International Journal of Engineering & Technology, 7 (4.38) (2018) 422-425

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



The Use of Psycholinguistic Patterns in Interactive Systems of Active Information Retrieval

Valery Evgenevich Sachkov[1]*, Dmitry Aleksandrovich Akimov[1], Sergey Aleksandrovich Pavelyev[1]

[1]Russian Technological University (MIREA), Vernadsky Avenue, 78, Moscow, 119454, Russia
**Corresponding author E-mail: sachkov.v.e@mail.ru*

Abstract

The article explores the possibility of using psycholinguistic patterns in a dialogue with the Internet visitors. The scheme of the semantic kernel is shown for the purpose-setting installation of the search system and the methodology for constructing patterns, taking into account the psycholinguistic features of constructing a dialogue for obtaining the required information. The model of building psycholinguistic patterns for revealing the semantic information in dialogues is given. Patterns are based on associative links of words and word combinations. Such associative connections allow expanding the list of related words and revealing key information in the best way from short messages. The use of such a method in interactive active search systems makes it possible to improve information exchange and achieve a higher level of identifying the purpose of the dialogue.

Keywords: *Active Search; AIML; Dialog systems; Psycholinguistic patterns; RDF; Semantic core.*

1. Introduction

When searching for information, search engines use text or other information that has already been published on the Internet in the form of content. There are problems when you need to apply the so-called active search [1]. Active search means a search in which a search program, registered on behalf of a visitor on a resource that allows you to post comments, conducts a dialogue with visitors on the 'problem of interest', integrates the answers and sends the result to the Customer.

At the same time, copies of such programs should be placed on resources whose visitors are geographically distributed. As a result, it becomes possible in real time to monitor processes (movement of equipment, people, and map of structures) on given territorial formations [2].

Sample questions:

– I'm going to tomorrow with my family to 11 in the city of X, how long will I have to stand on the ferry?

– Do not tell me, on the way to Y today you can drive. I was told that it was being repaired. Did any of you on it pass in the near future?

Questions with some modification can be re-assigned automatically when new resource visitors are discovered.

Software that actively searches for information (purposeful action) – an informational named dialogue system that is able to adapt to any input data to achieve the set goal - will be called an actor of the Internet.

The text in natural language is understandable to the user, but for the machine it is represented by not more than a set of encoded symbols. To extract valuable data the machine must solve many tasks for processing text in natural language. The special section of applied linguistics "Automatic processing of text in natural language" or "Natural Language Processing" (NLP) in an English-speaking audience is engaged in such tasks. To date, there is a wide variety of natural language processing tasks; the most common of them are the followings [3]:

• Search for text fragments – dividing the text into different elements of different types: words, sentences, paragraphs and so on, and so forth.

• Sentence Boundary Disambiguation (SBD) – definition of supply boundaries.

• Named entity recognition (NER) - the mechanism for searching for addresses, names, names, dates, or any other named entities.

• Definition of parts of speech (Parts of speech, POS) – classification of text elements at the level of the sentence. The sentence can be divided into separate words and phrases in such categories as nouns, verbs, adverbs, prepositions and so on, and so forth.

• Classification of texts and documents - the purpose of this classification is in assigning labels to fragments found in texts and documents.

• Sharing relationships – identifying the links between words or phrases, to build a semantic tree.

One of the priority tasks of automatic text processing is the semantic analysis or understanding of the content and semantic part of the text. In the creation of interactive systems in natural language semantics is widely used, because the first requirement of all these systems is a handling properly the user's request. A reference example is the first dialogue system "Eliza" created in 1966, that imitating a dialogue with a psychoanalyst [4]. For its time the program "Eliza" was a breakthrough, but its algorithm was based on simple rephrasing of questions and it was far from understanding the "meaning". A modern example of the development of interactive systems can be represented by the program "A.L.I.C.E" [5] – this is a virtual interlocutor, able to conduct a dialogue in a natural language. The basis of "A.L.I.C.E." is the AIML artificial intelligence markup language, which is discussed in more detail in the article below.

When creating "Eliza", "A.L.I.C.E", etc. systems, the key moment was always the possibility of understanding the "meaning" of the

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

user's query in natural language, respectively, the need to create a tool that can identify from the text, which the user asks the system and reorganize this information to a formal and comprehensible view handling by the computer is still an actual problem.

2. Analysis of the Problem

The most difficult stage of active search in natural language is considered to be meaningful analysis. To perform it successfully, you need to know what the meaning of the word and sentence is, how to describe these values formally, to represent and store the contents of the text in the computer's memory, to perform operations with values, to translate values from natural language to formal scripts and vice versa. The answers to these and many other questions are provided by computer semantics, which is responsible for the development of models of the semantic level of natural language [6].

The use of interactive dialogue systems (IDS) greatly simplifies the interaction of the computer with the user, due to the fact that communication between them occurs in natural language. The user of such a system should not have special skills and knowledge, and if possible, the IDS should conduct a meaningful dialogue, which increases the convenience and trust of the user to this system. The implementation of these IDS functions often requires the development of complex intelligent systems based on knowledge bases, rules, dictionaries, etc.

The current trend of the development of IDS is still built due to significant engineering and expert knowledge. In the article the authors present an overview of such an approach with detailed analysis of data sets, document hulls in natural language with the help of which it is possible to create interactive systems based on machine learning models for different topics. One of the important properties of dialog systems is the appropriacy of the generated response to the user's question, and it is very important to be able to evaluate the answer given to the user in natural language, where the final result should be the user satisfaction rating [7]; the comparison of several machine metrics Evaluation with the assessments of the person on the results of the dialogue system. The generation of natural language is a critical component of conversational dialogue and has a significant impact on both usability and perceived quality. Most of developers use rules and heuristics, generate stiff and stylized answers without a change in the natural human language; [2] considers a response generator based on semantically controlled structure of the "Long Short-term Memory" (LSTM) neural network [8], in order to try to approximate the machine responses to the natural language. Another example of constructing an interactive system can be found in, where large data sets are used to train recurrent neural networks.

To teach a computer to understand a natural language, it must be able to receive, extract and process the content of the document, be able to understand the meaning of words, phrases and individual sentences. Since the detection of the meaning and values involved in computer semantics, the computer needs some structure to store this knowledge.

An example of an attempt of the implementation is the concept of "Semantic Web" [7], on the basis of which a storing semantic structures model of documents "RDF" [9] was developed. An important problem of the RDF model is the complex structure of ontologies, and the SPARQL language, proposed as a standard one, has a number of significant shortcomings in the grammar and semantics of queries [10]. It was also influenced by not so wide dissemination of the concept among developers, due to the fact that it was necessary to create two documents identical in content, but one for "people" and the other for "machine".

Another modern approach in understanding the semantics of text, searching for related words and determining the subject matter of the text, is based on training models of neural networks. There are two competing approaches: one based on frequency entry into the body of documents, the other on the prediction model. A popular representative of the first approach is the latent-semantic analysis (LSA) [11]; the representative of the second is a set of algorithms word2vec [12].

The LSA model is based on frequency counts, where similar words have the same values in different documents. LSA is generally widely used in search engines, for indexing and searching for related words and documents. But the model has a number of disadvantages, namely: the sparse data and the ignoring of semantic associations between words, for example: "Hello, how are you", "Hi, how are you", "Hey, how are you". Also, the model does not work well with large data sets, because of the large memory consumption during calculations and loses in results' quality comparing with another model [13].

In the Word2Vec model, the word is represented as a vector, and the model itself predicts the set of nearest vectors to the original one by distributive attributes [14]. Advantage and at the same time a drawback of the model is the need for training on a large body of data, but it is more economical for computational resources. But even this model, despite the best results, cannot compare documents, only, according to individual words; it has the problem of frequency overlapping of words, which generates a semantic ambiguity.

The models considered have several advantages and disadvantages, a more detailed comparison can be found in [2015].

Summing up, we can single out that for the development of modern IDS a solution that takes into account the semantic associations is needed, based on the predicted model of word similarity, solves the problem of frequency overlapping, and also has an algorithm for comparing two documents in natural language. Based on these requirements for the semantic analysis of the text in the natural language set forth earlier, special software was developed: "Associative-semantic text preprocessor"; and in this paper the results of its application in the creation of modern IDS are given.

3. Methodology for Constructing Associative-Semantic Text Preprocessor

Important in determining the feature of the author’s dialogue is the mood and emotional state at the time of writing the message.

By cognitive psycholinguistic characteristics we will understand the study of the connection between speech messages and characteristics of participants in communication [4]. The concrete designing of the cognitive psycholinguistic model is based on the associative semantic vectors of word forms and the connection between them. As a technical basis for calculating psycholinguistic text constructs, let's take the semantic analysis apparatus and its representation in the form of an Associative-semantic text preprocessor.

Associative-semantic text preprocessor (ASTP) is intended for preliminary normalization of texts with the purpose of transformation into sets of associative semantic vectors with a given semantic annotation. Additionally, natural language processing functions in the processes of interaction of the natural language and languages of computer systems and robotic devices (stemming and tokenization) are supported for subsequent processing by the NLP tools of the body of documents for testing hypotheses, training and statistical linguistic analysis.

The ASTP preprocessor is intended for embedding in application packages as a third-party library and as an independent application. In order to interact with the IDS, let's take one of the ASTP algorithms responsible for semantic search of texts close in meaning Fig.1.

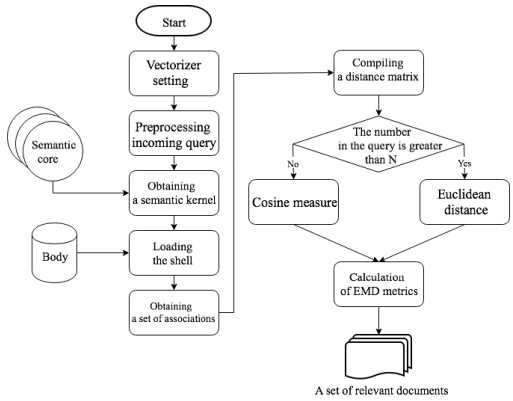


Fig. 1: Algorithm of associative-semantic search

Consider the most important elements of the algorithm:

• The semantic core is the vector space of the semantic field, in which the search for related words and associations is performed [16].

• The body is a set of specially prepared messages from the dialog system.

• The association assessment module is a module that evaluates and filters found associations in a semantic kernel.

• The EMD metric is a method for estimating the dissimilarity between two multidimensional distributions in some feature space, where a distance measure between single attributes is given [17]. The EMD metric calculates the minimum cost of changes or work required to convert one document to another. The calculation of EMD is based on the solution of the transport problem

• The distance matrix is the matrix of the weights of each word in the document to calculate the EMD metric.

4. Results of the Application of Psycholinguistic Patterns

As a basis for the creation of the IDS, a special standard for the Marketer Language was developed.

Artificial Intelligence Markup Language (AIML) is an artificial intelligence markup language that is compatible with XML, which allows you to create virtual interlocutors [1]. This markup language accelerates the creation of a dialog system, to which the ASTP preprocessor will be applied, for the associative-semantic search. An example of the AIML document is shown in Fig. 2.

```
<aiml version="1.0.1" encoding="UTF-8">

    <!—HELLO—>

    <category>
        <pattern> hello </pattern>
        <template>
            <random>
                <li> Hello! </li>
                <li>Good afternoon </li>
                <li>  Glad to see you  </li>
            </random>
            <random>
                <li>what is your name</li>
                <li> your name as I call you </li>
                <li> I will remember you as  </li>
            </random>
        </template>
    </category>

    <category>
        <pattern>*</pattern>
        <that>_what is your name </that>
        <template>
            I will remember you a:
            <set name="user">
                <star/>
            </set>
        </template>
    </category>
```

Fig. 2: AIML Template

To create an AIML document that would allow us to enter the dialog of the associative-search system with the user in natural language, the following semantic blocks were developed [18]:

• Greeting – this block is responsible for the greeting with the user, here the dialogue system remembers the user name, for further use in the dialog

• About myself – in this block the dialogue system tells about itself and its capabilities

• Associations – this unit is responsible for processing requests related to the search for associations, using the ASTP preprocessor

• Search – this block is responsible for the associative-search system

• Default response – this block is triggered only if none of the previous ones were able to process the user's request

A simplified diagram of the dialogue in natural language is presented in Fig. 3.

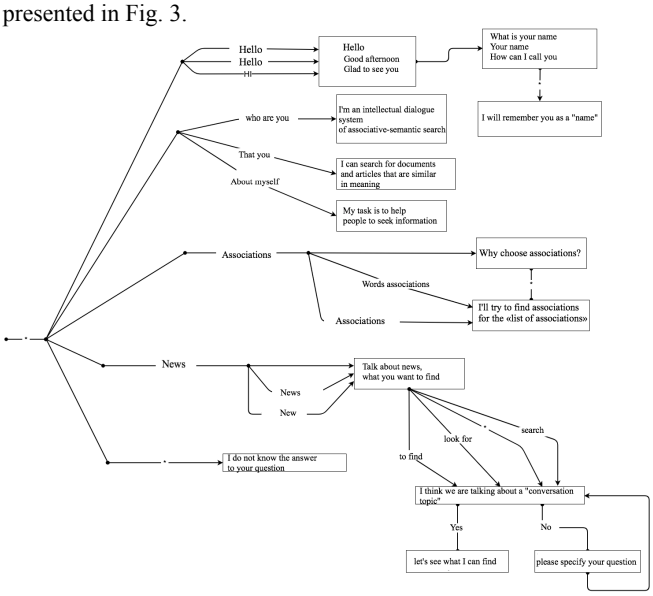


Fig. 3: Diagram of dialogue in natural language with the user

IDS greatly simplifies the search for similar meaning texts and associations, as it allows users to communicate with the search engine in a natural language, increasing the number of query formulations that can be used [19]. An example of the application of ASTP algorithms with AIML markup language can be the handling of a small application for finding associations in the semantic core of the message body from users, the output of which is presented in Fig. 4.

```
Loading bot/astp_bot.aiml...done (0.00 seconds)
User >>> привет
Bot  >>> Рада Вас видеть! Как вас зовут?
User >>> Watto
Bot  >>> Я запомню вас как  Watto
User >>> помоги мне найти ассоциации
Bot  >>> Скажите  Watto  к чему подобрать ассоциации?
User >>> истребитель
Bot  >>> Хорошо  Watto  попробую найти ассоциации для:  истребитель
Bot  >>>  Я думаю следующие ассоциации подходят вам: {'истребител', 'штурмовик',
 'миг', '30мки', '25см3'}
```

Fig. 4: General conversation with the dialogue system

In the end, we will demonstrate the accuracy of the IDS to find answers from the body of the message from users on the subject of the "Savior aircraft" dialogue; the graph is shown in Figure 5.

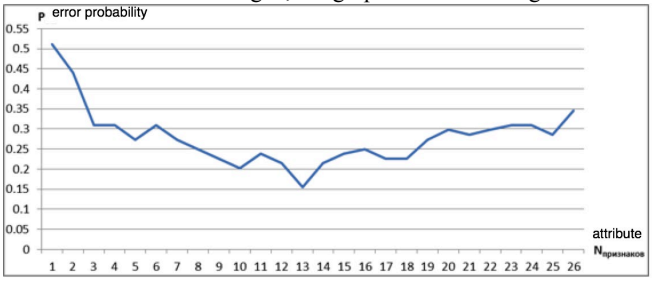


Fig. 5: Average error in message topic classification

5. Conclusion

As research shows the possibility of using specialized psycholinguistic patterns increases the level of interaction between the user and the machine, since they allow dialogue in the natural language. And unlike, for example, from the RDF model, the application of this software does not require much complexity in composing ontologies and complex search queries for document cases. They do not have the same great complexity in implementation and training, which makes it possible to give the IDS a minimal necessary understanding of the topic of dialogue for solving typical problems without attracting large computing and intellectual resources. The use of semantic, associative and psycholinguistic patterns shows that it allows improving the quality of dialogue sequences, as well as increasing the relevance of the requested information to Internet visitors. The results showed the fundamental possibility of creating a system of active information retrieval. The subject of further research can be the automatic modelling of personalized psycholinguistic patterns with self-learning possibilities.

References

[1] Marietto M das GB, de Aguiar RV, Barbosa G de O, Botelho WT, Pimentel E, França R dos S & da Silva VL (2013), Artificial Intelligence Markup Language: A Brief Tutorial. *International Journal of Computer science and engineering Survey* 4(3), 1-20.

[2] Liu C-W, Lowe R, Serban IV, Noseworthy M, Charlin L & Pineau J (2016), How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, 2122-2132.

[3] Sachkov VE, Gilmutdinova EF, Matyash ED & Akimov DA (2016), Processing and computer analysis of the text in natural languages. *Journal of Contemporary Science: Actual Problems of Theory and Practice, Series of Natural and Technical Sciences* 12, 57-64.

[4] Colbaugh R & Glass K (2013), Analyzing social media content for security informatics. *Proceeding of European Intelligence and Security Informatics Conference*, Uppsala, Sweden, 45–51.

[5] Weizenbaum J (1976), *Computer Power and Human Reason: From Judgment to Calculation*. New York: Freeman and Company.

[6] Wallace RS (2009), The Anatomy of A.L.I.C.E. In: Epstein R, Roberts G & Beber, G (Eds) *Parsing the Turing Test*. Springer, 181-210.

[7] Serban IV, Sordoni A., Bengio Y., Courville A., Pineau J (2016), Building End-To-End Dialogue Systems Using the Generative Hierarchical Neural Network Models. *Proceedings of the AAAI'16 Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence.* Phoenix, Arizona, 3776-3783.

[8] Henderson M, Thomson B & Williams J (2014), The Second Dialog State Tracking Challenge. *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Philadelphia, PA, 263-272.

[9] Resource Description Environment (RDF) (2004), Concepts and Abstract Syntax. *The World Wide Web Consortium (W3C),* 10 February. https://www.w3.org/TR/rdf11-concepts/

[10] W3C Semantic web activity (2013), *The World Wide Web Consortium (W3C)*. https://www.w3.org/2001/sw/

[11] Wikipedia, *Semantic core* https://en.wikipedia.org/wiki/Semantic_Channel

[12] Landauer T, Foltz PW & Laham D (1998), An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2-3), 259-284.

[13] Mikolov T, Le QV, Sutskever I (2013), Exploiting Similarities among Languages for Machine Translation. https://arxiv.org/abs/1309.4168

[14] Morozova YuI (2012), The construction of semantic vector spaces of different subject domains. *The Third School of Young Scientists of the IPI RAS. Collection of reports*. Moscow, 4-11.

[15] Levy O, Golberg Y & Dagan I, Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics* 3, 211–225.

[16] Altszyler E, Sigman M, Ribeiro S & Slezak DF (2017), Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. *Consciousness and Cognition* 56, 178-187.

[17] Serban IV, Lowe R, Henderson P, Charlin L & Pineau J (2018), A Survey of Available Corpora for Building Data-Driven Dialogue Systems. *Dialogue and Discourse* 9(1), 1-49.

[18] Doan S, Vo BKH & Collier N (2011), An analysis of Twitter messages in the 2011 Tohoku Earthquake. *International Conference on Electronic Healthcare*, 58–66

[19] Rennie JDM, Shih L, Teevan J & Karger DR (2003), Tackling the Poor Assumptions of Naive Bayes Text Classifiers. *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*. Washington, 616-623.