Automatic Selection of Verbs-Markers for Segmentation Task of Process Descriptions in Natural Language Texts

Varvara A. Krayvanova

Altai State Technical University, Barnaul, Russia
krayvanova@yandex.ru

Abstract. The paper presents the intermediate results of the research, the final goal of which is to develop the universal algorithm for process diagrams automatic visualization by text description of these processes. The purpose of this study is to check the use of verbs as markers for the semantic labeling of long fragments in scientific texts.

Keywords: automatic text fragmentation, text mining of scientific texts, verbs-markers, dynamic text parameters.

1 Introduction and Problem Statement

An effective system of collecting, storage and data processing of scientific observations will raise any natural-science research to essentially new level. The description of objects and procedures is presented in the form of natural language texts. Therefore automatic transformation of texts into more effective representations (such as activity diagrams, state diagrams, use case diagrams, IDEF0 diagrams, etc.) is required to reduce the cost of specialized information systems creation.

Current researches in the field of process descriptions extraction from natural language texts are oriented to work with the news bulletins [1] or with other objects from a very narrow areas [2,3]. The similar situation is beheld with the problem of process visualization [4,5]. These algorithms assume the texts of the small length containing concentrated information of a certain type. Researches in the field of text processing for arbitrary structure and size are usually oriented to extraction of objects, instead of processes, for example, on ontologies construction[6]. To generalize existing algorithms of processes extraction to the long natural language texts we have to use automatic segmentation and semantic marking of the text to find places, suitable for these algorithms usage. The objects of this research include long scientific, regulatory and educational texts (articles, tutorials, monographs).

To reach these goals, it is necessary to allocate text fragments with various assignments:

- static (descriptions of objects, definitions) for ontology extraction;

– dynamic (description of processes, techniques and research procedures) for activity diagrams and other process diagrams extraction.

A text can be divided into various fragments of these types with the use of clustering on the base of statistical analysis of parts of speech distribution[7]. We simulated the reading process by the sliding window method (window is the sequence of the length L of consecutive sentences). As clustering parameters, for each window the total number of words and the number of various words separately for nouns, verbs and adjectives are calculated. The studies of the various parameters distribution in long texts are focused mainly on the definition of the author [8]. This method allows to divide scientific texts into fragments of the types described above. For automatic illustration we have to find a way to define fragments types. One possible way of solving this problem is to analyze the distribution of verbs in the text. There are usually much less various verbs than nouns in the texts, especially in business and scientific ones[9]. Linguistic verbs classifications, e.g. the one given in the dictionary of linguistic terms by Rosenthal[10], are not good enough for extracting information from scientific texts. In scientific style, quite narrow verbs segments are applied, therefore linguistic classifications can be called excessive. Besides, used verbs and their meaning in the text significantly depend on concrete subject domain.

The purpose of the research is to check the possibility of using verbs as markers for different types of fragments.

For the illustrations we used *Bykov N. I., Popov E. S.* Observing the dynamics of snow cover in protected areas of the Altai-Sayan Ecoregion. Methodological guidance. Krasnoyarsk. 2011. 64 pages. The text consists of 1257 sentences. Parser identified 196 different verbs.

2 Mathematical model

Let V be the set of all natural language verbs. Scientific text T is represented as an ordered set of natural language sentences $T = \langle s_k \rangle$, where s_k is kth sentence in the text. Let $V_k \subset V$ be the set of verbs in the sentence s_k. For each verb let's define the list $E_v = \langle s_k | v \in V_k \rangle$. This is an ordered list of sentences that contain a verb v. $|E_v|$ is the number of occurrences of the verb v in the text T. Since the object of study is the verb distribution in the text, the cases of multiple use of a single verb within a sentence can not be ignored. Text neighborhood $T_v^\epsilon = \langle s_i \rangle$ of the verb v is an ordered set of sentences s_i, such that $\forall s_i \; \exists s_k \in E_v \; and \; |k - i| <= \epsilon$, ϵ is a non-negative integer. All the verbs from the text T are divided into three groups. The first group contains rare verbs V_{unic}. The number of occurrences $|E_v|$ in the text for these verbs is below the border β: $|E_v| < \beta$. The second group includes common verbs V_{common}. These verbs get the largest values of $|E_v|$, and are distributed relatively evenly within the text. Typically, these are parts of collocations from scientific speech style, such as "ОСУЩЕСТВЛЯТЬ" ("TO CARRY OUT"), "ПРОИЗВОДИТЬ" ("TO MAKE"). The third group contains verbs-markers V_{marker}. Those verbs-markers are present in the text in sufficient quantities and are unevenly distributed. These verbs can also be parts of collocations from scientific speech style.

Let each sentence s_k of the text T be assigned to some cluster c from a finite set of clusters C. For example, the set of clusters can be obtained by clustering on the base of the distribution of parts of speech along the text (described in detail in [7]). Clusters are obtained automatically, so their boundaries can be defined with an error margin. Let c_v^ϵ be a subset of textual neighborhood T_v^ϵ, which belonging to cluster c: $c_v^\epsilon = T_v^\epsilon \cap c$.

The verb v_m is marker of cluster c, if $|c_{v_m}^\epsilon|/|T_{v_m}^\epsilon| > \sigma$ and $\forall a \in C |a_{v_m}^\epsilon|/|T_{v_m}^\epsilon| \leq \sigma$. The values of ϵ and σ are parameters of marker detection algorithm and depend on the method of obtaining clusters C.

The text nest of verb-marker v_m is the set of verbs: $N_{v_m}^\mu = \{v | E_v \cap T_{v_m}^\mu \neq \emptyset\}$.

3 Results and Conclusion

The mathematical model described is realized in algorithms for automatic labeling of text fragments and for construction of text nests of verbs. In the software complex the sentence s_k is implemented as a parse tree of Dialing parser[1]. The table 1 presents the lists of verbs for the three clusters. The window size for clustering $L = 120$ sentences. The algorithm parameters values is $\epsilon = 7$ sentences and $\sigma = 0.9$.

Table 1. Verbs-markers for clusters

Cluster annotation (expert)	**Verbs-markers (automatic extraction)**
Cluster 1. Description of the research objects: introduction definitions and process of snow formation.	СЛУЖИТЬ, СМОТРЕТЬ, ЗАВИСЕТЬ, ЯВЛЯТЬСЯ, ОПРЕДЕЛЯТЬ, ИМЕТЬ, ПРОИСХОДИТЬ (TO SERVE, TO WATCH, TO DEPEND, TO BE, TO DEFINE, TO HAVE, TO HAPPEN)
Cluster 2. Chapter about calculations and laboratory processing of research results, different tables of classifications, fragments about parameters measurement.	ВЫЧИСЛЯТЬ, ВЫЧИСЛЯТЬСЯ, ЗАПИСЫВАТЬСЯ (TO CALCULATE, TO BE CALCULATE, TO REGISTER)
Cluster 3. Observation methodology: observation areas marking, equipment and recommendations.	СОСТОЯТЬ, ПРИНИМАТЬСЯ, ИСПОЛЬЗОВАТЬ, РЕКОМЕНДОВАТЬ, БЫТЬ (TO CONSIST, TO BE TAKEN, TO USE, TO RECOMMEND, TO BE as a link-verb)

Let's consider the example of a nest for marker "ВЫЧИСЛЯТЬ" ("TO CALCULATE") for $\mu = 7$ sentences:

[1] http://aot.ru/

`НАПОМНИТЬ, ОТСУТСТВОВАТЬ, РАССЧИТЫВАТЬСЯ, ОКРУГЛЯТЬ, ЗАПАСТИ, ПРЕДСТАВЛЯТЬ, УЧИТЫВАТЬСЯ, ОКАЗАТЬСЯ, ПРОБИВАТЬСЯ, ПОЗВОЛЯТЬ, ВЫБИРАТЬ, ПОДСЧИТЫВАТЬ (TO REMIND, TO BE ABSENT, TO BE CALCULATED, TO ROUND, TO STORE, TO PRESENT, TO BE CONSIDERED, TO APPEAR, TO BREAK THROUGH, TO ALLOW, TO CHOOSE, TO COUNT)`

The nest obtained shows that the set of verbs, which is located around the marker, belongs to the calculations and laboratory processing of research results for snow cover observations.

The algorithm of fragments labeling and nests construction has been checked using the test set containing 15 scientific texts of various authors and subjects. Selected verbs-markers are consistent with the expert annotation of the fragments content. Verbs-markers can be used for semantic labeling of automatically separated fragments, although some of them have no semantic value and are just stylistic features of a specific text. In the future we are planning to use verbs-markers to improve the accuracy of fragments boundaries determining. The nests of verbs received on the basis of the model presented will be used in algorithms of processes visualization using their text descriptions.

References

1. UzZaman, N., Allen, J.F.: Event and temporal expression extraction from raw text: First step towards a temporally aware system. Int. J. Semantic Computing **4**(4) (2010) 487–508
2. Wang, X., McKendrick, I., Barrett, I., Dix, I., French, T., Tsujii, J., Ananiadou, S.: Automatic extraction of angiogenesis bioprocess from text. Bioinformatics **27**(19) (2011) 2730–2737
3. Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F.: An Overview of Event Extraction from Text. In: Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011). Volume 779 of CEUR Workshop Proceedings., CEUR-WS.org (2011) 48–57
4. Johansson, R.: Natural Language Processing Methods for Automatic Illustration of Text. Licentiate Thesis. Department of Computer Science, Lund University, Lund, Sweden (2006)
5. Krayvanova, V., Kruychkova, E.: Automatic illustration of texts based on templates. In: Proceedings of All–Russian Conference ”Knowledge - Ontology - Theory” (KONT-13) with international participatio. Volume 1. (2013) 235–240
6. E., M.: Automatic ontology learning from text document collection. In: Proceedings of Russian Conference on Digital Libraries. (2011) 293–298
7. Krayvanova, V., Kruychkova, E.: Application of automatic fragmentation for the semantic comparison of texts. In: 15th International conference SPECOM 2013 Proceedings, September 1-5. Lecture Notes In Artificial Intelligence, Springer (2013) 46–53
8. Lvov, A.: Linguistic analysis of the text and author recognition (2008)
9. Homutova, T.: Research text: integral analysis of lexis. Language and culture (4) (2010)
10. Rozental, D., Telenkova, M.: Glossary of linguistic terms. 2 edn. Prosveshenie, Moscow, Russia (1976)

Автоматический выбор глаголов-маркеров для задачи выделения описаний процессов в текстах на естественном языке

Варвара А. Крайванова

Алтайский государственный технический университет, Барнаул, Россия
krayvanova@yandex.ru

Аннотация В статье представлены промежуточные результаты исследования, конечной целью которого является разработка универсального алгоритма для автоматической визуализации диаграмм процессов по текстовым описанием этих процессов. Цель данного исследования — проверка возможности использования глаголов в качестве маркеров для семантической маркировки длинных фрагментов в научных текстах.

Ключевые слова: автоматическое фрагментирование текста, text mining, глаголы-маркеры, динамические параметры текста.