. $A_1$] [However, while they do not directly contradict each other theoretically, $B_1$] [they are resistant to being incorporated within one cohesive model. $C_1$]

And now an example of French discourse of electrical engineering:

[Un autre grand succès de la mécanique quantique fut de résoudre le paradoxe de Gibbs: $D_2$] [en physique statistique classique, des particules identiques sont considérées comme étant discernables, $E_2$] [et l'entropie n'est alors pas une grandeur extensive. $F_2$]

In the course of summarization process, rather than analyze these spans of discourse ($A_1$, $B_1$, $C_1$, $D_2$, $E_2$, $F_2$) as a whole, I interpret them as rhetorical relations, which obtain between the pairs of discourse units. These rhetorical relations (*RR*) might be represented by means of certain sets, for example:

$RR_0 = Rhet\_rel(\text{JUSTIFICATION}_0, A_1, B_1)$
$RR_1 = Rhet\_rel(\text{JUSTIFICATION}_1, D_2, E_2)$
$RR_2 = Rhet\_rel(\text{EVIDENCE}, C_1, F_2)$

Strictly speaking, on the one hand, my rhetorical relations $A_1$ and $B_1$ provide some sort of context-dependent justification in English discourse of electrical engineering, while on the other hand, relations $D_2$ and $E_2$ provide context-dependent justification in French discourse of electrical engineering.

It goes without saying that the proposition expressed by the nucleus (*N*) of a rhetorical relation is more pertinent to the discoursal purpose than the satellite (*S*). This entails the fact that satellite is, in fact, impenetrable and opaque, if extracted without the nucleus. The remainder of this part of the paper is a sketchy descriptive exercise in the general structure of English and French electrical engineering discourses.

In what follows, rhetorical relations titles are labelled as unique identifiers in the form of subscripts ranging over the set of natural numbers (e.g. $_1, _2, _3 \ldots _n$). Let us now see the following examples:

English discourse: [This is not accomplished by introducing some new axiom to quantum mechanics, $A_1$] [but, on the contrary by *removing* the axiom of the collapse of the wave packet. $B_1$]

French discourse: [Le microcontrôleur ne dispose pas non plus de mémoire morte interne, $C_2$] [mais la communication avec l'EPROM se fait non pas par des broches d'entrée-sortie. $D_2$]

Taking into consideration that cohesive devices are relevant within summarization process on purely formal grounds, certain attention ought to be refocused to them. Both English and French cohesive devices (English *but* and French *mais* in my examples) occur in unit *i* of a given discourse. Echoing Marcu's framework, I adopt that these cohesive devices might signal rhetorical relations holding between one particular unit in the interval $[i - k, i - 1]$, while the other unit occurs in the interval $[i, i + k]$. Of course, *k* should be a sufficiently large constant. Similarly, cohesive devices, like English *but* and French *mais*, might

exhibit rhetorical relations between spans $[i - k_1, i - 1]$ and $[i, i + k_2]$.

From the perspective of the present topic, and adopting Marcu's model, the following linguistic features of electrical engineering discourse structure have been taken into account: 1. Elementary units of complex discourse structures are non-overlapping spans of discourse, 2. Discourse relations are valid between discourse units of various sizes, 3. Some discourse units play a more relevant role in discourse than the other, 4. The abstract structure of discourse might be represented as a tree.

Some of these features have been exhibited, if not explained. For instance, we have already seen that the elementary units in my examples have been delimited by square brackets. As regards discourse relations, they ought to be considered so as to provide an account dealing with meaning of the given discourse. I have also exemplified the fact that some discourse units are more pertinent and salient than the others, this being illustrated with the *paratactic* and *hypotactic* relations. With regard to the abstract representation, we might assert that trees are solid mathematical abstractions of discourse structure.

In addition to this, I have also taken into consideration rhetorical analysis, proposed by [5]. Contrary to Rhetorical Structure Theory, this approach might be interpreted as nonhierarchical, referring to discourse chunks at a lower level of granularity. However, my approach goes halfway between Marcu and Teufel/Moens approaches. Therefore, my model represents an eclectic fusion including both frameworks.

## III. Discourse Summarization

### A. Towards Measuring Discourse Structures – Relevance and Redundancy

In this section, I shall describe the discourse-based framework for electrical engineering discourse summarization. According to [6], "[a] more complex way to integrate discourse, cohesion, position, and other summarization-based methods I to consider that the structure of discourse is the most important factor in determining saliency […]". To put it simply, this approach does not treat discourse as flat sequences of discourse units, but rather as tree structures reflecting the nuclearity and rhetorical relations, which are typical of discourse spans.

While adopting [5] model, I have noticed that sometimes the vague definition of relevance cannot be replaced with a more operational rhetorical status in a straightforward manner. However, in no way does the imminent analysis rule out the possibility of sentences describing research goal or stating a difference with a rival scientific approach in the discourse of electrical engineering being used as relevance criteria. Nor does it aim at any prescriptive correction. Instead, this analysis would like to offer a discourse model in which each chunk of discourse receives both a rhetorical category label and also tag, implying either relevance or irrelevance. It would be interesting to see what further analytical results the tagging process would produce; nevertheless, it lies outside the scope of the present paper.

### B. Electronic Corpus

My electronic corpus consists of 88 scientific papers converted into an electronic form, so that they might be processed by means of a computer. It contains 16.888 sentences. All equations, tables, figures and illustrations have been removed, because they constitute the redundant part of this type of discourse (of course, redundant only from the point of view of the summarization process). Unfortunately, I have not added Extensible Markup Language to all chunks within my corpus, although it would be productive to see how titles, abstracts, headlines, and paragraphs behave after being marked up.

While treating this corpus, both manually and computationally, I have taken as axiomatic that discourse might be partitioned into a sequence of non-overlapping, elementary discoursal units.

### C. Manually-Generated Summaries

Manually-generated summaries were created by only four native speakers (i.e. one American, two British, and one French), and therefore are perhaps not representative enough to account for some specific problems pertaining to human summarization processes. However, this analysis has been provided also with other non-native annotators' contribution, these annotators being my students who also generated some of the summaries, but who are non-native speakers.

Before I provided these human annotators with actual discourse, I first determined manually the minimal discourse chunks of each discourse. These discourse chunks were enclosed within square brackets and labelled in increasing order with a natural number from 1 to N.

Some of the human annotators thought that important sentences are very frequently located either at the beginning or the end of paragraphs in the discourse of electrical engineering. Guided by this presumption, human annotators took these discourse-initial and discourse-final sentences as the most salient ones, and included them into their manually-generated summaries by inertia.

In a nutshell, the summary which annotators delimited in this way has been selected from the discourse source representation, the original representation being introduced by full sentences (such as the English sentence: [However, electrons normally remain in an unknown orbital path around the nucleus, defying classical electromagnetism. $EED_1$], or the French sentence: [Cependant, la dérivation de cette inégalité énergie-temps est assez différente de celle des inégalités position-impulsion. $EED_2$]). Whether this discourse representation is directly or indirectly linked to the source representation ought to be contextually determined by human annotators. Also, I have noticed that there is a strong correlation between the electrical engineering discourse structural pattern and the chunks which human annotators perceived as being pertinent. This seems to be hardly surprising bearing in mind that human annotators were either electrical engineers (4 native

speakers) or future electrical engineers (students).

Due to spatial limitations imposed upon the author, and also taking into account that this paper should describe the area of automatic summarization, I shall not include the examples of human annotators' summaries, but shall rather offer some examples of automatic summarization output in the part which follows.

### D. Examples of Automatic Summarization

I have compared the performance of discourse-based summarizer with that pertaining to a commercial summarizer, or more precisely, the one incorporated into Microsoft Office 2007 package. Then, I ran the summarization program on the texts from my electronic corpus.

Let us now see some of the examples from my corpus which have undergone automatic summarization, and then have been manually annotated and broken into discourse relations:

English: [The only difference is operation of the rectifier in the discontinuous conduction mode, $A_1$] [which cannot be utilized in the case of the switching current injection device, $B_1$] [due to high values of the input current THD. $C_1$] [Unfortunately, this significantly reduces applicability of the voltage-loaded resistance emulator in this case, $D_1$] [but all other methods are directly applicable. $E_1$]

French: [Des simulateurs sont disponibles pour certains microcontrôleurs, $F_2$] [comme l'environnement MPLAB de Microchip. $G_2$] [Les développeurs peuvent AINSI analyser le comportement du microcontrôleur et du programme, $H_2$] [comme s'il s'agissait du composant réel. $I_2$]

The neat examples from my corpus so far cited in this section are relatively straightforward cases of "direct" summarization, but they might obscure the picture, as there are lots of cases which do not easily lend themselves to such a clear-cut division, and, therefore, represent a potential problem for automatic summarizers.

### E. Towards Mathematics of Discourse Structures

I have utilized the following formalizations, found in the literature, which states: 1. Predicate *position*($u_i$, $j$) is true for a discourse unit $u_i$ in a sequence $U$ if and only if $u_i$ is the $j$-th element in the sequence. 2. Predicate *rhet_rel*(name, $u_i$, $u_j$) is true for discourse units $u_i$ and $u_j$ with respect to rhetorical relation *name* if and only if the definition $D$ of the rhetorical relation name is consistent with the relation between discourse units $u_i$ and $u_j$, the former being a satellite ($S$), and the latter being a nucleus ($N$).

In other words, a set of predicates may be represented as follows:

*rhet_rel*(JUSTIFICATION$_0$, $A_1$, $B_1$)
*rhet_rel*(JUSTIFICATION$_1$, $D_1$, $B_1$)
*rhet_rel*(RESTATEMENT, $E_1$, $B_1$)
position($A_1$, 1), position($B_1$, 2)
position($C_1$, 3), position($D_1$, 4)

Non-binary trees have been treated as collapsed version of binary trees. In addition to this, the status of the root, such as {N, S} indicates that discourse span [$A_1$, $I_1$] might only take over NUCLEUS or SATELLITE value in a given discourse span.

A sketchy answer I am suggesting at this point might lie in Marcu's intent to determine from the set of $n + (n - 1) + (n - 2) + \ldots + 1 = n(n + 1)/2$ potential discourse spans, bearing in mind that these spans belong to a sequence of $n$ discoursal units. In the following Table 1, I enumerate only some of indicator types that I spotted in my corpus:

TABLE 1: FREQUENCY OF INDICATOR TYPES.

| ***Indicator type*** | ***Number*** |
|---|---|
| SIMILARITY | 128 |
| CONTRAST | 46 |
| DEIXIS | 56 |
| **Total of 3 classes** | 230 |

## IV. CONCLUSION

In this paper I have reexamined certain aspects of summarization from a novel perspective connected with English and French discourses of electrical engineering. As far as practical and statistical implications are concerned, my results are still too far from perfect. Evident and obvious ways of improving summarization process performance are: 1. The use of more sophisticated classifiers, and, 2. Incorporating more training material to both human annotators and automatic summarization tools.

Apart from the notorious lack of cross-contrastive summarization studies, certain important aspects remain underexplored, perhaps one of them being human/computational interface. This paper aims at promoting an interest in the study of automatic summarization processes in English and French discourses of electrical engineering, which might significantly contribute to: 1. Discourse structure formalization, 2. General syntactic parsing, 3. Natural Language Generation (NLG), 4. Machine translation, and probably 5. Information extraction and retrieval. My assumptions, however, merit further elaboration.

## REFERENCES


[1] M. F. Bott, "Computational linguistics," in *New horizons in linguistics*, J. Lyons, Ed. Penguin: Harmondsworth, 1971, pp. 215–228.

[2] D. Marcu, *The theory and practice of discourse parsing and summarization*, Cambridge, Massachusetts: The MIT Press, 2000, ch. 1.

[3] D. R. Radev, E. Hovy and K. McKeown, "Introduction to the special issue on summarization," *Computational linguistics*, 2002 28(4), p. 399.

[4] D. Marcu, *The theory and practice of discourse parsing and summarization*, Cambridge, Massachusetts: The MIT Press, 2000, p. 187.

[5] S. Teufel and M. Moens "Summarizing scientific articles: experiments with relevance and rhetorical status," *Computational Linguistics*, 2002 28(4), p. 416.

[6] D. Marcu, *The theory and practice of discourse parsing and summarization*, Cambridge, Massachusetts: The MIT Press, 2000, p. 204.