Reconstruction of the phonetic composition the recognized word using lexical ontology

V S Moshkin[1], A I Armer[1] and N A Krasheninnikova[2]

[1]Ulyanovsk State Technical University, Severny Venets street 32, Ulyanovsk, Russia, 432027
[2]Ulyanovsk State University, Lev Tolstoy street 42, Ulyanovsk, Russia, 432017

Abstract. It is possible to improve the quality of speech recognition in noisy environments adapting the reconstruction algorithm for the recognized word to the certain peculiarities of usage and application. The article describes an approach to reconstruct the phonetic composition of a recognized word using lexical ontology. The lexical ontology contains lexical links among the words of the subject domain and their phonetic composition in terms of the SAMPA+ for the Russian language.

1. Introduction

Continuous speech recognition is a complex iterative process, based on the sequential division of the total acoustic signal into words, and then the words into phonemes. At the same time, many factors (e.g., noise) significantly complicate the recognition and reduce its accuracy. Thus, after preliminary speech signal processing many algorithms result in a matrix consisting of a phoneme set with their corresponding detection probability in the recognized word.

However, the task of recognizing and converting Russian continuous speech into a text is an urgent problem. It should be solved in different subject domains: medicine, litigation, automated detection of extremist materials, etc.

At the same time, a special set of the most frequently used terms corresponds to each subject domain. The a priori sampling of the basic terms in the analyzed subject domain helps to increase the probability accuracy while determining the patterns of certain phonetic combinations in the process of speech recognition.

One of the ways to adapt speech recognition algorithms to the corresponding subject domains is to apply lexical ontologies of subject domains.

2. The algorithm for the phoneme-by-phoneme recognition of the Russian words

For the initial extraction of speech units from the recognized speech signal, the algorithm of the phoneme-by-phoneme recognition of the Russian words is used. The initially detected speech units will be used to form word groups taking into account both subject domains and the analysis results of a certain lexical ontology.

For the phonemic recognition of the Russian words from an unlimited dictionary, we use the following algorithm: a speech segment is divided into constituents [1, 2]. For this purpose, it is preliminary transformed into a two-dimensional autocorrelation portrait. Then, in the sliding window, which size is equal to the corresponding portrait of a model speech unit, the distance is calculated. During the calculation, the distance between the windows is optimized using the discrete dynamic

programming. For each speech unit, a distance array along the portrait of the analyzed speech segment is determined. The distances corresponding to the same fragments of the analyzed speech segment portrait are compared with each other. As a result, speech unit portraits, which have the smallest distances, form the desired boundaries in such a way that the starting and final readouts of each speech signal are known. In such a method, the average error in determining the beginning of a speech unit is 3579 samples, in the interval [0; 10860] samples, the average error in determining the end of a speech unit is 3724 samples, in the interval [0; 12481] samples; the sampling rate is 44100 S/s.

Then, the speech signal of each speech unit is successively converted into an autocorrelation portrait in the following way. Let $s(i)$ be the *i-th* readout of a digital speech signal; then $s(i+k)$ is a readout spaced k readouts apart $s(i)$. Dependency factor of these readouts is expressed by a sample correlation coefficient:

$$R_s(k) = R[s(i), s(i+k)] = \frac{\mathrm{cov}[s(i), s(i+k)]}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} s^2(i) - m_{s(i)}^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N} s^2(i+k) - m_{s(i+k)}^2}},$$

$$\mathrm{cov}[s(i), s(i+k)] = \frac{1}{N}\sum_{i=1}^{N} s(i)s(i+k) - \left[\frac{1}{N}\sum_{i=1}^{N} s(i)\right]\frac{1}{N}\sum_{i=1}^{N} s(i+k), \quad (1)$$

where N is a number of readouts in the interval, in which the dependency is sought; $\mathrm{cov}[s(i), s(i+k)]$ is the sample covariance $s(i)$ and $s(i+k)$ when $i = 1..N$; $m_{s(i)}$ is a sample mean $s(i)$ when $i = 1..N$; $m_{s(i+k)}$ is a sample mean $s(i+k)$ when $i = 1..N$. Function determined by the sample correlation coefficient using (1) is an autocorrelation function of a signal. While its calculation we perform the transformation of a speech signal $s(i) i = 1..M$ (M is the number of readouts in a speech signal) into a two-dimensional image. For this purpose, $s(i)$ is divided into intervals including $N < M$ readouts. Then, using equation (1) we generate image lines:

$$\begin{matrix} s\left(i_m^j\right), s\left(i_m^j + k\right) & \begin{matrix} k=1..N \\ j=1,N,2N,\ldots \end{matrix} \\ X(j,k) = R \end{matrix}. \quad (2)$$

The two-dimensional image *X (j, k)* obtained from (2), where *i* is the line number, and *k* is the column number, is the autocorrelation portrait (ACP) of a speech signal *s (i)* dimensioned $N \times \frac{M}{N}$.

Model speech units with the same parameter value *N* are transformed into ACPs. These speech units are arranged from the examples of SAMPA + phonetic alphabet. Thus, it is possible to determine to what extent the ACP speech unit corresponds to the model ACP. As a result, the speech unit is considered to correspond to the model with the utmost similarity. The similarity of the ACP is determined by calculating the Euclidean distance between the ACP lines. The position of each line is determined in dynamics [3, 4, 5]. Nevertheless, if in ACPs the number of lines exceeds a certain threshold value they are considered different.

3. A model of lexical ontology

Ontology is a system consisting of many concepts, their definitions and axioms, which are necessary to limit the interpretation and use the concepts [6].

OWL (Ontology Web Language) is the Semantic Web language designed to describe classes and their interrelations. At the heart of the language lies the representation of reality in the "object-property" data model. OWL is a reformulation of the descriptive logic using XML syntax.

Subject domain ontology is a collection of RDF-triples: subject-predicate-object. In this research OWL-ontology was used to solve the problem under consideration [7, 8].

A special type of ontology is lexical (or linguistic). Its distinctive property is to use (lexicalized) concepts (words) together with their linguistic properties in one resource. The main source of concepts in such ontologies are the values of linguistic units. They are also distinguished by a set of relationships, which usually characterize linguistic elements: such as synonymy, hyponymy, meronymy, etc. [9, 10, 11].

To reconstruct the phonetic composition of the recognized word, the elements of phonetic alphabets, establishing letter-sound correspondence, were included in the lexical ontology structure. The most widely-used phonetic alphabets are the International Phonetic Alphabet (IPA) and X-SAMPA (as well as SAMPA + modification, including transcriptions of the Russian language) [12].

The formal model of a certain subject domain lexical ontology may look as follows:

$$O = \left\langle A, C^{A_i}, P^{IPA}, P^{SAMPA+_i}, R^{A_i} \right\rangle, \mathrm{i} = \overline{1, m},$$

where m is the number of subject domains covered by ontology; $A = \{A_1, A_2, K, A_t\}$ is the number of subject domains covered by ontology; $C^{A_i} = \left\{C_1^{A_i}, C_2^{A_i}, K, C_n^{A_i}\right\}$ is a set of terms within the *i-th* subject domain; P^{IPA}= is a set of phonemes peculiar to the Russian language according to IPA; $P^{SAMPA+} =$ is set of phonemes peculiar to the Russian language according to SAMPA+. This alphabet consists of 89 phonemes.

R^{A_i} is a set of ontology ratios within *i-th* subject domain:

$$R^{A_i} = \left\{R_C^{A_i}, R_{C-P}^{A_i}, R_P^{A_i}, R_{P_1P_2}^{A_i}\right\},$$

where $R_C^{A_i}$ is a set of links, which form the hierarchy of ontology terms within *i-th* subject domain; $R_p^{A_i}$ is a set of object properties and data type properties, which determine the relationship between the elements of P^{IPA}, P^{SAMPA+} sets, and also the corresponding properties of these objects. $R_p^{A_i} = \{hasIPA, hasSampaPlus, hasExample...\}$.

$R_{C-P}^{A_i}$ is a set of relations determining the links among the set objects C^{A_i} and P^{IPA}, P^{SAMPA+} within *i-th* subject domain. The given property determines whether the phoneme belongs to the phonetic representation of the corresponding term.

$R_{P_1P_2}^{A_i}$ is a set of links determining the probability that the phoneme P_2 follows the phoneme P_1 within the terms of the *i-th* subject domain. This value depends both on the frequency of the term in the texts of the *i-th* subject domain and on the phoneme sequence in the phonetic representation of a certain term:

$$R_{P_1P_2}^{A_i} = \mu_C^{A_i} \times \rho_{j,k}, \quad (3)$$

where $\mu_C^{A_i}$ is the grade of membership of the term *c* to the subject domain A_i; $\rho_{j,k}$ is the probability that the phoneme k follows the phoneme j within the terminological frames of the subject domain A_i.

The values of these relations are derived from statistical analysis of large text corpus on a given subject domain and from phonetic analysis of each marked term. The task of terminology extraction is solved with the help of semantic algorithms, i.e. thesaurus-based and nested link algorithms [13].

4. Thesaurus-based algorithm

The thesaurus-based algorithm for terminology extraction from a set of words, which belong to a certain text, using the OWL-ontology, calculates the degree of semantic proximity of input word-groups to the terms of the subject domain. This algorithm chooses from the set of incoming words/ word groups only those terms and expressions, which belong to the given subject domain.

The degree of semantic proximity of the input word-group to the subject domain k_{ont} can belong to the interval from 0 to 1: the closer is the obtained value to 1, the greater is the possibility that this word/ word group is a term.

The thesaurus-based algorithm suggests a direct search for input word lemmas and their combinations among the terms defined in a certain ontology. For this purpose, for every type of ontology, it is necessary to define the property “containsLemma”, which has a line value obtained by lemmatizing (reducing to the initial form) the object name with the help of Mystem (Yandex product). The lemmatization is carried out according to certain morphological peculiarities of the term.

The thesaurus-based algorithm consists of the following stages:

1. evaluation of the degree of proximity of the input word / word group to each ontology object;
2. search for the OWL-ontology core object, which is most closely associated with the input word / word group.

The scheme of the algorithm is shown in Figure 1.

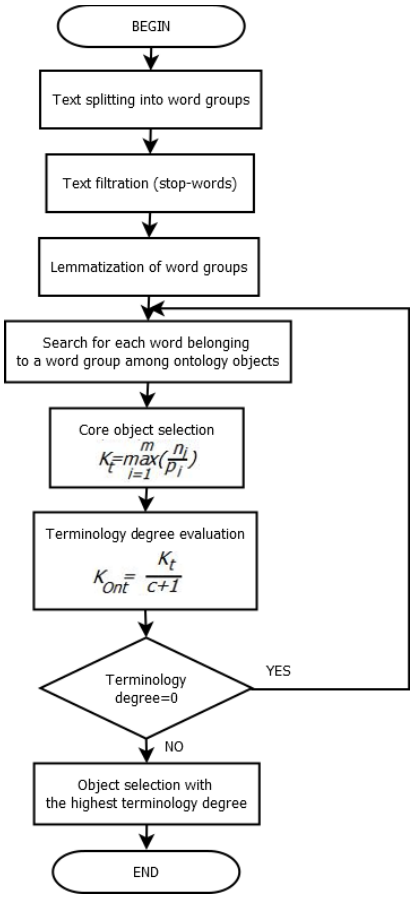

Figure 1. Schematic block diagram of the thesaurus algorithm for extracting terminology.

The reference core object of the expandable ontology, used in the further analysis, has a degree of proximity to the input word / word group, which is calculated by the following formula:

$$k_t = \max_{i=1}^{m}(\frac{n_i}{p_i}),$$

where m is the number of all objects of the OWL-ontology core; p_i is the number of words in the lemma of the reference OWL-ontology core object; n_i is the number of words from the input word group lemma, which are found in the lemma of the OWL ontology core object.

The general scheme to evaluate the degree of proximity of the word groups to the terms of the subject domain is shown in Figure 2.

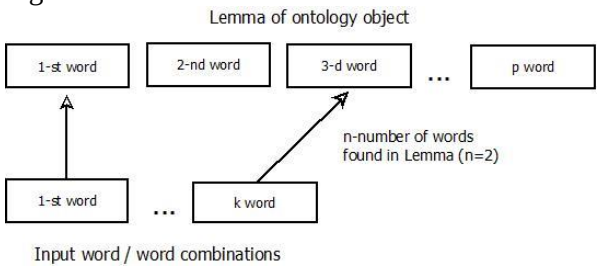

Figure 2. Thesaurus-based criterion. Search for reference object.

In this case, the word order in the group of words in the reference object should not change.

If several different ontology objects have the same value of the coefficient k_t,, then the object that corresponds to the maximum n_i. will be considered as a reference one. If there are several maximum objects, then all of them will be considered reference ones and the analysis according to ontological criterion will be carried out for each of them.

The ontology structure of the analyzed subject domain assumes that each of its objects has a Datatype Property “a Term”, which is of a logical type. This property is auxiliary and is determined by the expert who distinguishes to what extent this ontology object is peculiar for a certain subject domain.

Thus, the terminology degree of the input word group is calculated in the following way:

$$k_{Ont} = \frac{k_t}{c+1},$$

where k_t is a value obtained during the first stage of the algorithm execution; c is the number of ratios between the reference object and the nearest ontology object that has a Data Property “a Term”=true (if this property is true for the reference object, then c=0). The search scheme is shown in Figure 3.

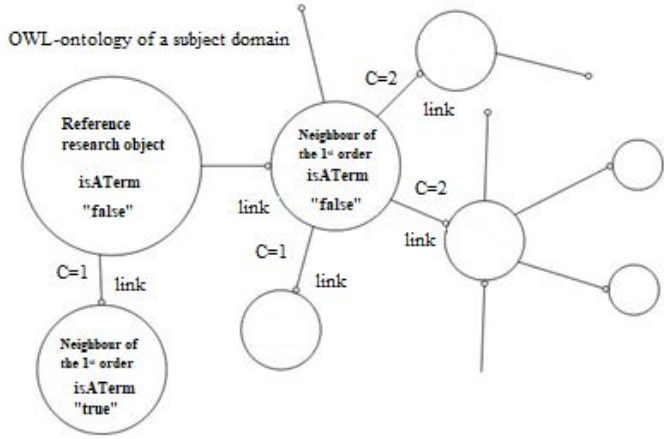

Figure 3. Terminology degree detection of input word groups.

If for the reference object “a Term”=false, and in addition the object has no links with other ontology objects, or all related objects are also distinguished as “false”, then it is necessary to find another reference object for the given word / word group and the evaluate it. In a similar situation with other reference objects, or in the case of their absence, the input word / word group is said to be “not a term” (k_{Ont}=0).

Thus, the process of the terminology degree evaluation of the input word group is a movement along the graph, at the nodes of which there are objects of the corresponding ontology classes.

5. Nested link algorithm

In addition to the terminology degree evaluation of a word / word group, the developed metrics makes it possible to extract terms from the text by comparing them with the existing objects and combinations of lemmas of the corresponding objects with the help of R_{add} ratios, which expand the set of objects of the described subject domain by combining lemmas of related objects. For example: the properties “has a Link” and “is a Part”.

Thus, while comparing the input word groups and the objects of the subject domain, which are related to each other by unidirectional relations R_{add}, the word group will be said to belong to a certain subject domain, if its lemma completely corresponds to the set of lemmas of the corresponding ontology objects.

The nested link method makes it possible to extract terms from the text by comparing them with the existing objects and combinations of lemmas of the corresponding objects using the links defined in the ontology.

The scheme of the algorithm is shown in Fig. 4. The peculiarity of this method is the necessity to represent the ontology objects mainly in the form of single words with the maximum number of links among objects. The determining factors for this method are R_{add}, links, and it is possible to form word combinations in a natural way using these links.

$$t_1 + R_1 + t_2 + R_2 + \ldots + t_i + R_j + \ldots + t_m + R_n,$$

where $R_i \in R_{add}$, $t_j \in T$,T are the terms of the application area, which the ontology describes.

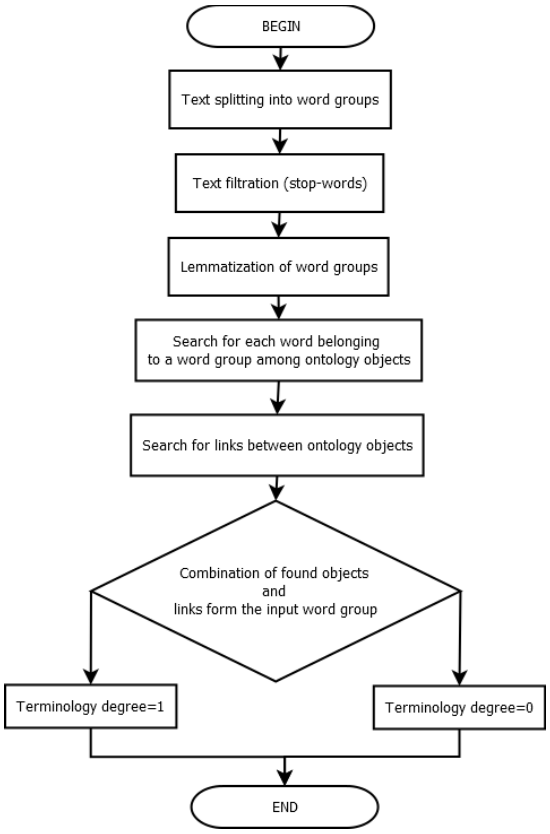

Figure 4. Diagram of the nested link algorithm of terminology extraction.

The scheme of the given algorithm is shown in Fig. 5.

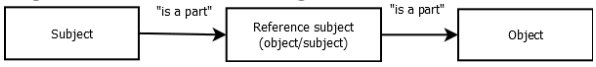

Figure 5. Nested link algorithm.

In this case, the extracted terms, which in its turn belong to terms consisting of a greater number of words, are not considered as terms to avoid redundancy [14].

6. Ontology development

In the course of the research, the linguistic ontology "Cardiovascular diseases" was developed for further recognition of the continuous speech, namely dictated medical diagnoses for this research area.

The ontology has 4 levels of hierarchy, and contains 743 terms belonging to the subject domain. Most of these words were automatically extracted from the texts using the nested link algorithm.

Figure 6 shows a fragment of the developed ontology that contains a description of all the phonemes used in the pronunciation of the extracted terminology.

Moreover, the probabilities of phoneme occurrence, which were found in the extracted terms of the subject domain under consideration, were calculated according to (3). These values will limit the set of selected phonemes while reconstructing the phonetic composition of the recognized words, the speech signal is divided into.

7. Conclusion

The use of lexical ontology implies the possibility to determine the basic set of terms for the analysed subject domain and, as a result, to increase the probability of accurate determination of certain phonetic combinations sequence in the process of speech recognition.

Within the framework of this research, it is planned to carry out a number of experiments to reconstruct the phonetic composition of recognized words using the developed model of the lexical ontology "Cardiovascular diseases" in order to validate the approach effectiveness.

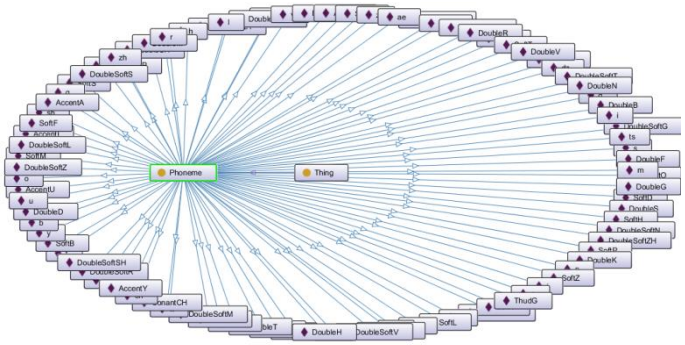

Figure 6. Fragment of lexical ontology. Phonemes.

8. References

[1] Andreev I A, Armer A I, Krasheninnikova N A and Moshkin V S 2017 Attacking the problem of continuous speech segmentation into basic units *III International conference Information Technology and Nanotechnology* 473-476

[2] Andreev I A, Armer A I, Krasheninnikova N A and Moshkin V S 2017 Attacking the problem of continuous speech segmentation into basic units *III International conference Information Technology and Nanotechnology* 6-9

[3] Pienado A and Segura J C 2006 *Speech recognition over digital channels: robustness and standards* (John Wiley & Sons Ltd.) p 257

[4] Keshet J and Bengio S 2009 *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods* (John Wiley & Sons Ltd.) p 253

[5] Gelbart D 2008 Ensemble Feature Selection for Multi-Stream Automatic Speech Recognition *Technical Report No. UCB/EECS-2008-160* (University of California at Berkeley) p 117

[6] Solov'ev V D, Dobrov B V, Ivanov V V and Lukashevich NV 2006 *Ontologies and thesauri* (Kazan', Mosow)

[7] Fellbaum C 1998 *WordNet: an Electronic Lexical Database* (MIT Press, Cambridge)

[8] Moshkin V S and Yarushkina N 2015 Methods for constructing fuzzy ontologies of complex subject domains *Open Semantic Technologies for the Design of Intelligent Systems* (Minsk: BGUIR) 401-406

[9] Khoroshevskiy V F 2008 Knowledge domains on the Internet and Semantic Web *Artificial Intelligence and Decision Making* **1**

[10] Mikhaylov D V, Kozlov A P and Emelyanov G M 2016 Extraction the knowledge and relevant linguistic means with efficiency estimation for formation of subject-oriented text sets *Computer Optics* **40(4)** 572-582 DOI: 10.18287/2412-6179-2016-40-4-572-582

[11] Mikhaylov D V, Kozlov A P and Emelyanov G M 2015 An approach based on tf-idf metrics to extract the knowledge and relevant linguistic means on subject-oriented text sets *Computer Optics* **39(3)** 429-435 DOI: 10.18287/0134-2452-2015-39-3-429-435

[12] Galunov V I and Solov'ev A N 2004 Modern issues in speech recognition *Information Technology and Computer Systems* 41-45

[13] Andreev I A, Bashaev V A, Kleyn V V, Moshkin V S and Yarushkina N G 2015 Estimation of the terminology of lexical units on the domain ontology basis *Open semantic technologies for the design of intelligent systems* (Minsk: BGUIR) 395-400

[14] Yarushkina N, Moshkin V, Klein V, Andreev I and Beksaeva E 2016 Hybridization of Fuzzy Inference and Self-learning Fuzzy Ontology-Based Semantic Data Analysis *Proceedings of the First International Scientific Conference "Intelligent Information Technologies for Industry"* 277-285

Acknowledgements

This work was supported by RFBR. Projects № 16-48-732046 and № 18-37-00450.