Complexity Class of Semantics-related Tasks of Text Processing

Oleg Bisikalo[1], Ilona Bogach[2]

Vinnytsia National Technical University, 95 Khmelnytske shose, Vinnytsia 21021, Ukraine
[1] obisikalo@gmail.com
[2] ilona.bogach@gmail.com

Abstract. Formal features of semantics-related tasks of text processing are reviewed; NP-complete procedural complexity of the class is substantiated. To diminish procedural complexity, the rationale behind applied formal linguistic knowledge is demonstrated, based on the analogy of the knapsack problem and automated text abstracting. Versatile approach for text processing is proposed, considering relations between entities; informational estimates are obtained and recommendations set forward.

Keywords. procedural complexity, *NP*-completeness, semantics-related tasks, text processing, informational estimate.

Key Terms. Text processing, computational linguistic, synthesis of natural language information.

1 Introduction

Some of the most complicated tasks in computational linguistics are those associated with semantics parsing and synthesis of natural language information. According to the authors, taking into account certain common formal features, such tasks are to be separately classified as semantics-related tasks of test processing, viz. text annotation and abstracting, searching key words, dialogue support etc. Sharper focus on this class and respective scientific research papers are driven by escalating demand for linguistic Internet technologies throughout the world.

Semantics-related tasks of text annotation and abstracting can save time for the experts, provided there is a proper quality of solutions. Summary is a coherent text, concisely depicting core topic as well as objectives, methods and findings of the research or insight, unlike annotation, which is a brief description of the content and general information on the topic. While the main purpose of annotation is to draw attention the text, the summary, containing just 10-23% of the text, allows users to arrive at conclusions as accurately as from the text, having spent twice as less time [1].

ICTERI 2016, Kyiv, Ukraine, June 21-24, 2016

The main challenges of the above-mentioned tasks are caused by polysemy of natural language, the topical issue in computational linguistics. It is important to obtain the assessment of complexity of different approaches to solutions of semantics-related tasks, from direct enumeration to such heuristic methods, which enable people to understand the sense of new text information quickly. This will identify the rationale and efficiency of additional procedures of linguistic text analysis, singling out the so-called stop-words, attracting expertise etc. Thus, crucial task is to identify formal properties, including informational and general assessment of the complexity of the class of semantics-related tasks of text processing.

The objective of the research is to assess procedural class complexity of semantics-related tasks and determine efficient approaches to solutions.

2 Analysis of subject domains

To formally include semantics-related tasks in separate class, the general notion of class complexity should be considered. In the theory of algorithms complexity class is a set of computational tasks, approximately similar in terms of computing complexity. Otherwise, complexity class is a predicate set (function, having a word at the entry and coming back with a result 0 or 1), which is used for computing approximately similar number of resources [2].

There is a category of "the most complicated" for each category of tasks. It means that any task from the class goes down to such one, and the task belongs to the class. Such tasks are called complete for the class. NP-complete tasks are the most common.

Usually complexity class is determined by predicate sets with certain properties. Common determination of the class is as follows: complexity class X is called predicate set $P(x)$, computed by Turing machines, using resource computation $O(f(n))$, wherein n is the length of word x.

In most cases computation time is selected for the resource (number of tact in Turing machine) or operation area. Languages, which are identified by predicates from any class (i.e. sets of words, for which predicate turns 1), are also called those that belong to the same class.

Class P (Engl. *Polynomial*) is a set of tasks, providing relatively quick algorithms of solutions. Class P is included in broader classes of algorithms complexity.

Examples of P class are integral addition, division, matrix multiplication, determination of graph connectivity, ranging of sets from n numbers.

Non-deterministic polynomial task is a set of recognition problems, where solutions can be promptly checked at Turing machine, providing certain additional data solutions certificates. Equivalently *NP class* can be identified as a class, which contains tasks, admitting polynomial time of solution at non-deterministic Turing machine. There are examples of tasks, which are currently either classified or not classified as P, but belonging to *NP*:

− Tasks with Boolean formulae – find out with the *Boolean formula*, if there is a set of input variable, which turns 1. Certificate is such a set.

− Tasks on complete subgraphs – according to graph data, find out, whether it contains complete subgraphs of specified size. Certificate is a number of vertex,making complete subgraph.

− Find out the availability of Hamilton cycle in graph. Certificate is a sequence of vertex, making Hamilton cycle.

NP-complete problems are the most complicated among class *NP*. If anyone could cope with any of them for polynomial time, all tasks of *NP class would be solved for polynomial time*. Some examples of *NP-complete problems* are travelling- salesman task problem, Steiner problem, independent set problem, games Sapper, Tetris, Knapsack problems etc. For the time being, all those problems require exponential algorithms of solution.

To assess complexity of semantics-related tasks of text processing, proposed to be included, significant specific criteria of the results on understanding text information should be taken into account. Therefore, let's consider the issue of polysemy of the words in natural language from the formal view on word meanings. Thesauri usually provide all possible meanings of each word form with respective lexeme sign, which combines a certain set of words. The same spelling of words, belonging to different word forms, is a driver of escalation of scope of searching in the process of determination of proper meaning (polysemantic) of the word in each sentence of the text. Formally for r_i lexeme signs in i-sentence of the selected text the general scope of search equals to all possible options of meanings $(k)^{r_i}$, with the only one correct according to the author (k – average polysemy coefficient of certain language).

Linguistic research substantiated the following hypothesis: the higher the level of analyticity in the language, the more frequently the same lexeme sign is used for different functions, and the larger is average polysemy coefficient. For example, Spanish language is more analytical than German, its polysemy coefficient makes the value of 6,9 of per lexeme, and for German – being less analytical language, polysemy coefficient is 5,6 per lexeme [3]. Average polysemy coefficient considerably varies for different parts of speech for most synthetic Slavic languages. For example, for nouns – 4,32 meanings per lexeme, for adjectives: 5– for specific and 3,5 for abstract ones; as for Russian language, average polysemy coefficient makes 3,1 meaning per lexeme [4]. Thus, it can be inferred that the lower limit of general scope of V search for the text is no less than

$$V \geq \sum_{i=1}^{m} 3^{r_i}, \qquad (1)$$

where, m is a number of sentences in the text.

Apart from the degree of language analyticity, character and subject domain of the text can affect average polysemy coefficient. The latter is reduced by terminological steadiness of certain subject domain and austere (scientific) writing style, and increase by a number of adverbs, metaphors, elements of the so-called Aesopian language etc. Anyway, it is clear that problems of text understanding are formally referred to *NP*-complete complexity due to a step function (1). Moreover, it is not difficult for people to understand familiar language, including unknown text, which testifies for natural mechanisms of effective selection of the most proper combinations of meanings of all lexemes, contrary to complete search of all possible meanings.

We also considered common approaches for semantic analysis of text information, which differentiate the notions of lexical functions and semantic ones. In terms of semantics of the separate sentence linguists revealed 40÷60 (depending on the

language) of lexical functions, which mostly connect separate pairs of words or collocations. Accurate differentiation of all possible cases means the following: complexity by number of pairs at least from r_i to 2 with the coefficient of 40, i.e. $V' \geq 40 \cdot \sum_{i=1}^{m} \frac{r_i!}{2! \cdot (r_i - 2)!}$. The next step of formal adumbration of the sentence content is the notion of semantic relation (scheme), e.g. in [5] aggregation of 21 relations in 6 types, assigned by 9 triadic (quadruple)- predicates. Complexity of such approach is proportional to a number of allocations from r_i to 3 with the coefficient of 9, viz. $V'' \geq 9 \cdot \sum_{i=1}^{m} \frac{r_i!}{(r_i - 3)!}$. It should be noted that nearly all people have never heard of lexical functions and semantic relations, or have never thought of them, but it has not prevented them from understanding their language.

Thus, the rationale behind separating the class of semantics-related tasks is as follows: on the one hand, it is characterized by *NP*-complete complexity, wherein $V'' \geq V' \geq V$, and, on the other hand, there is an objective existence of natural algorithms of thinking that enable to solve the tasks of the class efficiently.

3 Automated abstracting as an example of semantics-related tasks of text processing

Preliminary analysis is the ground to ascertain that certain semantics-related problems are not only classified as *NP*-complete by procedural complexity, but are also similar to them by the formulation. Let us consider the afore-mentioned Knapsack problem as a proof, demonstrating convenient analogy for comparison and estimation of procedural complexity of automated text abstracting tasks [6]. Generally, tasks can be formed as follows: we need to select a certain number of objects from assigned set of objects with properties value and weight so that we obtain a maximal aggregate value along with the limit for the aggregate weight.

Without taking into account additional information by parameter analogues “value” and “weight” in Knapsack problems, it is obvious that there are parameters “importance” and “size” of fragments in automated abstracting tasks. Thus, in general case, abstracting is to result in a minimal scope of text, provided that it contains the most important phrases (sentences), whereat the text is supposed to keep the essence, and the last additional requirement makes the tasks of automated abstracting even more complicated. We assume that by the analogy described above, the task of automated abstracting is related to *NP*-complete problems.

As we know, classifying certain computation problems as *NP*-complete brings finding approximate algorithms [7] to focus of the scientists, since the unavailability of polynomial solutions makes the scientific paper futile. The problem of combinatory optimization of knapsack packing is a classical example of unsatisfactory time for solution by precise methods of full enumeration (for the sake of increasing necessary memory), dynamic programming or branches and limits. It shifts the focus to obtaining approximate results by greedy algorithm, genetic algorithms or other methods of discreet optimization. Unlike its analogy, approaches in solutions of

automated abstracting tasks have been historically construed as approximate methods [8], which considered additional linguistic information depending on the specifics of the task, e.g. TRM – Text Relation Map.

Classical *TRM* method takes into account weighted word vectors, corresponding to fragments (sentences) of the selected document, wherein graph is used as a formal model of semantic relations between structural units of text. Graph vertexes are text fragments, edges connecting the vertexes with a high level of approximation (semantic relation). Identifying key text fragments (vertex of the graph) for abstracting is based on criteria of a number of semantic relations of some fragments with others (ribs, coming out from vertex of the graph). It is proposed to combine TRM method with statistics methods TFIDF and *TLTF* in different options to additionally identify the weight of separate words of the document [9].

Estimation of procedural complexity of traditional *TRM method*. Wherein n is a number of words in the text, and m is a number of fragments (e.g., sentences). Generally thinking, we assume that there is an equal number of words in each sentence making $n' = n/m \approx r_i$. Then one operation of finding scalar outcome of two vectors with dimension n' (for 2-x sentences) requires computation

$$k_1 = 2\frac{n}{m} - 1. \tag{2}$$

Since general number of fragments is *m*, the number of operation of scalar outcome of its vectors equals

$$k_2 = \sum_{j=1}^{m} j. \tag{3}$$

Sum of terms of arithmetic progression (3):

$$k_2 = \sum_{j=1}^{m} j = \frac{m(m+1)}{2}. \tag{4}$$

On the assumption of (2) and (4), general number of computation for identification of measures of semantic similarity of text fragments by TRM method equals to

$$K_2 = k_1 k_2 = (2\frac{n}{m} - 1)\frac{m(m+1)}{2}. \tag{5}$$

Thus [10], limiting the estimate of procedural complexity by TRM method *O(nm)* does not exceed the complexity of 2-classed polynom for the number of words in the text *n* and proves the efficiency of applying procedures of linguistic analysis. Though, we should admit that the best results of automated abstracting cede in authorship or expert options.

So, the following is effective for typical semantics-related tasks of automated abstracting: а) consideration of relevant linguistic properties and parameters of separate words, text or selection of texts; б) identifying the most informative metrics for estimation of abstracting quality taking into account peculiarities of the text.

4 Informational estimate of the approach to text abstracting with relations between entities taken into account

Information flow analysis can be deemed as an alternative method for estimating complexity of semantics-related tasks. As opposed to procedural complexity, which is identified as a general estimate without specification, and considering peculiarities of the method of solution, informational estimate is procedure-oriented. Therefore, we propose to review informational estimate of the universal approach to text processing with relations between its entities (lexemes) taken into account.

Key feature of semantics-related tasks is deemed to be in determination and processing of content entities of the text. From informational view, understanding the sense of the sentence by a person is accompanied by recognizing separate words of the sentence and relations between pairs of the words with respective construction of the relations tree [11]. It should be recognized that in general using calculus of probability and in particular the notion of entropy in building NLP system goes back to the works of academician Markov regarding mathematical analysis of literary texts [12] and Claude Shannon regarding information value of English alphabetical symbols [13]. Though, such works focus on determination of the probability of correct string of symbols on the level of one word or several consecutive words.

Thus, famous work [14] covers maximum-likelihood approach for automatically constructing maximum entropy models and describes how to implement this approach efficiently, using as examples several problems in natural language processing. Partial results for constructing context-dependent models are obtained, viz. for segmentation of sentences and optimization of other parameters of machine translation. The following is proposed in the work [15]: multilayer neural network architecture that can handle a number of NLP tasks with both speed and accuracy by entropy- based criteria. Proposed are unified neural network architecture and learning algorithm that can be applied to various natural language processing tasks including: part-of-speech tagging, chunking, named entity recognition, and semantic role labeling. Obtained results allow automating the processes of useful markers to the text, though they do not take into account the level of general understanding of the text, which can be achieved by people.

The difference of our approach is that we consider all understanding processes to be carried out by comparative analysis and attracting information from the general linguistic base of knowledge of the subject. If each of those stages is accompanied by increasing of information, the following will be hypothetical in the universal approach:

- Ration of general understanding of text T can vary from minimal to maximal depending on the scope and other parameters of general linguistic base of individual's knowledge;
- Quality of determination of contents entities is proportional to the level of general understanding of text, to be confirmed by formal properties.

Informational estimate of this information is as follows:

1. Hereby we determine the entropy scope of one word in the text, for the case when appearance of the word is an independent and accidental event x with l of the following possible states

$$H(x) = -\sum_{j=1}^{l} p(x_j) \cdot \log p(x_j),$$

And maximal average estimate is made for the equally likely case

$$H_w = \log_2 l \ \ [Bit].$$

A number of different words (lexemes) of the text T can be considered variable l, and it is obvious that $l \le n$.

2. Let's determine maximal estimate of entropy scope of all words of the sentence, provided that appearance of next word with n' words of this sentence does not depend on the previous one

$$H(x) = -n' \cdot \sum_{j=1}^{l} p(x_j) \cdot \log p(x_j),$$

or for equally likely case

$$H_{sw} = n' \cdot \log_2 l \ \ [Bit]. \qquad (6)$$

3. Let's determine entropy scope of paired association, provided that the words of the sentence in the form of a certain set $X = \{x_1,\ldots,x_{n'}\}$ are known and recognized by the individual. For the pair to appear independently as an accidental event, potential number of pairs *y* can be $n' \times (n'-1) = (n')^2 - n'$, where sentence with n' words makes parsing tree from n' pairs, taking into account bilateral relation of subject-predicate. On the other hand, key diagonal of such matrix is excluded, since the word in the sentence cannot be connected with itself. Thus

$$H(y) = -\sum_{i=1}^{(n')^2 - n'} p(y_i \mid X) \cdot \log p(y_i \mid X),$$

Accordingly, we get the following for equally likely cases

$$H_p \approx \log_2 (n')^2 = 2 \cdot \log_2 n' \ \ [Bit].$$

4. Entropy scope of all pairs of separate sentences can be determined by combinational properties of tree formation from n' pairs, which are selected from $n' \times (n'-1)$ possible. In case of the most rigid condition of independent combination of words into n' pairs we have the following

$$H(y) = -n' \cdot \sum_{i=1}^{(n')^2 - n'} p(y_i \mid X) \cdot \log p(y_i \mid X),$$

And for equally likely case

$$H_{sp} \approx n' \cdot \log_2 (n')^2 = 2n' \cdot \log_2 n' \ \ [Bit]. \qquad (7)$$

As a result of increasing basic scope of word entropy into sentences (6) by additional entropy of its pairs (7) we get maximal general entropy of one sentence of the text

$$H_{sent} = H_{sw} + H_{sp} = n' \cdot (\log_2 l + 2 \cdot \log_2 n') \ \ [Bit]. \qquad (8)$$

Thus, application of proposed universal approach to processing the texts with m sentences increases general linguistic knowledge base of the individual by $m \cdot n' \cdot (\log_2 l + 2 \cdot \log_2 n') \ \ [Bit]$.

Analysis of phrase (8) demonstrates that in case of inconsiderable fluctuation of a number of significant words n' in a sentence, l – a number of recognized word forms (lexemes) of the text- remains a key parameter of general linguistic knowledge of the individual. It can be also inferred that there is a common estimate range $O(m \cdot n' \cdot l)$ of procedural complexity of the universal approach, which is commensurate to procedural complexity of TRM method for typical semantics-related tasks of automated abstracting.

To diminish procedural complexity of the proposed approach, which appears to be promising for solving a number of semantics-related tasks [11], frequency characteristics of vocabulary stock of natural language should be considered. Since considerable word forms (lexemes) of the sentence carry the most comprehensive information, direct exclusion of the so-called stop-words in the process of text parsing and identification of co-references of the pronoun considerable decreases value l. Thus, according to [16, 17] for synthetic Russian language specific weight in the corpus of such parts of speech as parenthesis, pronouns (for nouns, adjectives and adverbs), prepositions, conjunctions, particles make 38,1%. The situation is different for analytical languages like English, viz. according to [18], specific weight of nouns, verbs, adjectives and adverbs from most frequently used words makes 96,4%. On the other hand, exclusion of loss- making relations for relatively small number of prepositions of English sentence enables to considerably decrease the value n'. All those processes are provided by modern parsers.

5 Conclusion

It has been substantiated that semantics-related tasks should be identified as separate class, characterized by *NP*-complete complexity along with algorithms of natural reasoning, which provide efficient solutions of tasks of this class. The analogy between knapsack problem and automated abstracting has demonstrated that using linguistic knowledge is traditionally included in text processing algorithms and enables to decrease procedural complexity to polynomial.

Based on the obtained informational estimate of universal approach in text processing with relations between entities taken into account, maximal general entropy of one sentence of the text has been determined. Since complexity of proposed method is polynomial, and technological possibilities of modern parsers provide respective procedures of linguistic analysis of text, processing of the latter with m sentences under that approach can practically increase general linguistic knowledge base of the individual by $m \cdot n' \cdot (\log_2 l + 2 \cdot \log_2 n')$ $[Bit]$.

Promising direction for development of the research is to determine how acquired knowledge about relations between relevant entities of the text affect the accuracy of likelihood estimation $p(y_i \mid X)$.

References

1. U.Hahn, I.Mani: The Challenges of Automatic Summarization. IEEE Computer Society. Vol. 33, № 11, 29-36 (2000)
2. T.H.Cormen, C.E.Leiserson, R.L.Rivest, C.Stein: Introduction to Algorithms (2nd ed.), MIT Press and McGraw-Hill. Chapter 34: NP–Completeness, 966–1021 (2001)
3. Contrastive analysis of substantive polysemy : Based on German nad Spanish languages, theme of the thesis on spec. 10.02.20, assistant professor of philology Tatiana Anatoliyevna Yakovleva, Majored in: comparative and historical, typological and comparative linguistic (2001)
4. L.V.Nikolayeva: The notion of polysemy in nominative systems of terms. Culture of the People K of Black Sea Region, № 110, T. 2, 65-67 (2007)
5. Major types of sematic relations between subject domain terms [E-resource] / Available: http://cyberleninka.ru/article/n/osnovnye-tipy-semanticheskih-otnosheniy-mezhdu-terminami-predmetnoy-oblasti
6. K. Riedhammer, D. Gillick, B. Favre, and D. Hakkani-Tur: Packing the Meeting Summarization Knapsack. In Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH), 2434– 2437 (2007)
7. Indentification of complexity classes [E-resource] / Available: http://www.cs.berkeley.edu/~luca/cs172/karp.pdf
8. N.B.Shakhovska: Automated text abstracting. National University Bulletin: Lviv polytechnics, № 743: Informational systems and networks, 210–218 (2012)
9. O.V.Kanischeva: Using maps of text ratios for automated abstracting. National University Bulletin Lviv Polytechnics, № 770: Informational systems and networks, 108–122 (2013)
10. O.V.Bisikalo: Automated annotation of texts based on linguistic images. Cybernatic management and informational, № 1, 46–51 (2014)
11. O.V.Bisikalo: Identification of pertinent parameters of text based on the relations between lexical units. National Technical University Bulletin «ХПІ», Series: Mechanical and technological systems and complexes, X: НТУ «ХПІ», № 21 (1130), 83–89 (2015)
12. Andrei Andreyevich Markov. [E-resource] / Available: http://www-gap.dcs.st-and.ac.uk/~history/Biographies/Markov.html
13. C. E. Shannon. Prediction and entropy of printed english. Bell Systems Technical Journal, 30:50–64, 1951.
14. A.Berger, S.Della Pietra, V.Della Pietra: A Maximum Entropy Approach to Natural Language Processing [E-resource] / Available: http://www.cs.columbia.edu/~jebara/6772/papers/maxent.pdf
15. Ronan Collobert, Jason Weston, L´eon Bottou - Natural Language Processing (almost) from Scratch [E-resource] / Available: http://arxiv.org/pdf/1103.0398.pdf
16. Russian National Corpus [E-resource] / Available: http://www.ruscorpora.ru/corpora-stat.html
17. New frequency dictionary of Russian language [E-resource] / Available: http://dict.ruslang.ru/freq.php?act=show&dic=freq_pos&title=%C4%E0%ED%ED%FB%E5%20%EE%20%F7%E0%F1%F2%EE%F2%ED%EE%F1%F2%E8%20%F7%E0%F1%F2%E5%F0%E5%F7%ED%FB%F5%20%EA%EB%E0%F1%F1%EE%E2%20%28%ED%E0%20%EC%E0%F2%E5%F0%E8%E0%EB%E5%20%EF%EE%E4%EA%EE%F0%EF%F3%F1%E0%20%F1%EE%20%F1%ED%FF%F2%EE%E9%20%E3%F0%E0%EC%EC%E0%F2%E8%F7%E5%F1%EA%EE%E9%20%EE%EC%EE%ED%E8%EC%E8%E5%E9%29
18. Adam Kilgarriff. BNC database and word frequency lists [E-resource] / Available: http://www.kilgarriff.co.uk/bnc-readme.html