Results

With the help of the developed systems were calculated parameters of such microelectromechanical devices:
MEMS sensor with a circular membrane;
Differential capacitive MEMS accelerometer.
Examples of output data presented in Fig. 4, Fig. 5.

Conclusion

Today there are quite a large number of commercial systems which allow calculating the parameters of MEMS in certain situations, such as the calculation of the geometry of sensor, how thick membrane is needed in order to don't get significant deflection after applying pressure, etc. Unfortunately, these systems are expensive, demanding of computing resources and not easy to use. Organization cloud computing will greatly simplify the calculation of parameters of MEMS, making them cheaper and faster. Therefore, the calculation of parameters of MEMS using cloud computing is extremely important and urgent task.

1. Годовицын И. В., Сайкин Д. А., Федоров Р. А., Ameличев В. В. Расчет и моделирование основных параметров дифференциального емкостного МЭМС-акселерометра // Сб. науч. тр. IV Всероссийской научно-технической конференции "Проблемы разработки перспективных микро– и наноэлектронных систем – 2010. 2. Mossaddequr Rahman, Sazzadur Chowdhury "A Highly Accurate Method to Calculate Capacitance of MEMS Sensors with Circular Membranes" Electro/Information Technology, 2009. eit '09. IEEE International Conference. – P. 178–181. 3. Tkachenko S., Kulpa R., Havryshko V. Training programs for calculating the parameters of MEMS using cloud computing / CADMD'2014, October 10–11, 2014, Lviv, UKRAINE. – P. 143–145.

UDC 004.93.1

A. Zayats, A. Romaniuk
Lviv Polytechnic National University, Computer Aided Design department

CREATION OF THE ANNOTATED TEXT CORPUS FOR AUTOMATIC RECOGNITION OF SEMANTIC RELATIONS WITHIN NAMED ENTITIES

СТВОРЕННЯ АНОТОВАНОГО КОРПУСУ ТЕКСТІВ ДЛЯ ЗДІЙСНЕННЯ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ СЕМАНТИЧНИХ ЗВ'ЯЗКІВ МІЖ СУТНОСТЯМИ

This article describes the general approach to creation and annotation of the medical text corpora for Ukrainian language for its future application in arelation extraction system. The annotation process was done with the help of General Architecture for Text Engineering (Gate) system.

Key words: information extraction, semantic relations extraction, annotations, annotation scheme, data corpora.

Описано загальний алгоритм створення та анотування корпусу медичних текстів українською мовою, для їх подальшого використання в системах видобування зв'язків. Анотування даних проводилося за допомогою системи Gate (Загальна архітектура обробки текстів).

Ключові слова: видобування даних, виділення семантичних зв'язків, анотація, схема розмітки, корпус даних.

Introduction

Rapid development of the information technologies causes improvement of already existing and creation of new tools and methods for natural language processing. The ability to use a variety of software innovations in linguistics allows presenting languages as formal systems with clear structural features.

Among the main challenges in the natural language processing sphere are the following: automatic morphological, syntactic and semantic analysis. More complex tasks are development of the machine translation, information retrieval, summarization, text editing systems, etc.

Almost all of the mentioned tasks in this field require understanding of the semantic relations among entities. Thus relation extraction task is an actual modern issue. The main difficulty lies in teaching the system to recognize text fragments based on the specific semantic data and create correct annotations. Therefore, this task is a significant step towards the qualitative and easy automated natural language processing.

Analysis of the existing researches and publications

Among the most prominent scientists, who dedicated their researches to relation extraction field are Johann Petrak, Ken Williams, Eric Sword, Andrew Borthwick (GATE system developers). There are some researches done by national scholars as well, for example, S. Katrenko`s PhD thesis "A Closer Look at Learning Relations from Text" is an outstanding work in the field, Petro Zhezhnych, S. Buk.

Number of the developed IE systems constantly grows. Majority of them are free and has open source code, thus can be developed and improved. The most widely used systems and tools are GATE (General Architecture for Text Engineering), Callisto annotation tool, @Note, Ellogon, LingPipe, BioNotate, UAM CorpusTool. All of them has both advantages and disadvantages but their main purpose and destination is to reach and satisfy user`s needs – accomplish quick and quality text processing tasks.

Despite the fact that over the last two decades there have been a lot of researches in this sphere, this issue still remains actual. The main reason for such attention is the possibility of practical results application. Since after semantic relations are identified they can be used for questionnaires, ontologies and hypotheses developing.

Another reason for such interest to this sphere is a large number of semantic relations types. According to the root classification, all relations can be divided into general and domain specific. A significant number of existing systems was created to work with the first relations type [1]. However, when used in new subject areas, these systems require significant improvements, since for getting correct results for the specific subject area system should be able to indicate this specific industry terminology [2]. Now the largest number of relation extraction systems are developed for medical, economic, business and information technologies spheres.

Main Objectives

Nawadays there are a lot of methods for extracting information from text data. Among the most widely used are manually written patterns, supervised semi supervised and unsupervised automatic learning [3]. Each of these approaches has its advantages and disadvantages, and depends on the research goals and range of available resources. However, the initial stage, which does not depend on the recognition algorithm chosen for the system development, is the same for all approaches – to build annotated text corpora.

Thus, the main objective of this article is to describe an algorithm for creation of the annotated text corpora for further development of the relations extraction systems in medical sphere. The domain area was chosen due to the need in such kind of system for facilitation and improvement of the health systems performance while processing large amounts of text data.

Thedevelopment process can be divided into the following stages:

1. Collection and systematization of text documents.
2. Choice of software tools for annotation process implementation.
3. Annotation`s scheme development.
4. Manual annotation of the collected data.

Collection and systematization of text documents

The first issue to addressed, when creating annotated text corpus is to chose the domain area. As it was mentioned above, statistics and analysis of the latest researches show that creation of RE system can be helpful in the medical industry. According to this it was decided to focus on medical texts.

Materials for further annotation were taken from several Internet resources: reviews and questions from medical columns, medical articles in Lviv medical journal [4], online consultations, forums on medical sites [5], [6]. The total sample counts 300 units.

Choice of software tools for text data annotation

Modern information technologies market offers a wide range of software tools for automated texts processing in general and their previous annotation in particular. Among the most popular systems in free access and with open-source code are the following: Callisto annotation tool, @Note, Ellogon, LingPipe, GATE, BioNotate, UAM CorpusTool [7]. Each of the above mentioned systems has some specific text annotation means. After some investigation, it was decided to use the GATE (General Architecture for Text Engineering) system of processing natural text with open source and a set of applications written in Java.

Annotation scheme development

Creation and realization of the annotation scheme was done using the built-in module CREOLE (a Collection of Reusable Objects for Language Engineering), namely, resource Annotation scheme. With the help of this tool specific annotation type for further annotation process were distinguished. For this function GATE uses XML markup language that is supported by W3C. The schema annotation editor component (gate.gui.SchemaAnnotationEditor) was also applied. It is managed by the created xml file, and performs annotating process in the way which ensures that only specified schemes are used for annotation. Gate provides the possibility to use other components that do not limit data processing [8].

Table 1

Semantic types of named entities in medical texts

Physical Object	[Physical Object] (continued)
Organism	[Substance] (continued)
Plant	[Chemical] (continued)
Fungus	Chemical Viewed Structurally
Virus	Organic Chemical
Bacterium	Nucleic Acid, Nucleoside, or Nucleotide
Archaeon	Organophosphorus Compound
Eukaryote	Amino Acid, Peptide, or Protein
Animal	Carbohydrate
Vertebrate	Lipid
Amphibian	Steroid
Bird	Eicosanoid
Fish	Inorganic Chemical
Reptile	Element, Ion, or Isotope
Mammal	Body Substance
Human	Food
Anatomical Structure	Conceptual Entity
Embryonic Structure	Idea or Concept
Anatomical Abnormality	Temporal Concept
Congenital Abnormality	Qualitative Concept
Acquired Abnormality	Quantitative Concept
Fully Formed Anatomical Structure	Functional Concept
Body Part, Organ, or Organ Component	Body System
Tissue	Spatial Concept
Cell	Body Space or Junction
Cell Component	Body Location or Region
Gene or Genome	Molecular Sequence
Manufactured Object	Nucleotide Sequence
Medical Device	Amino Acid Sequence
Drug Delivery Device	Carbohydrate Sequence
Research Device	Geographic Area
Clinical Drug	Finding
Substance	Laboratory or Test Result
Chemical	Sign or Symptom
Chemical Viewed Functionally	Organism Attribute
Pharmacologic Substance	Clinical Attribute
Antibiotic	Intellectual Product
Biomedical or Dental Material	Classification

Continued table 1

Biologically Active Substance	Regulation or Law
Neuroreactive Substance or Biogenic Amine	Language
Hormone	Occupation or Discipline
Enzyme	Biomedical Occupation or Discipline
Vitamin	Organization
Immunologic Factor	Health Care Related Organization
Receptor	Professional Society
Indicator, Reagent, or Diagnostic Acid	Self-help or Relief Organization
Hazardous or Poisonous Substance	Group Attribute
	Group
	Professional or Occupational Group
	Population Group
	Family Group
	Age Group
	Patient or Disabled Group

Table 2

Semantic types of relations in medical texts

is_a	
associated_with	[associated_with] (continued)
physically_related_to	[functionally_related_to] (continued)
part_of	performs
consists_of	carries_out
contains	exhibits
connected_to	practices
interconnects	occurs_in
branch_of	process_of
tributary_of	users
ingredient_of	manifestation_of
spatially_related_to	indicates
location_of	result_of
adjacent_to	temporally_related_to
surrounds	co-occurs_with
traverses	precedes
functionally_related_to	conceptually_related_to
affects	evaluation_of
manages	degree_of
treats	analyzes
disrupts	assesses_effect_of
complicates	measurement_of
interacts_with	measures
prevents	diagnoses
brings_about	property_of
produces	derivative_of
causes	developmental_form_of
	method_of
	conceptual_part_of
	issue_in

Table 3

Structure of tags of type

TYPE	FEATURES	EXAMPLES
token	pos	N (Noun – Noun)
		V(Verb -Verb)
		A (Adjective – Adjective)
		P(Pronoun -Pronoun)
		R (Adverb -Adverb)
		S(Preposition -A Preposition)
		C (Conjunction -Binder)
		M (Numeral --Numeral)
		Q (Particle -Particle)
		I (Interjection -Whoop)
		Y (Abbreviation -Reduction)
		X (Residual -Balance)
	lemma	
	morphology	Tag in accordance with the morfosintakično? specification[11]

The next stage of annotation scheme development process requires determining types of named entities and relations for further annotation. The complexity of this solution lies in the chosen domain sphere – medical industry – itself. Today scientists make significant efforts in designing relation extraction systems for medical sphere, though a set of semantic entities deeply depends on the future system specification. Large number of researchers use in their experiments the Unified Medical Language System (UMLS), containing organized medical texts database, dictionaries and handbooks, etc. The system also provides entities and relations hierarchy specific for medical texts [9, 10]. After deep analysis of the existing possibilities, it was decided to build the annotation scheme based on these hierarchies (Table 1, 2).

organism.xml

```
<?xml version="1.0"?>
<schema>
    <element name="organism">
      <complexType>
        <attribute name="plant"/>
        <attribute name="fungus"/>
        <attribute name="virus"/>
        <attribute name="bacterium"/>
        <attribute name="archaeon"/>
        <attribute name="eukaryote" use="optional">
          <simpleType>
            <restriction base="string">
              <enumeration value="amphibian"/>
              <enumeration value="bird"/>
              <enumeration value="fish"/>
              <enumeration value="reptile"/>
              <enumeration value="mammal"/>
            </restriction>
          </simpleType>
        </attribute>
      </complexType>
    </element>
</schema>
```

Fig. 1. Structure of the organism.xml file

creole.xml

```
<?xml version="1.0"?>
<!-- $Id: creole.xml 16181 2012-10-29 18:20:32Z markagreenwood $ -->
<CREOLE-DIRECTORY>
    <!-- LANGUAGE RESOURCES -->

    <!-- Annotation schema -->
    <RESOURCE>
        <CLASS>gate.creole.AnnotationSchema</CLASS>


        <HIDDEN-AUTOINSTANCE>
            <PARAM NAME="xmlFileUrl"
                VALUE="resources/schema/anatomical_structure.xml" />
        </HIDDEN-AUTOINSTANCE>
        <HIDDEN-AUTOINSTANCE>
            <PARAM NAME="xmlFileUrl"
                VALUE="resources/schema/associated_with_relation.xml" />
        </HIDDEN-AUTOINSTANCE>
        <HIDDEN-AUTOINSTANCE>
            <PARAM NAME="xmlFileUrl"
                VALUE="resources/schema/biomedical_occupation_or_discipline.xml" />
        </HIDDEN-AUTOINSTANCE>
        <HIDDEN-AUTOINSTANCE>
            <PARAM NAME="xmlFileUrl"
                VALUE="resources/schema/body_substance.xml" />
        </HIDDEN-AUTOINSTANCE>
        <HIDDEN-AUTOINSTANCE>
            <PARAM NAME="xmlFileUrl"
                VALUE="resources/schema/chemical_substance.xml" />
        </HIDDEN-AUTOINSTANCE>
        <HIDDEN-AUTOINSTANCE>
            <PARAM NAME="xmlFileUrl"
                VALUE="resources/schema/finding.xml" />
        </HIDDEN-AUTOINSTANCE>
```

Fig. 2. Structure of the creole.xml file

It was decided to add another annotation type to the above mentioned named entities– token, which allows specifying parts of speech and some morphological features. This information is important for further development of the relation extraction system. Detailed description of tags is listed in table 3.

As it was mentioned previously, annotation scheme was developed with the use of AnnotationSchema class. Each annotation is described in a separate xml file, an example is shown on fig. 1. In general, there are 17 xml files for describing separate annotations. They all are defined by the configuration file creole.xml, where also are references to other resources required for data processing. Creole.xml file structure can be seen on fig. 2.

Data annotation with program means

After text data collection, all articles and materials were converted into txt format and marked with the help of Gate environment. The system interface can be found on fig. 3.

After manual annotation documents were saved in xml format for their further use. Example of the annotated text can be seen in fig. 4.

Within the obtained results it can be seen that structure of annotated units consists of attribute values (if any) and beginning and ending positions in the text. This information provides additional possibilities for further analysis and researches.

Conclusions

Creating of the annotated text corpora is the initial stage in the semantic relations recognition system developing process. The implementation of this task depends on the goals and specific features of the developing system.

This article presents the algorithm for developing and annotating medical text corpus for Ukrainian language. The domain area was chosen due to the need in such kind of system for facilitation and improvement of the health systems performance while processing large amounts of text data.

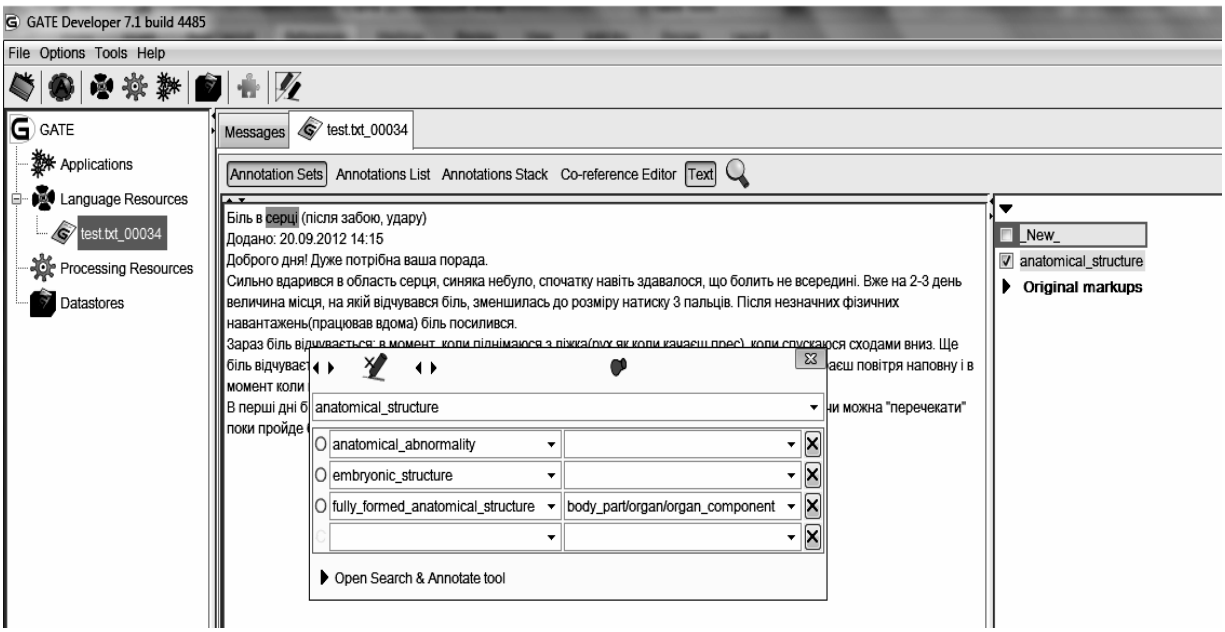
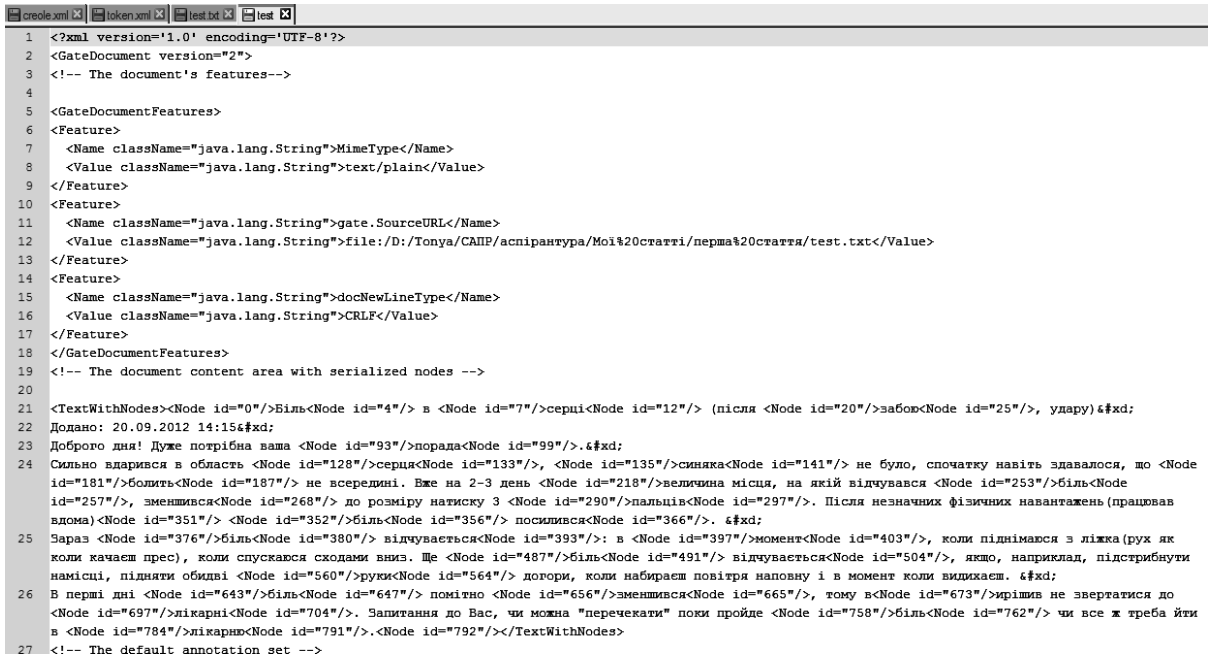


Fig. 3. Annotation using GATE with anatomic_structure tag



creole.xml | token.xml | test.txt | test

```
<?xml version='1.0' encoding='UTF-8'?>
<GateDocument version="2">
<!-- The document's features-->

<GateDocumentFeatures>
<Feature>
  <Name className="java.lang.String">MimeType</Name>
  <Value className="java.lang.String">text/plain</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">gate.SourceURL</Name>
  <Value className="java.lang.String">file:/D:/Tonya/САПР/аспірантура/Мої%20статті/перша%20стаття/test.txt</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">docNewLineType</Name>
  <Value className="java.lang.String">CRLF</Value>
</Feature>
</GateDocumentFeatures>
<!-- The document content area with serialized nodes -->

<TextWithNodes><Node id="0"/>Біль<Node id="4"/> в <Node id="7"/>серці<Node id="12"/> (після <Node id="20"/>забою<Node id="25"/>, удару)&#xd;
Додано: 20.09.2012 14:15&#xd;
Доброго дня! Дуже потрібна ваша <Node id="93"/>порада<Node id="99"/>.&#xd;
Сильно вдарився в область <Node id="128"/>серця<Node id="133"/>, <Node id="135"/>синяка<Node id="141"/> не було, спочатку навіть здавалося, що <Node id="181"/>болить<Node id="187"/> не всередині. Вже на 2-3 день <Node id="218"/>величина місця, на якій відчувався <Node id="253"/>біль<Node id="257"/>, зменшився<Node id="268"/> до розміру натиску 3 <Node id="290"/>пальців<Node id="297"/>. Після незначних фізичних навантажень(працював вдома)<Node id="351"/> <Node id="352"/>біль<Node id="356"/> посилився<Node id="366"/>. &#xd;
Зараз <Node id="376"/>біль<Node id="380"/> відчувається<Node id="393"/>: в <Node id="397"/>момент<Node id="403"/>, коли піднімаєш з ліжка(рух як коли качаєш прес), коли спускаєш сходами вниз. Ще <Node id="487"/>біль<Node id="491"/> відчувається<Node id="504"/>, якщо, наприклад, підстрибнути намісці, підняти обидві <Node id="560"/>руки<Node id="564"/> догори, коли набираєш повітря наповну і в момент коли видихаєш. &#xd;
В перші дні <Node id="643"/>біль<Node id="647"/> помітно <Node id="656"/>зменшився<Node id="665"/>, тому в<Node id="673"/>ирішив не звертатися до <Node id="697"/>лікарні<Node id="704"/>. Запитання до Вас, чи можна "перечекати" поки пройде <Node id="758"/>біль<Node id="762"/> чи все ж треба йти в <Node id="784"/>лікарню<Node id="791"/>.<Node id="792"/></TextWithNodes>
<!-- The default annotation set -->
```

Fig. 4. Result of the annotated text in xml format

At the first stage text date dedicated to the chosen sphere were gathered and structured. Analysis of the software tools for annotation process implementation was conducted. As a result GATE (General Architecture for Text Engineering) system was selected for further development. The last stage included annotation schema development and actual annotation process.

As a result, the annotated text corpus, which can be used for further research and development activities in entities and relation extraction sphere for medical domain, was obtained.

1. R. Grishman. Message Understanding Conference – 6: A brief history / Ralph Grishman, Beth Sundheim. – In Proceedings of the 16th International Conference on Computational Linguistics, 1996. – p3. 2. Katrenko S. A Closer Look at Learning Relations from Text. PhD thesis, University of Amsterdam. – 2009, – 222 p. 3. M.Mintz. Distant supervision for relation extraction without labeled data. / Mike Mintz

Steven Bills, Rion Snow, Dan Jurafsky. – 2008. – Режим доступу: http://www.stanford.edu/~jurafsky/mintz.pdf. 4. Львівський медичний часопис. Режим доступу: http://www.aml.lviv.ua/redakce/tisk.php?lanG=uk&portal=164&slozka=1303&clanek=1591& 5. Поставити запитання лікарю. Режим доступу: http://roddom.cn.ua/messages1/index.php?guestbook_page=all 6. Медісвіт – Форум. Режим доступу: http://www.medisvit.com/ua/forum. 7. M.Neves. Tools for Annotating Biomedical Texts / Neves, Leser, – Poster in the Fifth International Biocuration Conference, 2012, Washington, USA. 8. Developing Language Processing Components with GATE Version 7. Режим доступу: http://gate.ac.uk. 9. UMLS – current semantic types. Режим доступу: http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html. 10. UMLS – current relations. Режим доступу: http://www.nlm.nih.gov/research/umls/META3_current_relations.html. 11. MULTEXT-East Morphosyntactic Specifications, Version 4. Режим доступу: http://nl.ijs.si/ME/V4/msd/html/msd-uk.html.