Detecting Information-Dense Texts: Towards an Automated Analysis

Danguolė Kalinauskaitė
Department of Lithuanian Studies, Vytautas Magnus University, Kaunas, Lithuania
Baltic Institute of Advanced Technology, Vilnius, Lithuania
danguole.kalinauskaite@bpti.lt

***Abstract*—Determining information density has become a central issue in natural language processing. While information density is seen as too complex to measure globally, a study of lexical and syntactic features allows a comparison of information density between different texts or different text genres. This paper provides a part of methodology proposed for automatic analysis of information density based on lexical and syntactic levels of language.**

Keywords—lexical density, information density, natural language processing, computational linguistics

I. INTRODUCTION

Determining information density of text is a big challenge in natural language processing (NLP). The use of more fragments of text to train statistical NLP systems may not necessarily lead to improved performance. Recent developments in this field have spawned a number of solutions to evaluate information density. Nevertheless, a shortfall of most of these solutions is their dependency on the genre and domain of the text. In addition, most of them are not efficient regardless of the NLP problem areas [1].

It is worth to note that the notion of information is only formal here, i.e. information is defined as semantic, pragmatic, and only measurable in relative terms. A definition of information density is elaborated involving informativity (a relative measure of semantic and pragmatic information) per clause (following [2]). So in terms of semantics, information density is a measure of the extent to which the writer or speaker is making assertions (or asking questions) rather than just referring to entities [3]. Texts contain various elements ranging from characters to sentences, that are supposed to have reasonable discriminating strength in evaluating information density in natural language text. Examples of such elements are the use of simple words, complex words, function words, content words, syllables, and so on.

The paper starts with a theoretical background on the measurement of information density, followed by a presentation of the research, and ends with a conclusion and future work plans.

The goal of this paper is to present a part of methodology proposed for automatic analysis of information density based on lexical and syntactic levels of language.

II. THEORETICAL BACKGROUND

A. Information Density in Computational Linguistics

Information-dense texts report important factual information in direct, succinct manner. There were various attempts to determine and evaluate information density of texts. Earlier works did it manually. Later various programs began to appear, and now this process can be done automatically. However, all programs are different in nature, as well as in their productivity and principles based on which information density of texts is determined. Worth mentioning issue here is that there are a lot of confusion in determining what are the indicators of information-dense texts, and therefore there is no unified methodology for measuring information density.

In computational linguistics, one of the most common characteristics employed to detect information-dense texts is lexical density (see *B. Lexical Density* below). In some works it is even suggested as the main indication of how informative a text is, and used as a synonym for information density. However, lexical density, i.e. only one level of text that basically points to the vocabulary, is not sufficient to talk about the whole text informativeness. Therefore it does not seem convincing to link these two terms, it is more likely that one is a part of another, as lexical density measures only one level of texts, namely, vocabulary, and the whole text informativeness depends not only on the content but also on the structure.

It is worth to note that the applicability of research results in this area is very extensive. Numerous psychological experiments have related information density to readability [4], [5], memory, e.g., [6], quality of students' writing, e.g., [7], aging [8], [9], and prediction of Alzheimer's disease [10], [11], [12].

High information density signals complex interrelationships expressed. Low information density means relatively little information per sentence, therefore low information density in speech or writing can indicate mental disorders, including Alzheimer's disease.

B. Lexical Density

Lexical density is the term most often used for describing the proportion of content words to the total number of words [13], [14]. The result is a percentage for each text in the corpus. Content words give a text its meaning and provide information regarding what the text is about. More precisely, content words are simply nouns, verbs, adjectives, and adverbs. Nouns tell us the subject, adjectives tell us more about the subject, verbs tell us what they do, and adverbs tell us how they do it [13].

Other kinds of words such as articles (*a, the*), prepositions (*on, at, in*), conjunctions (*and, or, but*) and so forth, are more grammatical in nature and, by themselves, give little or no information about what a text is about [15]. These non-lexical words are also called function words. Auxiliary verbs, such as *to be (am, are, is, was, were, being), do (did, does, doing), have (had, has, having)* and so forth, are also considered non-lexical as they do not provide additional meaning.

It is worth first to determine the lexical density of an ideal example:

(1) *The **quick brown fox jumped swiftly** over the **lazy dog**.*

The lexical words (nouns, adjectives, verbs, and adverbs) are **bold**.
There are precisely 7 lexical words out of 10 total words. The lexical density of the above passage is therefore 70%.

Another simple example:

(2) *She **told** him that she **loved** him.*

The lexical density of the above sentence is 2 lexical words out of 7 total words, for a lexical density of 28.57%.

The meaning of the first sentence is quite clear. It is not difficult to imagine what happened when "the quick brown fox jumped swiftly over the lazy dog". On the other hand, it is not so easy to imagine what the second sentence means - due to the use of vague personal pronouns (*she* and *him*), this sentence has multiple interpretations and is, therefore, quite vague.

Lexical density is a reflection of the above observations. The sentence (1) has a rather high lexical density (70%), whereas, the sentence (2) has a lexical density which is quite low (28.57%).

The reason that the sentence (1) has a high lexical density is that it explicitly names both the subject (fox) and the object (dog), gives us more information about each one (the fox being quick and brown, and the dog being lazy), and tells us how the subject performed the action of jumping (swiftly).

The reason that the sentence (2) has such low lexical density is that it doesn't do any of the things that the first sentence does: we don't know who the subject (she) and the object (him) really are; we don't know how *she* told *him* or how *she* loves *him*; we don't even know if the first *she* and *him* mean the same people as the second *she* and *him*. This sentence tells us almost nothing, and its low lexical density is an indicator of that, contrary to the first sentence which is packed with information and its high lexical density is a reflection of that.

However the information lies here only on the lexical level of text. The lexical level is related with syntactic one, i.e. how words behave within a text, how they are connected with each other in a sentence, etc. Finally the vocabulary and the form of text highly depend on the genre of text.

III. RESEARCH

The research was conducted to investigate which features of text mostly characterize information-dense texts. Lexical density mentioned above is one of them here, however the analysis was performed on the basis of the form of texts, too.

Lexical density has the advantage of being easy to operationalise, and also practical to apply in computer analyses of large data corpora.

The research sought to compare journal abstracts and their research papers from the point of view of their linguistic features and specificity of the genre, and in this way identify textual features of abstracts based on their similarities and differences with regard to research papers. It was raised a hypothesis that abstracts are characterized by a higher information density than their research papers. The comparison was performed on the basis of two corpora, compiled from the research papers and their abstracts in the journal of "Pragmatics"[1], from the period of 2000-2017. They have been collected specifically for the purposes of this research. Both corpora will be available in CLARIN-LT Repository[2].

The research consisted of qualitative and quantitative analysis, and in this way the contents and the form of the corpus of abstracts (containing 85 616 running words), and the corpus of research papers (containing 3 479 442 running words) were analysed with the help of corpus management and analyses tool - Sketch Engine[3], and WordSmith Tools version 6[4]. The focus of research was on the abstracts, and full length papers were compared with their abstracts.

A. Qualitative Analysis

The following are the components of qualitative analysis:

- keywords for each corpus (frequency lists normalized for 1000 text words);
- terms for each corpus;
- contents of each corpus by parts of speech: the proportion of content words; the proportion of functional words.

Both keyword and term lists revealed more similarities than differences between abstracts and research papers, therefore further analysis of the most frequent terms from both corpora together was used to show the overall dynamics of topics of the journal over time.

[1] https://benjamins.com/#catalog/journals/prag/main.
[2] https://clarin.vdu.lt/xmlui.
[3] https://www.sketchengine.co.uk/.
[4] http://www.lexically.net/wordsmith/version6/.

The top 10 notional words from the keyword lists for each corpus (see Figure 1) were the basis for the analyses of their context, i.e. grammatical constructions and lexical collocations.

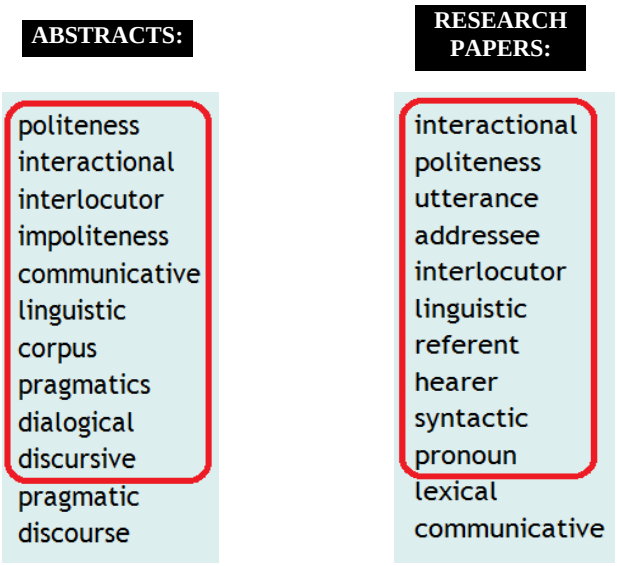

Fig. 1. The top keywords from each corpus

In this way the formal features of research papers and their abstracts were observed: such contextual analyses revealed linguistic ways to condense information in the abstracts that were absent in their full length counterparts. One of them is nominalisation (the use of nominal phrases instead of verbal phrases) allowing to merge a few sentences into one. Nominalisation, in turn, is associated with higher lexical density in abstracts than in their research papers (see Figure 2), i.e. it decreases the number of functional words and in this way increases lexical density in general.

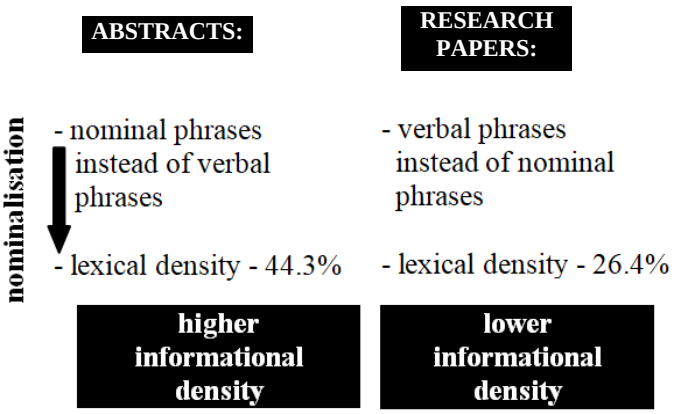

Fig. 2. The relation between formal and content features of texts

B. Quantitative Analysis

The following are the components of quantitative analysis:

- overall statistics (see Table 1 summary);
- lexical density: the proportion of content words to the total number of words (the corpus of abstracts and the corpus of research papers separately).

TABLE I. OVERALL STATISTICS OF CORPORA

	Abstracts	Research papers
Tokens (running words)	85 616	3 479 442
Types (distinct words)	8 295	71 038
Type/token ratio (TTR)	9.90	2.13
Standardized TTR	36.50	38.99
Standardized TTR std.dev.	58.35	60.85
Sentences	3 305	116 735
Average sentence length (words)	25.37	28.62

IV. CONCLUDING REMARKS AND FUTURE WORK

The qualitative analysis showed that contents are similar in case of abstracts and their research papers.

Quantitative analysis revealed that abstracts and their research papers are more different than similar in terms of formal features: formal features of both corpora manifested tangible differences in abstracts. Thus the research proposed that lexical density depends strongly on the form of texts.

Lexical density is useful and applicable measurement for different text genres, however, lexical level alone is not sufficient to measure information density of texts, while lexical and syntactic features together appear to be particularly well suited for the task.

With the above in mind, future work is to develop the methodology for measuring information density by analysing syntactic level of texts, and later - combining both lexical and syntactic features for implementing the results into the automatization of text analysis.

REFERENCES

[1] R. Shams, “Identification of informativeness in text using natural language stylometry”, Doctoral thesis, The University of Western Ontario, 2014.

[2] C. R. Mills, “Information density in French and Dagara folktales: a corpus-based analysis of linguistic marking and cognitive processing”, Doctoral thesis, Queen’s University Belfast, 2014.

[3] C. Brown, T. Snodgrass, S. J. Kemper, R. Herman, M. A. Covington, “Automatic measurement of propositional idea density from part-of-speech tagging”, Behavior Research Methods 40(2), pp. 540-545, 2008.

[4] W. Kintsch, J. Keenan, “Reading rate and retention as a function of the number of propositions in the base structure of sentences”, Cognitive Psychology 5, pp. 257-274, 1973.

[5] W. Kintsch, “Comprehension: A paradigm for cognition”, Cambridge, UK: Cambridge University Press, 1998.

[6] E. Thorson, R. Snyder, “Viewer recall of television commercials: Prediction from the propositional structure of commercial scripts”, Journal of Marketing Research 21, pp. 127-136, 1984.

[7] A. Y. Takao, W. A. Prothero, G. J. Kelly, “Applying argumentation analysis to assess the quality of university oceanography students’ scientific writing”, Journal of Geoscience Education 50, pp. 40-48, 2002.

[8] S. Kemper, J. Marquis, M. Thompson, “Longitudinal change in language production: Effect of aging and dementia on grammatical complexity and propositional content”, Psychology and Aging 16, pp. 600-614, 2001.

[9] S. Kemper, A. Sumner, “The structure of verbal abilities in young and older adults”, Psychology and Aging 16, pp. 312-322, 2001.

[10] D. A. Snowdon, S. J. Kemper, J. A. Mortimer, L. H. Greiner, D. R. Wekstein, W. R. Markesbery, “Linguistic ability in early life and cognitive function and Alzheimer’s disease in late life: Findings from the Nun Study”, JAMA 275, pp. 528-532, 1996.

[11] D. A. Snowdon, L. H. Greiner, S. J. Kemper, N. Nanayakkara, J. A. Mortimer, “Linguistic ability in early life and longevity: Findings from the Nun Study”, Berlin, Germany: Springer-Verlag, 1999.

[12] D. A. Snowdon, L. H. Greiner, W. R. Markesbery, “Linguistic ability in early life and the neuropathology of Alzheimer’s disease and cerebrovascular disease: Findings from the Nun Study”, Annals of the New York Academy of Sciences 903, pp. 34-38, 2000.

[13] D. Didau, “Black space: improving writing by increasing lexical density”, The Learning Spy: Brain Food for the Thinking Teacher, 2013.

[14] V. Johansson, “Lexical diversity and lexical density in speech and writing: a developmental perspective”, Working Papers 53, pp. 61-79, 2008.

[15] J. Ure, “Lexical density and register differentiation”. In G. E. Perren & J. L. M. Trim (eds.). Applications of linguistics. Selected papers of the Second International Congress of Applied Linguistics, Cambridge 1969, pp. 443-452. Cambridge: Cambridge University Press, 1971.